SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 1, January 1986

PHASE PORTRAIT OF THE MATRIX RICCATI EQUATION*

MARK A. SHAYMAN†

Abstract. The matrix Riccati equation which arises from optimal control and filtering problems is a quadratic differential equation on the space of real symmetric $n \times n$ matrices. It is closely related, via compactification of the phase space, to the differential equations on the Grassmann manifold and the Lagrange-Grassmann manifold whose flows are generated by the action of one-parameter subgroups of the general linear group and of the symplectic group respectively. We determine the complete phase portraits of the Riccati equations on all three spaces. The asymptotic behavior of *every* solution is described. The phase portraits are characterized topologically as well as set-theoretically. Although the Riccati equation is not generally a Morse-Smale vector field, we are able to show that it possesses suitable generalizations of many of the important properties of Morse-Smale vector fields. In particular, the Riccati equation satisfies a generalized version of the Morse inequalities for a Morse-Smale dynamical system. In fact, for the Riccati equation, the inequalities are actually equalities.

Key words. Riccati differential equation, phase portrait, Grassmann manifold, Lagrange-Grassmann manifold, Schubert cell decomposition, Morse-Smale flows

CONTENTS

1. Introduction.

By the matrix Riccati differential equation (RDE) we mean the quadratic differential equation

$$\dot{K} = B_{21} + B_{22}K - KB_{11} - KB_{12}K$$

* Received by the editors March 17, 1983, and in revised form August 10, 1984. This research was partially supported by the National Science Foundation under grant ECS-8301015. A summary of the results described in this paper was presented at the Conference on Information Sciences and Systems, the Johns Hopkins University, March 1983.

† Department of Systems Science and Mathematics, Washington University, St. Louis, Missouri 63130.

defined on the vector space $\mathbb{R}^{m\times n}$ of real $m\times n$ matrices. B_{21}, B_{22}, B_{11}, B_{12} are constant real matrices of dimensions $m\times n$, $m\times m$, $n\times n$, and $n\times m$ respectively.

The matrix Riccati equation plays a critical role in a wide variety of applications which include transmission line phenomena, the theory of stochastic processes, optimal control and filtering, diffusion problems, and invariant imbedding [32]. It is also of independent mathematical interest because it is the description in local coordinates of the differential equation on the Grassmann manifold whose flow is given by the action of a 1-parameter subgroup of the general linear group. This is discussed below.

We are particularly interested in the Riccati equation which arises in optimal control and filtering problems. It has the form

$$\dot{K} = -Q - A'K - KA + KLK$$

defined on the space $\mathbb{R}^{n\times n}$ of real $n\times n$ matrices. A, L, Q are constant real $n\times n$ matrices with L and Q symmetric and L nonnegative definite. The vector space $S(n)$ of real symmetric $n\times n$ matrices is an invariant manifold for this differential equation, and it is the restriction to $S(n)$ which is important in the applications. Thus, we will regard this differential equation as defined on $S(n)$, and call it the *symplectic Riccati differential equation* (SRDE) for reasons which will become apparent. The role of the SRDE in linear quadratic control problems is described in [5].

It has been known at least since the time of Poincaré that the topology of certain differential equations in the plane could be clarified by extending the domain to the projective plane. It was observed by C. Schneider [33] that the natural compactification of the domain $\mathbb{R}^{m\times n}$ for the RDE is the Grassmann manifold $G^n(\mathbb{R}^{n+m})$ of n-dimensional subspaces of $\mathbb{R}^{n+m}$. We will refer to the Riccati equation on $G^n(\mathbb{R}^{n+m})$ as the *extended Riccati differential equation* (ERDE). The natural compactification of the domain $S(n)$ for the SRDE has been described by R. Hermann and C. Martin [16], [27]. It is the so-called Lagrange-Grassmann manifold $\mathscr{L}(n)$ consisting of those subspaces belonging to $G^n(\mathbb{R}^{2n})$ on which a certain skew-symmetric bilinear form vanishes identically. We will refer to the symplectic Riccati equation on $\mathscr{L}(n)$ as the *extended symplectic Riccati differential equation* (ESRDE). The compactification of the domains of the RDE and SRDE is useful when investigating the behavior of trajectories at infinity. However, the principal advantage of the compactifications is that for both the ERDE and the ESRDE, the flow is obtained from the action of a 1-parameter subgroup of a matrix Lie group. For the ERDE, the Lie group is the general linear group $\mathrm{Gl}(n+m,\mathbb{R})$ while for the ESRDE, the Lie group is the symplectic group $\mathrm{Sp}(n,\mathbb{R})$.

The primary purpose of this paper is to give the complete phase portrait for the Riccati equation arising from the control and filtering applications, i.e. the symplectic Riccati differential equation on $S(n)$. However, the procedure we follow leads us to first determine the complete phase portraits for the ERDE and the ESRDE. Our results describe the asymptotic behavior of *every* solution for the ERDE, ESRDE, and SRDE. Under generic assumptions, the distinctive features of the phase portraits of Riccati equations are (1) a nonwandering set which consists of tori of various dimensions, and (2) stable and unstable manifolds which are unions of so-called Schubert cells. It was shown recently [18] that matrix Riccati equations on $G^n(\mathbb{R}^{n+m})$ are not generically Morse-Smale.[1] Our results show that although Riccati equations are generally not Morse-Smale, their phase portraits reflect the topology of the underlying manifold in an analogous way. For example, we show that Riccati vector fields satisfy a suitable

[1] This corrects the contrary announcement in [19].

generalization of the Morse-Smale inequalities. In fact, for the Riccati equation, the inequalities are actually equalities. We also use the phase portrait of the Riccati equation to obtain new derivations of the mod 2 Betti numbers for the Grassmann manifold and the Lagrange-Grassmann manifold.

In the penultimate section, we extend our results to include almost all of the Riccati equations which are excluded by the generic assumptions in force throughout the earlier sections. Under the relaxed assumptions, the connected components of the nonwandering set are products of Grassmannians, while the stable and unstable manifolds remain unions of Schubert cells.

Much of the research on Riccati equations has focused on the "stabilizing" equilibrium point of the SRDE, i.e. the equilibrium point K^+ which has the property that every eigenvalue of the closed loop plant matrix $A-LK^+$ has negative real part. However, recent work has also involved study of periodic solutions [29], [18], [19], [37] and convergence (and nonconvergence) of trajectories to equilibrium points other than K^+ [8]. Although there is a vast literature on Riccati equations, as far as we are aware, our results provide for the first time a complete description of the asymptotic behavior of every solution under assumptions which are satisfied by almost every SRDE which arises from control and filtering problems.

We have used similar techniques to obtain the phase portrait for the matrix Riccati equation in which the coefficient matrices are periodic functions of time [38], [40].

2. Extension of the phase space. In this section, we review how the RDE is extended to the ERDE on $G^n(\mathbb{R}^{n+m})$ [33], and how the SRDE is extended to the ESRDE on $\mathscr{L}(n)$ [16], [27]. Let $\psi:\mathbb{R}^{m\times n}\to G^n(\mathbb{R}^{n+m})$ be defined by $\psi(K)=\operatorname{Sp}\left[\begin{smallmatrix} I_n \\ K \end{smallmatrix}\right]$, the column space of the $(n+m)\times n$ full rank matrix $\left[\begin{smallmatrix} I_n \\ K \end{smallmatrix}\right]$. Let $G_0^n(\mathbb{R}^{n+m})$ consist of those subspaces in $G^n(\mathbb{R}^{n+m})$ which are complementary to the m-dimensional subspace $\operatorname{Sp}\left[\begin{smallmatrix} 0 \\ I_m \end{smallmatrix}\right]$. Then ψ embeds the Euclidean space $\mathbb{R}^{m\times n}$ in $G^n(\mathbb{R}^{n+m})$ as the open and dense subset $G_0^n(\mathbb{R}^{n+m})$. In fact, $(G_0^n(\mathbb{R}^{n+m}), \psi^{-1})$ is one of the standard charts for the manifold $G^n(\mathbb{R}^{n+m})$. (See Appendix.) Thus, $G^n(\mathbb{R}^{n+m})$ can be viewed as a compactification of $\mathbb{R}^{m\times n}$.

Let $K(t, K_0)$ denote the solution of the RDE for the initial point K_0. Let B denote the $(n+m)\times(n+m)$ matrix $\left[\begin{smallmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{smallmatrix}\right]$. Define a flow on $G^n(\mathbb{R}^{n+m})$ by $S(t, S_0)=e^{Bt}(S_0)$, where $e^{Bt}(S_0)$ is the image of the subspace S_0 under $e^{Bt}\in \mathrm{Gl}(n+m, \mathbb{R})$. It is straightforward to verify that we have

$$\psi(K(t, K_0)) = S(t, \psi(K_0))$$

whenever $K(t, K_0)$ exists. This means that the RDE is the local expression with respect to the chart $(G_0^n(\mathbb{R}^{n+m}), \psi^{-1})$ for the differential equation on $G^n(\mathbb{R}^{n+m})$ which corresponds to the flow $S(t, S_0)$. To say this another way, if we use the embedding ψ to identify $\mathbb{R}^{m\times n}$ with $G_0^n(\mathbb{R}^{n+m})$, then the restriction to $G_0^n(\mathbb{R}^{n+m})$ of the flow $S(t, S_0)$ on $G^n(\mathbb{R}^{n+m})$ is identified with the flow of the RDE. $K(t, K_0)$ ceases to exist (due to finite escape time) precisely when $S(t, \psi(K_0))$ leaves the subset $G_0^n(\mathbb{R}^{n+m})$. By the *extended Riccati differential equation* (ERDE), we mean the differential equation on $G^n(\mathbb{R}^{n+m})$ whose flow is given by $S(t, S_0)=e^{Bt}(S_0)$. Thus, the flow of the ERDE is given by the action of a one-parameter subgroup of $\mathrm{Gl}(n+m, \mathbb{R})$ on $G^n(\mathbb{R}^{n+m})$.

Next we describe how the symplectic Riccati differential equation is extended to the Lagrange-Grassmann manifold $\mathscr{L}(n)$. 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 $\omega(x, y)=x'Jy$. Then $\mathscr{L}(n)$ is defined to be the subset $\{S\in G^n(\mathbb{R}^{2n})\colon \omega(x, y)=0, \forall x, y\in S\}$ of $G^n(\mathbb{R}^{2n})$. If the standard inner product is assigned to $\mathbb{R}^{2n}$, then $\mathscr{L}(n)=\{S\in G^n(\mathbb{R}^{2n})\colon J(S)=S^\perp\}$. We will say that a subspace S of $\mathbb{R}^{2n}$ (of any dimension) is Lagrangian iff $J(S)\perp S$.

Let $\mathrm{Sp}(n, \mathbb{R})$ denote the symplectic group, i.e. $\mathrm{Sp}(n, \mathbb{R}) = \{P \in \mathrm{Gl}(2n, \mathbb{R}): P'JP = J\}$. Let $\mathrm{sp}(n, \mathbb{R})$ denote its Lie algebra. $\mathrm{sp}(n, \mathbb{R})$ consists of all $2n \times 2n$ real matrices H which satisfy $JH + H'J = 0$. This is equivalent to requiring that H have the partitioned form $\left[\begin{smallmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{smallmatrix}\right]$ with each submatrix $n \times n$, $H_{12} = H'_{12}$, $H_{21} = H'_{21}$, and $H_{22} = -H'_{11}$. If $S \in \mathscr{L}(n)$ and $P \in \mathrm{Sp}(n, \mathbb{R})$, then $P(S) \in \mathscr{L}(n)$, so $\mathrm{Sp}(n, \mathbb{R})$ acts on $\mathscr{L}(n)$. In fact, it can be shown [16] that $\mathscr{L}(n)$ is a homogeneous space of $\mathrm{Sp}(n, \mathbb{R})$. It can also be shown [17] that $\mathscr{L}(n)$ is the homogeneous space $U(n)/O(n)$, where $U(n)$ and $O(n)$ are the unitary and orthogonal groups respectively.

It is easy to check that if K is an $n \times n$ matrix, then $\mathrm{Sp}\left[\begin{smallmatrix} I_n \\ K \end{smallmatrix}\right]$ belongs to $\mathscr{L}(n)$ iff K is symmetric. Thus we can define $\phi: S(n) \to \mathscr{L}(n)$ by $\phi(K) = \mathrm{Sp}\left[\begin{smallmatrix} I_n \\ K \end{smallmatrix}\right]$. Let $\mathscr{L}_0(n)$ consist of those subspaces in $\mathscr{L}(n)$ which are complementary to the n-dimensional subspace $\mathrm{Sp}\left[\begin{smallmatrix} 0 \\ I_n \end{smallmatrix}\right]$. Then ϕ embeds the Euclidean space $S(n)$ in $\mathscr{L}(n)$ as the open and dense subset $\mathscr{L}_0(n)$. Thus, $\mathscr{L}(n)$ can be viewed as a compactification of $S(n)$.

Let $K(t, K_0)$ denote the solution of the SRDE for the initial point K_0, and let $H \in \mathrm{sp}(n, \mathbb{R})$ denote the $2n \times 2n$ matrix

$$\begin{bmatrix} A & -L \\ -Q & -A' \end{bmatrix}.$$

H is often referred to as the Hamiltonian matrix associated with the Riccati equation. Define a flow on $\mathscr{L}(n)$ by $S(t, S_0) = e^{Ht}(S_0)$. We have

$$\phi(K(t, K_0)) = S(t, \phi(K_0))$$

whenever $K(t, K_0)$ exists. Thus, $S(t, S_0)$ is the extension to $\mathscr{L}(n)$ of the flow of the SRDE on $S(n)$. By the *extended symplectic Riccati differential equation* (ESRDE), we mean the differential equation on $\mathscr{L}(n)$ whose flow is given by $S(t, S_0) = e^{Ht}(S_0)$. Thus, the flow of the ESRDE is given by the action of a one-parameter subgroup of $\mathrm{Sp}(n, \mathbb{R})$ on $\mathscr{L}(n)$.

In the next three sections, we successively determine the phase portraits of the ERDE, ESRDE, and SRDE. When we consider the ESRDE, we will drop the assumption that the symmetric matrix L be nonnegative definite. However, when we consider the SRDE, we will reinstate this assumption.

The use of the Grassmann manifold in the theory of the Riccati equation is closely related to the use of the state-costate equations in the study of the Riccati equation. Partition the matrix e^{Ht} as

$$\begin{bmatrix} P_{11}(t) & P_{12}(t) \\ P_{21}(t) & P_{22}(t) \end{bmatrix}.$$

Then it is well known [5, p. 156] that the solution of the SRDE with initial condition K_0 is given by

$$(*) \qquad K(t, K_0) = (P_{21}(t) + P_{22}(t)K_0)(P_{11}(t) + P_{12}(t)K_0)^{-1}.$$

This formula is valid in the largest time interval containing 0 on which the indicated inverse exists. This formula is equivalent to the formula

$$(**) \qquad \phi(K(t, K_0)) = S(t, \phi(K_0))$$

given above, which relates the flow of the ESRDE on $\mathscr{L}(n)$ to the flow of SRDE on

$S(n)$. To see this, note that

$$\begin{aligned} S(t, \phi(K_0)) &= e^{Ht}\left(\mathrm{Sp}\begin{bmatrix} I \\ K_0 \end{bmatrix}\right) \\ &= \mathrm{Sp}\begin{bmatrix} P_{11}(t)+P_{12}(t)K_0 \\ P_{21}(t)+P_{22}(t)K_0 \end{bmatrix} \\ &= \mathrm{Sp}\begin{bmatrix} I \\ (P_{21}(t)+P_{22}(t)K_0)(P_{11}(t)+P_{12}(t)K_0)^{-1} \end{bmatrix} \\ &= \phi((P_{21}(t)+P_{22}(t)K_0)(P_{11}(t)+P_{12}(t)K_0)^{-1}). \end{aligned}$$

Thus, (**) is equivalent to the formula

$$\phi(K(t, K_0)) = \phi((P_{21}(t)+P_{22}(t)K_0)(P_{11}(t)+P_{12}(t)K_0)^{-1}).$$

Since ϕ is injective, this is equivalent to (*).

Since (*) is equivalent to (**), we could use (*) and thereby avoid introducing the Grassmann manifold. This would have two drawbacks. The first is that the ERDE and the ESRDE turn out to be extremely interesting differential equations in themselves, when studied from the point of view of differentiable dynamical systems on compact manifolds. The second drawback is that the phase portrait of the Riccati equation (even when considered in the usual sense as a differential equation on $S(n)$) is closely related to the topology of the Grassmann manifold. In particular, we shall see that the Schubert cell decomposition of the Grassmann manifold plays an essential role in the description of the phase portrait. If we were to avoid introducing the Grassmann manifold, we would be losing the key insight obtainable from its topology.

3. Phase portrait of the extended Riccati differential equation. In this section, we determine the complete phase portrait for the ERDE on $G^n(\mathbb{R}^{n+m})$. We make the following assumptions which we denote collectively as Assumption A1: (1) the $n+m$ eigenvalues of B are distinct, and (2) if λ_i and λ_j are a pair of eigenvalues with the same real part, then $\lambda_i = \bar{\lambda}_j$. (Overbar denotes complex conjugation.) Each of these assumptions holds generically. These assumptions are relaxed considerably in § 6.

We fix some notation. Let p denote the number of real eigenvalues of B, and let q denote the number of conjugate pairs of nonreal eigenvalues of B. (So $p+2q = n+m$.) Let $r = p+q$, and let $E_1, \cdots, E_r$ denote the primary components of B ordered according to increasing real part of the corresponding eigenvalue(s). Thus, each E_j is either one-dimensional or two-dimensional. Also, if $i<j$ and if λ_i and λ_j are eigenvalues which correspond to E_i and E_j respectively, then $\mathrm{Re}\,\lambda_i < \mathrm{Re}\,\lambda_j$.

In the very special case where B has only real eigenvalues (i.e. $q=0$), the phase portrait of the ERDE was described by C. R. Schneider [33]. This reference also considers the corresponding discrete dynamical system associated with the Riccati equation in which the coefficient matrices are periodic. Similar results in the discrete case were obtained by S. Batterson [1]. N. Kuiper has considered the discrete case in the special case when $n=1$ or $m=1$ [22].

3.1. Nonwandering set, stable and unstable manifolds. The Grassmann manifold $G^n(\mathbb{R}^{n+m})$ can be given the structure of a metric space by defining on it the so-called "gap metric" ρ. (See [14] for details concerning this metric.) If $S_1, S_2 \in G^n(\mathbb{R}^{n+m})$ and if P_1, P_2 are the orthogonal projections onto S_1, S_2 respectively, then $\rho(S_1, S_2) =^{\mathrm{def}} \|P_1 - P_2\|$ (operator norm). $G^n(\mathbb{R}^{n+m})$ is a compact (and hence complete) metric space in the metric ρ.

The gap metric is widely used by analysts. However, topologists define a topology on $G^n(\mathbb{R}^{n+m})$ in a different way. Let $V_n(\mathbb{R}^{n+m})$ denote the set of all $(n+m)\times n$ full rank matrices with real entries. $V_n(\mathbb{R}^{n+m})$ is an open subset of $\mathbb{R}^{(n+m)\times n}$ and thus has the standard Euclidean topology. ($V_n(\mathbb{R}^{n+m})$ is known as a Stiefel manifold.) Define a mapping q: $V_n(\mathbb{R}^{n+m})\to G^n(\mathbb{R}^{n+m})$ with $q(Y)$ the column space Sp Y of the matrix Y. Then $G^n(\mathbb{R}^{n+m})$ is given the quotient topology induced by the surjective map q. In other words, a subset U of $G^n(\mathbb{R}^{n+m})$ is open if and only if $q^{-1}(U)$ is open in $V_n(\mathbb{R}^{n+m})$. It is not hard to show that the quotient topology is in fact the same as the topology induced by the gap metric.

The standard definition of a complex-valued almost periodic function [12] can be generalized to a function which takes values in a complete metric space. Let (X, ρ_X) 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 Appendix B.

Since the flow of the ERDE is given by $S(t, S_0) = e^{Bt}(S_0)$, it follows immediately that S_0 is an equilibrium point iff S_0 is B-invariant. Given our assumptions on B, it is clear that the equilibrium points consist of those n-dimensional subspaces which are direct sums of various of the primary components. If B has only real eigenvalues, there are $\binom{n+m}{n}$ equilibrium points, but in general there will be fewer.

One of the principal features of the phase portrait of the ERDE is the presence of invariant tori of various dimensions. The existence of invariant tori of dimension at least two was first described in a recent paper by Hermann and Martin [18] on the periodic orbits of the ERDE. It will prove convenient to view equilibrium points and isolated periodic orbits as zero-dimensional and one-dimensional invariant tori, respectively.

Let $l=(l_1,\cdots,l_r)$ be an (unordered) partition of n into r parts such that $0\leqq l_j\leqq \dim E_j$, $j=1,\cdots,r$. Let $G^{l_j}(E_j)$ denote the Grassmann manifold of all l_j-dimensional subspaces of E_j, and let $T(l)=\{S_1\oplus\cdots\oplus S_r\colon S_j\in G^{l_j}(E_j), j=1,\cdots,r\}$. Then $T(l)$ is isomorphic to the product $G^{l_1}(E_1)\times\cdots\times G^{l_r}(E_r)$. Since E_j has dimension equal to one or two, $G^{l_j}(E_j)$ consists of a single point unless $\dim E_j=2$ and $l_j=1$ in which case it is the projective line, which is topologically the circle S^1. Thus, $T(l)$ is a torus, and the dimension of $T(l)$ is equal to $\sum_{j=1}^r l_j(\dim E_j - l_j)$. Let $S\in T(l)$. Then $S=S_1\oplus\cdots\oplus S_r$ with S_j an l_j-dimensional subspace of E_j. $e^{Bt}(S)=e^{Bt}(S_1)\oplus\cdots\oplus e^{Bt}(S_r)$ which belongs to $T(l)$ since $S_j\subset E_j$ and E_j is B-invariant. Thus, the torus $T(l)$ is both positively and negatively invariant with respect to the flow of the ERDE.

Next we consider the flow on the invariant torus $T(l)$. Let k denote the dimension of $T(l)$. If $k=0$, then $T(l)$ is an equilibrium point. Thus, we suppose that $k>0$. Let $j_1,\cdots,j_k$ denote the elements of the set $\{j\colon l_j=1 \text{ and } \dim E_j=2\}$. For each $\nu=1,\cdots,k$, let $\sigma_{j_\nu}\pm i\omega_{j_\nu}$ be the conjugate pair of eigenvalues corresponding to the 2-dimensional primary component E_{j_ν}. We can choose a basis for E_{j_ν} such that the matrix for the restriction $B|E_{j_\nu}$ is

$$\begin{bmatrix} \sigma_{j_\nu} & \omega_{j_\nu} \\ -\omega_{j_\nu} & \sigma_{j_\nu} \end{bmatrix}.$$

Then the matrix for $e^{Bt}|E_{j_\nu}$ is

$$e^{\sigma_{j_\nu}t}\begin{bmatrix} \cos\omega_{j_\nu}t & \sin\omega_{j_\nu}t \\ -\sin\omega_{j_\nu}t & \cos\omega_{j_\nu}t \end{bmatrix}.$$

It follows that if S_{j_ν} is any 1-dimensional subspace of E_{j_ν}, then S_{j_ν} is e^{BT}-invariant iff T is an integer multiple of π/ω_{j_ν}. Let $S = S_1 \oplus \cdots \oplus S_r \in T(l)$. If $j \notin \{j_1, \cdots, j_k\}$, then either $S_j = 0$ or $S_j = E_j$. In either case $e^{BT}(S_j) = S_j$, $\forall t$. Thus, $e^{BT}(S) = S$ iff $e^{BT}(S_{j_\nu}) = S_{j_\nu}$, $\nu = 1, \cdots, k$. Thus, $e^{BT}(S) = S$ iff T is an integer multiple of each of the numbers π/ω_{j_ν}, $\nu = 1, \cdots, k$. Such a T exists iff $\{\pi/\omega_{j_\nu}: \nu = 1, \cdots, k\}$ are commensurable, or equivalently iff $\{\omega_{j_\nu}: \nu = 1, \cdots, k\}$ are commensurable. Thus, if $\{\omega_{j_\nu}: \nu = 1, \cdots, k\}$ are commensurable, then every orbit on $T(l)$ is periodic with minimum period equal to the least common multiple of the numbers π/ω_{j_ν}, $\nu = 1, \cdots, k$. Otherwise, no orbit on $T(l)$ is periodic.

Suppose that $\{\omega_{j_\nu}: \nu = 1, \cdots, k\}$ are not necessarily all commensurable. By choice of basis, we may assume that $B = \operatorname{diag}\{B_1, \cdots, B_r\}$ where $B_i = [\sigma_i]$ if $\dim E_i = 1$ and $B_i = \left[\begin{smallmatrix} \sigma_i & \omega_i \\ -\omega_i & \sigma_i \end{smallmatrix}\right]$ if $\dim E_i = 2$. Let $\tilde{B}$ be the matrix obtained from B by setting every entry on the main diagonal equal to 0. For fixed t, the restriction $e^{\tilde{B}t}|E_i$ differs from $e^{Bt}|E_i$ only by a constant factor. It follows that if $S_0 \in T(l)$, $e^{\tilde{B}t}(S_0) = e^{Bt}(S_0)$. Given such an S_0, let $X_0 \in V_n^0(\mathbb{R}^{n+m})$, the compact subset of $V_n(\mathbb{R}^{n+m})$ consisting of those matrices which have orthonormal columns, with $q(X_0) = S_0$. Then $S(t, S_0) = q(e^{\tilde{B}t}X_0)$. Since each entry of $e^{\tilde{B}t}X_0$ is a periodic function, it follows from AP 3 in Appendix B that $e^{\tilde{B}t}X_0$ is almost periodic. Since $e^{\tilde{B}t}$ is orthogonal, $e^{\tilde{B}t}X_0 \in V_n^0(\mathbb{R}^{n+m})$ for all t. Since the restriction of the continuous mapping q to the compact set $V_n^0(\mathbb{R}^{n+m})$ is uniformly continuous, it follows from AP 5 that $q(e^{\tilde{B}t}X_0)$ is almost periodic. Hence, $S(t, S_0)$ is almost periodic.

The preceding argument shows that if $\{\omega_{j_\nu}: \nu = 1, \cdots, k\}$ are not all commensurable, then every motion on $T(l)$ is almost periodic, but not periodic. It is easy to see that if no pair of $\{\omega_{j_\nu}: \nu = 1, \cdots, k\}$ are commensurable, then every trajectory on $T(l)$ is dense in $T(l)$. However, if at least two of the ω_{j_ν} are commensurable, then no trajectory is dense.

There is one additional property of the motion on $T(l)$ which we will need to use later. Let $S_0, \tilde{S}_0 \in T(l)$. Since $S(t, S_0)$ and $S(t, \tilde{S}_0)$ are almost periodic, it follows from AP 6 that if $\rho(S(t, S_0), S(t, \tilde{S}_0)) \to 0$ as $t \to \infty$ or $t \to -\infty$, then $S_0 = \tilde{S}_0$.

We summarize the preceding analysis as a theorem, which is mostly due to Hermann and Martin [18].

THEOREM 1. *$T(l)$ is a torus of dimension $k = \sum_{j=1}^r l_j(\dim E_j - l_j)$ which is both positively and negatively invariant. If $k = 0$, $T(l)$ is an equilibrium point. If $k > 0$ and $\{\omega_{j_\nu}: \nu = 1, \cdots, k\}$ are all commensurable, then each $S \in T(l)$ generates a periodic motion with period equal to the least common multiple of $\{\pi/\omega_{j_\nu}: \nu = 1, \cdots, k\}$. Otherwise, each $S \in T(l)$ generates an almost periodic motion which is dense in $T(l)$ iff no pair of the ω_{j_ν} are commensurable. In all cases, if $S, \tilde{S} \in T(l)$ with $\rho(e^{Bt}(\tilde{S}), e^{Bt}(S)) \to 0$ as $t \to \infty$ or $t \to -\infty$, then $S = \tilde{S}$.*

Let $T(l)$ be an invariant torus, and let $W^s(T(l))$ and $W^u(T(l))$ be the *stable* and *unstable manifolds* respectively for $T(l)$. In other words, $W^s(T(l)) = \{S_0 \in G^n(\mathbb{R}^{n+m}): S(t, S_0) \to T(l) \text{ as } t \to \infty\}$ and $W^u(T(l)) = \{S_0 \in G^n(\mathbb{R}^{n+m}): S(t, S_0) \to T(l) \text{ as } t \to -\infty\}$. Define a flag (i.e. strictly increasing sequence) of subspaces $0 = M_0 \subset M_1 \subset \cdots \subset M_r = \mathbb{R}^{n+m}$ by $M_k = \bigoplus_{j=1}^k E_j$, $k = 1, \cdots, r$. Define a second flag of subspaces $0 = N_0 \subset N_1 \subset \cdots \subset N_r = \mathbb{R}^{n+m}$ by $N_k = \bigoplus_{j=1}^k E_{r-j+1}$, $k = 1, \cdots, r$. We refer to $\{M_k\}_1^r$ as the *stable flag* associated with the ERDE, and we refer to $\{N_k\}_1^r$ as the *unstable flag* associated with the ERDE. The next theorem describes the sets $W^s(T(l))$ and $W^u(T(l))$.

THEOREM 2.

(a)

$$W^s(T(l)) = \left\{ S \in G^n(\mathbb{R}^{n+m}): \dim S \cap M_k = \sum_{i=1}^k l_i, k = 1, \cdots, r \right\}.$$

(b)

$$W^u(T(l)) = \left\{ S \in G^n(\mathbb{R}^{n+m}) \colon \dim S \cap N_k = \sum_{i=1}^{k} l_{r-i+1},\ k = 1, \cdots, r \right\}.$$

Proof. (a) Let

$$\tilde{W}^s(T(l)) = \left\{ S \in G^n(\mathbb{R}^{n+m}) \colon \dim S \cap M_k = \sum_{i=1}^{k} l_i,\ k = 1, \cdots, r \right\}.$$

The collection of sets $\{\tilde{W}^s(T(l))\}_l$ (where l ranges over all unordered partitions of n into r parts such that $0 \leqq l_i \leqq \dim E_i$, $i = 1, \cdots, r$) is a partition of $G^n(\mathbb{R}^{n+m})$. Hence, it suffices to show that $\tilde{W}^s(T(l)) \subseteq W^s(T(l))$ for all l. Let $S_0 \in \tilde{W}^s(T(l))$. By starting with a basis for $S_0 \cap M_1$ and successively extending to bases for $S_0 \cap M_2, \cdots, S_0 \cap M_r = S_0$, it is straightforward to show that there exists a basis for S_0 of the form $\{v_{ij} + w_{ij} \colon i$ is such that $l_i \neq 0$ and $j = 1, \cdots, l_i\}$ where $v_{ij} \in E_i$, $w_{ij} \in M_{i-1}$, and such that $\{v_{ij}\}$ is linearly independent. Letting $S_1 = \operatorname{Sp}\{v_{ij}\}$, we have $S_1 \in T(l)$. By choice of basis, we may assume that $B = \operatorname{diag}\{B_1, \cdots, B_r\}$ where $B_i = [\sigma_i]$ if $\dim E_i = 1$ and $B_i = \left[\begin{smallmatrix} \sigma_i & \omega_i \\ -\omega_i & \sigma_i \end{smallmatrix}\right]$ if $\dim E_i = 2$, and where $\sigma_1 < \sigma_2 < \cdots < \sigma_r$. Using the standard inner product on $\mathbb{R}^{n+m}$, it follows that $E_i \perp E_k$ whenever $i \neq k$. Consequently, $v_{ij} \perp v_{kp}$ 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 $v_{ij} \perp v_{ip}$ and such that v_{ij} has unit length. Thus, without loss of generality, we may assume that $\{v_{ij}\}$ is an orthonormal basis for S_1.

Let $\tilde{B}$ be the matrix obtained from B by setting every entry on the main diagonal equal to 0. Then $e^{Bt}x = e^{\sigma_i t} e^{\tilde{B}t}x$, $\forall x \in E_i$. Extend $v_{i1}, \cdots, v_{il_i}$ to an orthonormal basis for E_i by adding vectors $\{u_{ik} \colon k = l_i + 1, \cdots, \dim E_i\}$. We use $\{v_{ij}, u_{ik}\}$ to denote the resulting orthonormal basis for $\mathbb{R}^{n+m}$. Another basis for $\mathbb{R}^{n+m}$ is $\{v_{ij} + w_{ij}, u_{ik}\}$. Let $X(t)$, $Y(t)$, $Z(t)$ be the matrices whose columns are $\{e^{\tilde{B}t}v_{ij}\}$, $\{e^{\tilde{B}t}u_{ik}\}$, $\{e^{-\sigma_i t} e^{Bt}w_{ij}\}$ respectively. Then $S(t, S_1) = \operatorname{Sp} X(t)$ and $S(t, S_0) = \operatorname{Sp}(X(t) + Z(t))$. Since $e^{\tilde{B}t}$ and $[X(0), Y(0)]$ are both orthogonal matrices, $[X(t), 0]$, $[X(t), Y(t)]^{-1}$ is the orthogonal projection onto $S(t, S_1)$. Since $[X(t) + Z(t), 0]$, $[X(t) + Z(t), Y(t)]^{-1}$ is a projection onto $S(t, S_0)$ (in general not orthogonal), it follows from a basic property of the gap metric [14, p. 361] that

$$\begin{aligned}
\rho(S(t, S_1), S(t, S_0)) &\leqq \|[X(t), 0][X(t), Y(t)]^{-1} \\
&\quad - [X(t) + Z(t), 0][X(t) + Z(t), Y(t)]^{-1}\| \\
&= \|[X(t), 0]([X(t), Y(t)]^{-1} - [X(t) + Z(t), Y(t)]^{-1}) \\
&\quad - [Z(t), 0][X(t) + Z(t), Y(t)]^{-1}\| \\
&\leqq \|[X(t), 0]\|\, \|[X(t), Y(t)]^{-1}\|\, \|[X(t) + Z(t), Y(t)]^{-1}\|\, \|[Z(t), 0]\| \\
&\quad + \|[Z(t), 0]\|\, \|[X(t) + Z(t), Y(t)]^{-1}\|.
\end{aligned}$$

Since $[X(t), Y(t)]$ is an orthogonal matrix and the orthogonal group is compact, it follows that $[X(t), 0]$ and $[X(t), Y(t)]^{-1}$ are bounded. The compactness of the orthogonal group together with the fact that $Z(t) \to 0$ as $t \to \infty$ implies that $[X(t) + Z(t), Y(t)]^{-1}$ is bounded. Then the fact that $Z(t) \to 0$ as $t \to \infty$ implies that $\rho(S(t, S_1), S(t, S_0)) \to 0$ as $t \to \infty$. Thus, $\tilde{W}^s(T(l)) \subseteq W^s(T(l))$ which shows that $\tilde{W}^s(T(l)) = W^s(T(l))$ and completes the proof of (a).

The proof of (b) is completely analogous to the proof of (a). □

Remark 1. We have defined $W^s(T(l))$ only in terms of asymptotic convergence. However, it follows easily from the inequality in the proof of Theorem 2 that $W^s(T(l))$ has the additional property that given any $\varepsilon > 0$, there exists $\delta > 0$ such that if $S_0 \in W^s(T(l))$ with $d(S_0, T(l)) < \delta$, then $d(S(t, S_0), T(l)) < \varepsilon$, for all $t \geqq 0$. Corresponding statements apply to $W^u(T(l))$.

COROLLARY 1. *The sets* $\{W^s(T(l))\}_l$ *are a partition of* $G^n(\mathbb{R}^{n+m})$. *The sets* $\{W^u(T(l))\}_l$ *are also a partition of* $G^n(\mathbb{R}^{n+m})$.

For a flow ϕ_t on a manifold M, $x \in M$ is called a *wandering point* [2], [42] if there exists a neighborhood U of x in M and some $t_0 > 0$ such that $\phi_t(U) \cap U$ is empty whenever $|t| > t_0$. The subset of M consisting of all points which are not wandering points is called the *nonwandering set.* The next corollary follows immediately from Theorem 1 and Corollary 1 of Theorem 2.

COROLLARY 2. *The nonwandering set of the* ERDE *is the disjoint union* $\bigsqcup_l T(l)$ *of the invariant tori.*

We will use Ω to denote the nonwandering set of the ERDE.

From the proof of Theorem 2, we know that if $S_0 \in W^s(T(l))$, then there exists $S_1 \in T(l)$ such that $S(t, S_0) \to S(t, S_1)$ as $t \to \infty$. Similarly, if $S_0 \in W^u(T(l'))$, then there exists $S_2 \in T(l')$ such that $S(t, S_0) \to S(t, S_2)$ as $t \to -\infty$. Let η_j denote the projection onto the subspace E_j along the subspace $\bigoplus_{i=1, i\neq j}^{r} E_i$, $j = 1, \cdots, r$. Define a mapping $\Pi_+ : G^n(\mathbb{R}^{n+m}) \to \Omega$ by

$$\Pi_+(S) = \eta_1(S \cap M_1) \oplus \eta_2(S \cap M_2) \oplus \cdots \oplus \eta_r(S \cap M_r).$$

From the construction of the subspace S_1 in the proof of Theorem 2, it follows that $S_1 = \Pi_+(S_0)$. In other words, $S(t, S_0) \to S(t, \Pi_+(S_0))$ as $t \to \infty$. Similarly, we define a mapping $\Pi_- : G^n(\mathbb{R}^{n+m}) \to \Omega$ by

$$\Pi_-(S) = \eta_1(S \cap N_r) \oplus \eta_2(S \cap N_{r-1}) \oplus \cdots \oplus \eta_r(S \cap N_1).$$

It is easily seen that $S_2 = \Pi_-(S_0)$. Thus, $S(t, S_0) \to S(t, \Pi_-(S_0))$ as $t \to -\infty$.

The problem we now consider is: given $S_1 \in T(l)$, describe the subset $\Pi_+^{-1}(S_1)$ of $W^s(T(l))$ ($\Pi_-^{-1}(S_1)$ of $W^u(T(l))$ which contains those subspaces S_0 with the property that $S(t, S_0) \to S(t, S_1)$ as $t \to \infty$ ($t \to -\infty$).

Let $T(l)$ be an invariant torus, and let S_1 be a given point in $T(l)$. Let $W^s(S_1) = \Pi_+^{-1}(S_1)$, and let $W^u(S_1) = \Pi_-^{-1}(S_1)$.

THEOREM 3.

(a)

$$W^s(S_1) = \left\{ S \in W^s(T(l)) \colon \dim S \cap [M_{k-1} \oplus (S_1 \cap E_k)] = \sum_{i=1}^{k} l_i,\ k = 1, \cdots, r \right\}.$$

(b)

$$W^u(S_1) = \left\{ S \in W^u(T(l)) \colon \dim S \cap [N_{k-1} \oplus (S_1 \cap E_{r-k+1})] = \sum_{i=1}^{k} l_{r-i+1},\ k = 1, \cdots, r \right\}.$$

Proof. (a) Suppose $S \in W^s(S_1)$. Then it follows from the proof of Theorem 2 that there exists a basis for S of the form $\{v_{ij} + w_{ij} \colon 1 \leqq i \leqq r$ such that $l_i \neq 0;\ j = 1, \cdots, l_i\}$ where $v_{ij} \in E_i$, $w_{ij} \in M_{i-1}$, and $\{v_{ij}\}$ is a basis for S_1. It follows immediately that $S \cap [M_{k-1} \oplus (S_1 \cap E_k)] = \operatorname{Sp}\{v_{ij} + w_{ij} \colon 1 \leqq i \leqq k$ such that $l_i \neq 0;\ j = 1, \cdots, l_i\}$ which has dimension $\sum_{i=1}^{k} l_i$.

Conversely, suppose $S \in W^s(T(l))$ and $\dim S \cap [M_{k-1} \oplus (S_1 \cap E_k)] = \sum_{i=1}^{k} l_i$, $k = 1, \cdots, r$. Then $\dim S \cap [M_{k-1} \oplus (S_1 \cap E_k)] = \dim S \cap M_k$, so $S \cap [M_{k-1} \oplus (S_1 \cap E_k)] =$

$S \cap M_k$. Hence, $\eta_k(S \cap M_k) = \eta_k(S \cap [M_{k-1} \oplus (S_1 \cap E_k)]) \subseteq \eta_k(M_{k-1} \oplus (S_1 \cap E_k)) = S_1 \cap E_k$, which implies that $\Pi_+(S) = S_1$, completing the proof of (a).

The proof of (b) is analogous to the proof of (a). □

It is worth noting that the last conclusion in Theorem 1 implies that if $S_1 \neq S_2$, then $W^s(S_1)$ and $W^s(S_2)$ ($W^u(S_1)$ and $W^u(S_2)$) are disjoint.

3.2. Topology of the stable and unstable manifolds. Next we consider the topological structure of the subsets $\{W^s(T(l))\}$ and of the subsets $\{W^u(T(l))\}$, and their relationship to the topology of the Grassmann manifold $G^n(\mathbb{R}^{n+m})$.

Let X be a locally compact topological space. A *cell decomposition* of X is partition $\{X_i\}_{i \in I}$ of X into disjoint subsets such that (1) $\{X_i\}_{i \in I}$ is locally finite; (2) X_i is homeomorphic to $\mathbb{R}^{n_i}$ for some n_i; (3) $\bar{X}_i - X_i$ is the union of some of the cells $\{X_j: \dim X_j < \dim X_i\}$.

There is a well-known cell decomposition of $G^n(\mathbb{R}^{n+m})$ which is sometimes referred to as the *Schubert cell decomposition* of the Grassmann manifold. (See e.g. [15, p. 194].) It is constructed by first choosing a flag of subspaces $0 = V_0 \subset V_1 \subset V_2 \subset \cdots \subset V_{n+m-1} \subset V_{n+m} = \mathbb{R}^{n+m}$ such that $\dim V_j = j, j = 0, \cdots, n+m$. To emphasize that this flag contains a subspace for each dimension between 0 and $n+m$, we will refer to it as a *complete flag*. (If B has any nonreal eigenvalues, then $r < n+m$, so the flags $\{M_j\}$ and $\{N_j\}$ are not complete.) For each $(n+m)$-tuple $a = (a_1, \cdots, a_{n+m})$ such that $a_j = 0$ or $a_j = 1$ and $\sum_{j=1}^{n+m} a_j = n$, let $U(a) = \{S \in G^n(\mathbb{R}^{n+m}): \dim S \cap V_j = \sum_{i=1}^{j} a_i,\ j = 1, \cdots, n+m\}$. Then the $\binom{n+m}{n}$ sets $\{U(a)\}$ partition $G^n(\mathbb{R}^{n+m})$, and it is not hard to show that $U(a)$ is homeomorphic to Euclidean space of dimension $\sum_{j=1}^{n+m} a_j(j - \sum_{i=1}^{j} a_i)$. To show this, we observe that by choosing a basis, we may assume that $V_j = \operatorname{Sp}\{e_1, \cdots, e_j\}$, where $e_1, \cdots, e_{n+m}$ are the standard basis vectors for $\mathbb{R}^{n+m}$. Let $j_1 < j_2 < \cdots < j_n$ be the n elements of the set $\{j: a_j = 1\}$. Then each $S \in U(a)$ has a unique $(n+m) \times n$ basis matrix Z_S in column echelon form. Specifically, rows $j_1, \cdots, j_n$ of Z form an $n \times n$ identity submatrix, and $z_{ik} = 0$ if $i > j_k$. Otherwise z_{ik} is arbitrary. For example, if $n = 2$, $m = 3$, and $a = (0, 1, 0, 0, 1)$, then each $S \in U(a)$ is spanned by a unique matrix of the form

$$Z_S = \begin{bmatrix} z_{11} & z_{12} \\ 1 & 0 \\ 0 & z_{32} \\ 0 & z_{42} \\ 0 & 1 \end{bmatrix}.$$

The correspondence $S \leftrightarrow Z_S$ gives a homeomorphism (in fact, a real-analytic isomorphism) between $U(a)$ and Euclidean space of dimension $\sum_{j=1}^{n+m} a_j(j - \sum_{i=1}^{j} a_i)$. This dimension is also given by the formula $\sum_{k=1}^{n} (j_k - k)$.

The decomposition $G^n(\mathbb{R}^{n+m}) = \bigsqcup_a U(a)$ is the Schubert cell decomposition of $G^n(\mathbb{R}^{n+m})$. The closure of the cell $U(a)$ is the set $\{S \in G^n(\mathbb{R}^{n+m}): \dim S \cap V_j \geqq \sum_{i=1}^{j} a_i,\ j = 1, \cdots, n+m\}$. It is called a *Schubert variety*.

Let $T(l)$ be a given invariant torus. Refine $\{M_j\}$ to a complete flag by inserting a subspace M_j' between M_{j-1} and M_j whenever $\dim E_j = 2$. Since $W^s(T(l)) = \{S \in G^n(\mathbb{R}^{n+m}): \dim S \cap M_j = \sum_{i=1}^{j} l_i,\ j = 1, \cdots, r\}$, it follows that $W^s(T(l))$ is a union of one or more of the $\binom{n+m}{n}$ cells $U(a)$ which correspond to the complete flag. Let $S \in W^s(T(l))$. Then the dimensions of each of $S \cap M_j$ $(j = 1, \cdots, r)$ are completely determined by the assumption that S belongs to $W^s(T(l))$. Let j be given, and suppose that $\dim E_j = 2$. Then there is a subspace M_j' which is between M_{j-1} and M_j in the complete flag. There are three cases to consider.

(1) If $l_j = 0$, then $\dim S \cap M_j = \dim S \cap M_{j-1}$, and this must also be the dimension of $S \cap M_j'$.

(2) If $l_j = 2$, then $\dim S \cap M_j = \dim S \cap M_{j-1} + 2$, which implies that $\dim S \cap M_j' = \dim S \cap M_{j-1} + 1$.

(3) If $l_j = 1$, then $\dim S \cap M_j = \dim S \cap M_{j-1} + 1$, and there are two possibilities: $\dim S \cap M_j' = \dim S \cap M_{j-1}$ or $\dim S \cap M_j' = \dim S \cap M_j$.

Suppose that the given invariant torus $T(l)$ is k-dimensional. Then $\{j\colon l_j = 1$ and $\dim E_j = 2\}$ contains exactly k elements, say $j_1, \cdots, j_k$. If $S \in W^s(T(l))$, the analysis in the preceding paragraph shows that there are 2^k possible choices for the set of dimensions in which S intersects the $n+m$ subspaces in the complete flag. Thus, $W^s(T(l))$ is the union of exactly 2^k cells $U(a)$ which correspond to the complete flag. Each cell is given by fixing a vector $b = (b_1, \cdots, b_k)$ with b_j equal to 0 or 1, and setting $W^s(T(l)), b) = \{S \in W^s(T(l))\colon \dim S \cap M_{j_\nu}' = \dim S \cap M_{j_\nu - 1} + b_\nu,\ \nu = 1, \cdots, k\}$.

To determine the dimension of $W^s(T(l), b)$, we determine the vector $a = (a_1, \cdots, a_{n+m})$ which corresponds to the given vectors $l = (l_1, \cdots, l_r)$ and $b = (b_1, \cdots, b_k)$. Let $j \in \{1, \cdots, r\}$ and set $j' = \sum_{i=1}^{j-1} \dim E_i + 1 = \dim M_{j-1} + 1$. If $\dim E_j = 1$, then $a_{j'} = l_j$. If $\dim E_j = 2$ and $l_j = 0$, then $a_{j'} = 0$ and $a_{j'+1} = 0$. If $\dim E_j = 2$ and $l_j = 2$, then $a_{j'} = 1$ and $a_{j'+1} = 1$. If $\dim E_j = 2$ and $l_j = 1$, then $j = j_\nu$ for some ν, and we have $a_{j_\nu'} = b_\nu$ and $a_{j_\nu'+1} = 1 - b_\nu$. With this definition of a, it follows that $W^s(T(l), b)$ is analytically isomorphic to Euclidean space of dimension $\sum_{\gamma=1}^{n+m} a_\gamma(\gamma - \sum_{i=1}^{\gamma} a_i)$. It is straightforward to show that the dimension of $W^s(T(l), b)$ can be reexpressed as

$$\sum_{j=1}^{r} l_j\left(\dim M_{j-1} - \sum_{i=1}^{j-1} l_i\right) + k - \sum_{\nu=1}^{k} b_\nu.$$

This proves part (a) of the next theorem. The proof of part (b) is completely analogous to the proof of (a).

THEOREM 4. *Let $T(l)$ be a k-dimensional invariant torus. Then* (a) *$W^s(T(l))$ is the disjoint union of 2^k cells. Exactly $\binom{k}{\nu}$ of these cells have dimension $\sum_{j=1}^{r} l_j(\dim M_{j-1} - \sum_{i=1}^{j-1} l_i) + \nu$, $\nu = 0, \cdots, k$.* (b) *$W^u(T(l))$ is the disjoint union of 2^k cells. Exactly $\binom{k}{\nu}$ of these cells have dimension $\sum_{j=1}^{r} l_{r-j+1}(\dim N_{j-1} - \sum_{i=1}^{j-1} l_{r-i+1}) + \nu$, $\nu = 0, \cdots, k$.*

The next result describes the topology of the sets $W^s(S_1)$ and $W^u(S_1)$.

THEOREM 5. *Let $T(l)$ be an invariant torus, and let $S_1 \in T(l)$. Then* (a) *$W^s(S_1)$ is analytically isomorphic to Euclidean space of dimension $\sum_{j=1}^{r} l_j(\dim M_{j-1} - \sum_{i=1}^{j-1} l_i)$.* (b) *$W^u(S_1)$ is analytically isomorphic to Euclidean space of dimension $\sum_{j=1}^{r} l_{r-j+1}(\dim N_{j-1} - \sum_{i=1}^{j-1} l_{r-i+1})$.*

Proof. Let $k = \dim T(l)$, and let $j_1, \cdots, j_k$ be the elements of $\{j\colon l_j = 1$ and $\dim E_j = 2\}$. Refine $\{M_j\}_1^r$ to a complete flag in two steps. First, define $M_{j_\nu}' = M_{j_\nu - 1} \oplus (S_1 \cap E_{j_\nu})$ and insert it between $M_{j_\nu - 1}$ and M_{j_ν}, $\nu = 1, \cdots, k$. This refines $\{M_j\}_1^r$ at each j such that $\dim E_j = 2$ and $l_j = 1$. Second, if $\dim E_j = 2$ and $l_j = 0$ or 2, insert any subspace M_j' between M_{j-1} and M_j. The refined flag is now a complete flag, and with respect to this flag, we have $W^s(S_1) = W^s(T(l), b)$ provided that $b = (1, 1, \cdots, 1)$. It then follows from the proof of Theorem 4 that $W^s(S_1)$ is analytically isomorphic to Euclidean space of dimension $\sum_{j=1}^{r} l_j(\dim M_{j-1} - \sum_{i=1}^{j-1} l_i)$.

The proof of (b) is analogous to the proof of (a). □

The next result shows that the ERDE has either a unique asymptotically stable equilibrium point or a unique orbitally asymptotically stable periodic orbit, but not both. In fact, somewhat more than this is true.

THEOREM 6. (a) *If $\sum_{j=\nu+1}^{r} \dim E_j = n$ for some ν, then there exists an equilibrium point whose stable manifold is open and dense. Otherwise, there exists a periodic orbit whose stable manifold is open and dense.* (b) *If $\sum_{j=1}^{\nu} \dim E_j = n$ for some ν, then there*

exists an equilibrium point whose unstable manifold is open and dense. Otherwise, there exists a periodic orbit whose unstable manifold is open and dense.

Proof. (a) *Suppose that* $\sum_{j=\nu+1}^{r} \dim E_j = n$ for some ν. Let $l_j = 0$, $j = 1, \cdots, \nu$ and let $l_j = \dim E_j$, $j = \nu+1, \cdots, r$. Then $T(l)$ is a 0-dimensional invariant torus, and hence an equilibrium point. By Theorem 2, $S \in W^s(T(l))$ iff $\dim S \cap M_k = \sum_{j=1}^{k} l_j$, $k = 1, \cdots, r$, which implies that $W^s(T(l)) = \{S \in G^n(\mathbb{R}^{n+m})\colon \dim S \cap M_\nu = 0\}$. Since $\dim M_\nu = m$, the complement of $W^s(T(l))$ is therefore a proper subvariety of $G^n(\mathbb{R}^{n+m})$, which shows that $W^s(T(l))$ is open and dense in $G^n(\mathbb{R}^{n+m})$.

Now suppose that there does not exist ν such that $\sum_{j=\nu+1}^{r} \dim E_j = n$. Since $\dim E_j$ is 1 or 2, there must exist ν such that $\dim E_{\nu+1} = 2$ and $\sum_{j=\nu+1}^{r} \dim E_j = n+1$. Let $l_j = 0$, $j = 1, \cdots, \nu$, let $l_{\nu+1} = 1$, and let $l_j = \dim E_j$, $j = \nu+2, \cdots, r$. Then $T(l)$ is a 1-dimensional invariant torus, and hence a periodic orbit. $S \in W^s(T(l))$ iff $\dim S \cap M_\nu = 0$, $\dim S \cap M_{\nu+1} = 1$, and $\dim S \cap M_k = 1 + \sum_{j=\nu+2}^{k} \dim E_j$, $k = \nu+2, \cdots, r$, or equivalently $W^s(T(l)) = \{S \in G^n(\mathbb{R}^{n+m})\colon \dim S \cap M_\nu = 0, \dim S \cap M_{\nu+1} = 1\}$. Since $\dim M_{\nu+1} = m+1$, it follows that $\dim S \cap M_{\nu+1} \geqq 1$ for every $S \in G^n(\mathbb{R}^{n+m})$. Thus, the complement of $W^s(T(l))$ is $\{S \in G^n(\mathbb{R}^{n+m})\colon \dim S \cap M_\nu \geqq 1\} \cup \{S \in G^n(\mathbb{R}^{n+m})\colon \dim S \cap M_{\nu+1} \geqq 2\}$, which is a subvariety of $G^n(\mathbb{R}^{n+m})$. Since $\dim M_\nu = m-1$ and $\dim M_{\nu+1} = m+1$, this is a proper subvariety. (Alternatively, note that $W^s(T(l))$ is nonempty since it contains the periodic orbit $T(l)$.) This completes the proof of (a). The proof of (b) is completely analogous. □

If $T(l)$ is a k-dimensional invariant torus, Theorem 4 describes the stable manifold $W^s(T(l))$ of $T(l)$ as the union of 2^k cells. We now describe the topology of $W^s(T(l))$ in greater detail. To avoid cumbersome notation and to serve as an illustration, we consider a concrete example. However, the example contains all of the features of the general case. Furthermore, it will become obvious that the proof for the general result follows step-by-step the analysis of the example.

Let $n = 3$ and $m = 5$, so the ERDE is a differential equation on $G^3(\mathbb{R}^8)$, which is 15-dimensional. Let $r = 6$, and suppose that $\dim E_1 = 1$, $\dim E_2 = 2$, $\dim E_3 = 1$, $\dim E_4 = 2$, $\dim E_5 = 1$, $\dim E_6 = 1$. Let $l = (l_1, \cdots, l_6) = (0, 1, 0, 1, 0, 1)$. Then $l_2 = 1$, $l_4 = 1$, $\dim E_2 = 2$, $\dim E_4 = 2$, so $T(l)$ is a 2-dimensional invariant torus. By Theorem 4, $W^s(T(l))$ is the union of 4 open cells of dimensions 9, 10, 10, 11. By changing basis, we may assume that when the flag $M_0 \subset M_1 \subset M_2 \subset M_3 \subset M_4 \subset M_5 \subset M_6$ is refined to a complete flag by inserting subspaces M_2' and M_4' between M_1, M_2 and between M_3, M_4 respectively, the flag $V_0 \subset V_1 \subset V_2 \subset V_3 \subset V_4 \subset V_5 \subset V_6 \subset V_7 \subset V_8$ is obtained, where $V_j = \mathrm{Sp}\{e_1, \cdots, e_j\}$. ($e_j$ is the jth standard basis vector of $\mathbb{R}^8$.) The 4 cells are given by

$$
\begin{aligned}
U_1 = \{S \in G^3(\mathbb{R}^8)\colon\ & \dim S \cap V_1 = 0, \dim S \cap V_2 = 1,\\
& \dim S \cap V_3 = 1, \dim S \cap V_4 = 1, \dim S \cap V_5 = 2,\\
& \dim S \cap V_6 = 2, \dim S \cap V_7 = 2, \dim S \cap V_8 = 3\},\\
U_2 = \{S \in G^3(\mathbb{R}^8)\colon\ & \dim S \cap V_1 = 0, \dim S \cap V_2 = 0,\\
& \dim S \cap V_3 = 1, \dim S \cap V_4 = 1, \dim S \cap V_5 = 2,\\
& \dim S \cap V_6 = 2, \dim S \cap V_7 = 2, \dim S \cap V_8 = 3\},\\
U_3 = \{S \in G^3(\mathbb{R}^8)\colon\ & \dim S \cap V_1 = 0, \dim S \cap V_2 = 1,\\
& \dim S \cap V_3 = 1, \dim S \cap V_4 = 1, \dim S \cap V_5 = 1,\\
& \dim S \cap V_6 = 2, \dim S \cap V_7 = 2, \dim S \cap V_8 = 3\},
\end{aligned}
$$

$$U_4 = \{S \in G^3(\mathbb{R}^8): \dim S \cap V_1 = 0, \dim S \cap V_2 = 0,$$
$$\dim S \cap V_3 = 1, \dim S \cap V_4 = 1, \dim S \cap V_5 = 1,$$
$$\dim S \cap V_6 = 2, \dim S \cap V_7 = 2, \dim S \cap V_8 = 3\}.$$

These cells are isomorphic to $\mathbb{R}^9$, $\mathbb{R}^{10}$, $\mathbb{R}^{10}$, $\mathbb{R}^{11}$ respectively. They are parametrized using column echelon form as follows:

$$\operatorname{Sp}\begin{bmatrix} x_1 & y_1 & z_1 \\ 1 & 0 & 0 \\ 0 & y_3 & z_3 \\ 0 & y_4 & z_4 \\ 0 & 1 & 0 \\ 0 & 0 & z_6 \\ 0 & 0 & z_7 \\ 0 & 0 & 1 \end{bmatrix}, \quad \operatorname{Sp}\begin{bmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ 1 & 0 & 0 \\ 0 & y_4 & z_4 \\ 0 & 1 & 0 \\ 0 & 0 & z_6 \\ 0 & 0 & z_7 \\ 0 & 0 & 1 \end{bmatrix}, \quad \operatorname{Sp}\begin{bmatrix} x_1 & y_1 & z_1 \\ 1 & 0 & 0 \\ 0 & y_3 & z_3 \\ 0 & y_4 & z_4 \\ 0 & y_5 & z_5 \\ 0 & 1 & 0 \\ 0 & 0 & z_7 \\ 0 & 0 & 1 \end{bmatrix}, \quad \operatorname{Sp}\begin{bmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ 1 & 0 & 0 \\ 0 & y_4 & z_4 \\ 0 & y_5 & z_5 \\ 0 & 1 & 0 \\ 0 & 0 & z_7 \\ 0 & 0 & 1 \end{bmatrix}.$$

U_1 U_2 U_3 U_4

These 4 cells can be modified in an obvious way to obtain 4 charts W_1, W_2, W_3, W_4 which cover $W^s(T(l))$.

$$\operatorname{Sp}\begin{bmatrix} x_1 & y_1 & z_1 \\ 1 & 0 & 0 \\ x_3 & y_3 & z_3 \\ 0 & y_4 & z_4 \\ 0 & 1 & 0 \\ 0 & y_6 & z_6 \\ 0 & 0 & z_7 \\ 0 & 0 & 1 \end{bmatrix}, \quad \operatorname{Sp}\begin{bmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ 1 & 0 & 0 \\ 0 & y_4 & z_4 \\ 0 & 1 & 0 \\ 0 & y_6 & z_6 \\ 0 & 0 & z_7 \\ 0 & 0 & 1 \end{bmatrix}, \quad \operatorname{Sp}\begin{bmatrix} x_1 & y_1 & z_1 \\ 1 & 0 & 0 \\ x_3 & y_3 & z_3 \\ 0 & y_4 & z_4 \\ 0 & y_5 & z_5 \\ 0 & 1 & 0 \\ 0 & 0 & z_7 \\ 0 & 0 & 1 \end{bmatrix}, \quad \operatorname{Sp}\begin{bmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ 1 & 0 & 0 \\ 0 & y_4 & z_4 \\ 0 & y_5 & z_5 \\ 0 & 1 & 0 \\ 0 & 0 & z_7 \\ 0 & 0 & 1 \end{bmatrix}.$$

W_1 W_2 W_3 W_4

Each of these charts is a submanifold chart relative to one of the standard charts for $G^3(\mathbb{R}^8)$. (The standard charts for the Grassmann manifold are described in Appendix A.) Thus, $W^s(T(l))$ is an embedded submanifold of $G^3(\mathbb{R}^8)$.

$T(l)$ is itself covered by 4 submanifold charts $\bar{W}_1$, $\bar{W}_2$, $\bar{W}_3$, $\bar{W}_4$ defined by

$$\operatorname{Sp}\begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ x_3 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & y_6 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \operatorname{Sp}\begin{bmatrix} 0 & 0 & 0 \\ x_2 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & y_6 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \operatorname{Sp}\begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ x_3 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & y_5 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \operatorname{Sp}\begin{bmatrix} 0 & 0 & 0 \\ x_2 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & y_5 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

$\bar{W}_1$ $\bar{W}_2$ $\bar{W}_3$ $\bar{W}_4$

The mapping $\Pi_+: G^n(\mathbb{R}^{n+m}) \to \Omega$ maps $W^s(T(l))$ onto $T(l)$. We will also use Π_+ to denote its restriction to $W^s(T(l))$. If $S_1 \in T(l)$, then $\Pi_+^{-1}(S_1) = W^s(S_1)$. Consequently,

the fiber $\Pi_+^{-1}(S_1)$ is isomorphic to Euclidean space of dimension $\sum_{j=1}^{r} l_j(\dim M_{j-1} - \sum_{i=1}^{j-1} l_i)$, which in this example is equal to 9. It is clear that $\Pi_+^{-1}(\bar{W}_i) = W_i$, $i = 1, 2, 3, 4$. Also there is an obvious isomorphism γ_i: $W_i \to \bar{W}_i \times \mathbb{R}^9$ with the property that if p_i: $\bar{W}_i \times \mathbb{R}^9 \to \bar{W}_i$ is the natural projection, then $p_i \circ \gamma_i$ is the restriction of Π_+ to W_i. For example, let $i = 1$, let S denote the element of W_1 whose coordinates are $(x_1, x_3, y_1, y_3, y_4, y_6, z_1, z_3, z_4, z_6, z_7)$, and let $\bar{S}$ denote the element of $\bar{W}_1$ whose coordinates are (x_3, y_6). Then $\gamma_1(S) = (\bar{S}, (x_1, y_1, y_3, y_4, z_1, z_3, z_4, z_6, z_7))$.

The analysis given for the preceding example can be applied essentially unchanged to describe the structure of $W^s(T(l))$ for an arbitrary invariant torus $T(l)$. If k denotes the dimension of $T(l)$, then by Theorem 4, $W^s(T(l))$ is the union of 2^k cells, the largest of which has dimension $\sum_{j=1}^{r} l_j(\dim M_{j-1} - \sum_{i=1}^{j-1} l_i) + k$. From these cells we obtain 2^k submanifold charts for $W^s(T(l))$. Thus, $W^s(T(l))$ is an embedded submanifold of $G^n(\mathbb{R}^{n+m})$ of dimension $\sum_{j=1}^{r} l_j(\dim M_{j-1} - \sum_{i=1}^{j-1} l_i) + k$. The projection Π_+: $W^s(T(l)) \to T(l)$ is defined as above. If $S_1 \in T(l)$, then $\Pi_+^{-1}(S_1) = W^s(S_1)$ which is isomorphic to Euclidean space of dimension $d_s = \sum_{j=1}^{r} l_j(\dim M_{j-1} - \sum_{i=1}^{j-1} l_i)$. Each chart W_i $(i = 1, \cdots, 2^k)$ for $W^s(T(l))$ is the inverse image of a chart $\bar{W}_i$ for $T(l)$. Also, there exist real-analytic isomorphisms γ_i: $W_i \to \bar{W}_i \times \mathbb{R}^{d_s}$ such that $p_i \circ \gamma_i = \Pi_+ | W_i$. Thus, Π_+: $W^s(T(l)) \to T(l)$ is a locally trivial bundle with $\mathbb{R}^{d_s}$ as fiber. It is easy to see that the transition functions for this bundle are invertible polynomial mappings of $\mathbb{R}^{d_s}$ (rational functions in which the denominators are functions only of base point).

The next theorem summarizes these conclusions as well as the analogous results for the unstable manifolds.

THEOREM 7. *Let $T(l)$ be a k-dimensional invariant torus. Then*

(a) *$W^s(T(l))$ is an embedded submanifold of $G^n(\mathbb{R}^{n+m})$ of dimension $k + d_s$, where $d_s = \sum_{j=1}^{r} l_j(\dim M_{j-1} - \sum_{i=1}^{j-1} l_i)$.*

(b) *Π_+: $W^s(T(l)) \to T(l)$ is a locally trivial bundle with fiber $\mathbb{R}^{d_s}$ and polynomial transition functions.*

(c) *$W^u(T(l))$ is an embedded submanifold of $G^n(\mathbb{R}^{n+m})$ of dimension $k + d_u$, where $d_u = \sum_{j=1}^{r} l_{r-j+1}(\dim N_{j-1} - \sum_{i=1}^{j-1} l_{r-i+1})$.*

(d) *Π_-: $W^u(T(l)) \to T(l)$ is a locally trivial bundle with fiber $\mathbb{R}^{d_u}$ and polynomial transition functions.*

3.3. Morse theory and structural stability. Next, we wish to determine exactly when the ERDE is a *Morse-Smale vector field.* By definition [42], a Morse-Smale vector field is a smooth vector field on a compact manifold with the following 3 properties: (1) The nonwandering set is the union of a finite number of equilibria $x_1, \cdots, x_m$ and a finite number of closed orbits $\gamma_1, \cdots, \gamma_n$ of the flow. (2) Every equilibrium point and every closed orbit is hyperbolic. (3) The stable and unstable manifolds of the x_i, γ_j intersect each other only transversally.

We start with a lemma. Suppose that $\mathbb{R}^k = F_1 \oplus \cdots \oplus F_k$ with $\dim F_j = 1$ for each j. Define two complete flags $\{V_j\}_1^k$, $\{W_j\}_1^k$ such that $V_j = \bigoplus_{i=1}^{j} F_i$, $W_j = \bigoplus_{i=1}^{j} F_{k-i+1}$. Let $\alpha = (\alpha_1, \cdots, \alpha_k)$ and $\beta = (\beta_1, \cdots, \beta_k)$ be such that $\alpha_i = 0$ or 1, $\beta_i = 0$ or 1, and $\sum_{i=1}^{k} \alpha_i = n = \sum_{i=1}^{k} \beta_i$. Define $X(\alpha) = \{S \in G^n(\mathbb{R}^k)$: $\dim S \cap V_j = \sum_{i=1}^{j} \alpha_i$, $j = 1, \cdots, k\}$ and $Y(\beta) = \{S \in G^n(\mathbb{R}^k)$: $\dim S \cap W_j = \sum_{i=1}^{j} \beta_{k-i+1}$, $j = 1, \cdots, k\}$.

LEMMA 1. (a) *$X(\alpha) \cap Y(\beta)$ is nonempty iff $\sum_{i=1}^{j} \alpha_i \leqq \sum_{i=1}^{j} \beta_i$, $j = 1, \cdots, k$.*

(b) *$X(\alpha)$ and $Y(\beta)$ intersect transversally.*

Proof. (a) First suppose that the intersection is nonempty, and let $S \in X(\alpha) \cap Y(\beta)$. By construction, $V_j \oplus W_{k-j} = \mathbb{R}^k$. Thus, $\dim S \cap V_j + \dim S \cap W_{k-j} \leqq \dim S = n$, which implies that $\sum_{i=1}^{j} \alpha_i + \sum_{i=1}^{k-j} \beta_{k-i+1} \leqq n$. Hence, $\sum_{i=1}^{j} \alpha_i \leqq n - \sum_{i=1}^{k-j} \beta_{k-i+1} = \sum_{i=1}^{j} \beta_i$.

Conversely, suppose that $\sum_{i=1}^{j}\alpha_i \leqq \sum_{i=1}^{j}\beta_i$, $j=1,\cdots,k$. Let $j_1<\cdots<j_n$ be the elements of $\{j\colon \alpha_j=1\}$, and let $l_1<\cdots<l_n$ be the elements of $\{l\colon \beta_l=1\}$. Then $j_p=\min\{j\colon \sum_{i=1}^{j}\alpha_i=p\}$ and $l_p=\min\{l\colon \sum_{i=1}^{l}\beta_i=p\}$. Since $\sum_{i=1}^{j}\alpha_i\leqq\sum_{i=1}^{j}\beta_i$, it follows that $l_p\leqq j_p$ for all p. Let f_i be a basis vector for the 1-dimensional subspace F_i, $i=1,\cdots,k$. Define a subspace S by $S=\mathrm{Sp}\{f_{j_1}+f_{l_1},\ f_{j_2}+f_{l_2},\cdots,f_{j_n}+f_{l_n}\}$. We claim that $S\in X(\alpha)\cap Y(\beta)$.

Since $j_1<\cdots<j_n$ and $l_i\leqq j_i$ for all i, it follows immediately that $\dim S\cap V_j=0$ for $j<j_1$, $\dim S\cap V_j=1$ for $j_1\leqq j<j_2,\cdots,\dim S\cap V_j=n-1$ for $j_{n-1}\leqq j<j_n$, $\dim S\cap V_j=n$ for $j_n\leqq j$. But this is equivalent to the condition that S belong to $X(\alpha)$. We also have $k-l_n+1<\cdots<k-l_1+1$ and $k-j_i+1\leqq k-l_i+1$ for all i. This implies that $\dim S\cap W_j=0$ for $j<k-l_n+1$, $\dim S\cap W_j=1$ for $k-l_n+1\leqq j<k-l_{n-1}+1,\cdots,\dim S\cap W_j=n-1$ for $k-l_2+1\leqq j<k-l_1+1$, $\dim S\cap W_j=n$ for $k-l_1+1\leqq j$. But this is equivalent to the condition that S belong to $Y(\beta)$. Thus, $S\in X(\alpha)\cap Y(\beta)$, so the intersection is nonempty.

(b) If $X(\alpha)$ and $Y(\beta)$ are disjoint, the assertion is trivially satisfied, so we can assume that $X(\alpha)\cap Y(\beta)$ is nonempty. Then (a) implies that $\sum_{i=1}^{j}\alpha_i\leqq\sum_{i=1}^{j}\beta_i$, $i=1,\cdots,k$. Define $j_1,\cdots,j_n$ and $l_1,\cdots,l_n$ as in the proof of (a). Also define $f_1,\cdots,f_k$ as in (a). Let $S_0\in X(\alpha)\cap Y(\beta)$. Let $J=\{j_1,\cdots,j_n\}$. Since $S_0\in X(\alpha)$, S_0 has a basis of the form $\{f_{j_1}+\sum_{i=1}^{j_1-1}z_{i1}f_i, f_{j_2}+\sum_{i=1,i\notin J}^{j_2-1}z_{i2}f_i,\cdots,f_{j_n}+\sum_{i=1,i\notin J}^{j_n-1}z_{in}f_i\}$. Let Z_0 be the $k\times n$ rank n matrix whose columns are the coordinates of these basis vectors with respect to the basis $\{f_1,\cdots,f_k\}$ for $\mathbb{R}^k$. Then the ipth entry of Z_0 is 1 if $i=j_p$, 0 if $i>j_p$, z_{ip} if $i<j_p$ and $i\notin J$, and 0 if $i<j_p$ and $i\in J$.

Now, let $M=\mathrm{Sp}\{f_{j_1},\cdots,f_{j_n}\}$ and let $N=\mathrm{Sp}\{f_i\colon i\notin J\}$. Corresponding to the indicated bases for M and N is a chart for $G^n(\mathbb{R}^k)$ which parametrizes all the subspaces which are complementary to N. In particular, S_0 is contained in this chart. This chart associates the $(k-n)n$ coordinates $\{a_{ip}\colon 1\leqq i\leqq k \text{ with } i\notin J, 1\leqq p\leqq n\}$ with the subspace $\mathrm{Sp}\{f_{j_p}+\sum_{i=1,i\notin J}^{k}a_{ip}f_i\colon p=1,\cdots,n\}$. It follows that if v is an arbitrary tangent vector to $G^k(\mathbb{R}^n)$ at S_0, then there exist $\{a_{ip}\}$ such that $v=d/dt|_{t=0}S(t)$, where $S(t)$ is the curve in $G^k(\mathbb{R}^n)$ with $S(0)=S_0$ described by

$$S(t)=\mathrm{Sp}\left\{f_{j_p}+\sum_{\substack{i=1\\ i\notin J}}^{j_p-1}z_{ip}f_i+t\sum_{\substack{i=1\\ i\notin J}}^{k}a_{ip}f_i\colon p=1,\cdots,n\right\}.$$

Let

$$S_1(t)=\mathrm{Sp}\left\{f_{j_p}+\sum_{\substack{i=1\\ i\notin J}}^{j_p-1}(z_{ip}+ta_{ip})f_i\colon p=1,\cdots,n\right\},$$

$$S_2(t)=\mathrm{Sp}\left\{f_{j_p}+\sum_{\substack{i=1\\ i\notin J}}^{j_p-1}z_{ip}f_i+t\sum_{\substack{i=j_p+1\\ i\notin J}}^{k}a_{ip}f_i\colon p=1,\cdots,n\right\},$$

$$v_1=\frac{d}{dt}\bigg|_{t=0}S_1(t),\qquad v_2=\frac{d}{dt}\bigg|_{t=0}S_2(t).$$

Then $S_1(0)=S_2(0)=S_0$, and $v_1+v_2=v$.

It is clear that $S_1(t)\in X(\alpha)$ for all t. Thus, v_1 is a tangent vector to the submanifold $X(\alpha)$ at S_0. We claim that $S_2(t)\in Y(\beta)$ for all t, which implies that v_2 is a tangent vector to the submanifold $Y(\beta)$ at S_0. This would show that $T_{S_0}(X(\alpha))+T_{S_0}(Y(\beta))=T_{S_0}(G^n(\mathbb{R}^k))$ and thereby complete the proof.

To prove the claim, let $Z_2(t)$ be the $k\times n$ rank n matrix whose columns are the coordinates of the indicated basis vectors for $S_2(t)$ with respect to the basis $\{f_1,\cdots,f_k\}$

for $\mathbb{R}^k$. Then the ipth entry of $Z_2(t)$ is 1 if $i=j_p$, 0 if $i\in J$ but $i\neq j_p$, z_{ip} if $i<j_p$ and $i\notin J$, and ta_{ip} if $i>j_p$ and $i\notin J$. Let $Z_2^q(t)$ be the submatrix of $Z_2(t)$ consisting of its first q rows, $q=1,\cdots,k$. The matrix $Z_2(t)$ maps $\mathbb{R}^n$ isomorphically onto the n-dimensional subspace of $\mathbb{R}^k$ consisting of all k-tuples which give the coordinates (with respect to the basis $\{f_1,\cdots,f_k\}$ for $\mathbb{R}^k$) for vectors belonging to $S_2(t)$. If $y\in\mathbb{R}^n$, then $Z_2(t)y$ represents a vector in $S_2(t)\cap W_{k-q}$ iff $y\in\ker Z_2^q(t)$. Thus, $\dim S_2(t)\cap W_{k-q}=\dim\ker Z_2^q(t)=n-\operatorname{rank} Z_2^q(t)$.

To show that $S_2(t)\in Y(\beta)$ is equivalent to showing that $\dim S_2(t)\cap W_j=\dim S_0\cap W_j$, $j=1,\cdots,k$. In turn, this is equivalent to showing that $\operatorname{rank} Z_2^q(t)=\operatorname{rank} Z_2^q(0)$ for all t. Observe that rows $j_1,\cdots,j_n$ of $Z_2(t)$ form an identity submatrix. Consequently, row operations can be performed on $Z_2(t)$ to remove every entry of the form ta_{ip} and thus obtain the matrix $Z_2(0)$. Furthermore, note that in each column, the entry "1" occurs above every entry of the form ta_{ip}. This means that the row operations used to reduce $Z_2(t)$ to $Z_2(0)$ are such that the row operations applied to the submatrix $Z_2^q(t)$ involve only rows in the submatrix. This implies that the rank of each such submatrix is preserved. Hence, $\operatorname{rank} Z_2^q(t)=\operatorname{rank} Z_2^q(0)$ for all t, $q=1,\cdots,k$. This completes the proof. □

It follows easily from Lemma 1 that stable and unstable manifolds for the ERDE always intersect transversally.

PROPOSITION 1. *Let $T(l)$ and $T(l')$ be invariant tori. Then $W^s(T(l))$ and $W^u(T(l'))$ intersect transversally.*

Proof. We have $E_1\oplus\cdots\oplus E_r=\mathbb{R}^{n+m}$. Obtain a refined direct sum decomposition $F_1\oplus\cdots\oplus F_{n+m}=\mathbb{R}^{n+m}$ by decomposing E_j into a direct sum of 2 1-dimensional subspaces whenever $\dim E_j=2$. In doing this, we preserve the ordering so that it still corresponds to increasing real part of the eigenvalues. Define complete flags $\{V_j\}_1^{n+m}$, $\{W_j\}_1^{n+m}$ by $V_j=\bigoplus_{i=1}^j F_i$, $W_j=\bigoplus_{i=1}^j F_{n+m-i+1}$. Then $\{V_j\}$ and $\{W_j\}$ refine the stable and unstable flags $\{M_j\}$ and $\{N_j\}$ respectively. By Theorem 4, $W^s(T(l))$ and $W^u(T(l'))$ are each disjoint unions of finitely many cells corresponding respectively to the complete flags $\{V_j\}$ and $\{W_j\}$. In the notation of Lemma 1, each cell for $W^s(T(l))$ is of the form $X(\alpha)$ for some α, while each cell for $W^u(T(l'))$ is of the form $Y(\beta)$ for some β. By Lemma 1, $X(\alpha)$ and $Y(\beta)$ intersect transversally. Since $X(\alpha)$ and $Y(\beta)$ are embedded submanifolds of $W^s(T(l))$ and $W^u(T(l'))$ respectively, we conclude that $W^s(T(l))$ and $W^u(T(l'))$ intersect transversally. □

The next question we consider is whether or not every equilibrium point of the ERDE is hyperbolic. Suppose that S_0 is an equilibrium point. Then S_0 is of the form $S_0=E_{j_1}\oplus\cdots\oplus E_{j_k}$ for some k and indices $j_1,\cdots,j_k$. Let $J=\{j_1,\cdots,j_k\}$ and let $W_0=\bigoplus_{j=1,j\notin J}^r E_j$. Then W_0 is an m-dimensional B-invariant subspace which is complementary to the n-dimensional B-invariant subspace S_0. Let $\{\alpha_i\}_1^n$ and $\{\beta_j\}_1^m$ denote the eigenvalues of the restrictions $B|S_0$ and $B|W_0$ respectively. A calculation of Hermann and Martin [19] shows that the eigenvalues of the linearization of the ERDE at S_0 are $\{\beta_j-\alpha_i\colon i=1,\cdots,n;\ j=1,\cdots,m\}$. It follows that if B satisfies Assumption A1, then none of these nm eigenvalues have zero real part. Thus, we have

PROPOSITION 2. *Every equilibrium point of the* ERDE *is hyperbolic.*

If $T(l)$ is a 1-dimensional invariant torus, then $T(l)$ is a periodic orbit. For the ESRDE, Hermann and Martin [19] have discussed the Poincaré map associated with such a periodic orbit. We use a similar approach here to analyze the Poincaré maps for the ERDE. It is worth noting that periodic orbits may lie on invariant tori of dimension 2 or greater. However, if this is the case, there are uncountably many periodic orbits.

Let $T(l)$ be a 1-dimensional invariant torus. Then there exist indices $j_0, \cdots, j_k$ such that the elements of $T(l)$ are the subspaces of the form $\tilde{S} \oplus E_{j_1} \oplus \cdots \oplus E_{j_k}$ where $\sum_{i=1}^{k} \dim E_{j_i} = n-1$, $\dim E_{j_0} = 2$, and $\tilde{S}$ is any 1-dimensional subspace of E_{j_0}. Let $\alpha_1 = a + ib$ and $\beta_1 = a - ib$ denote the pair of complex conjugate eigenvalues of $B|E_{j_0}$ chosen so that $b > 0$, and let $\tau = \pi/b$. Let $\alpha_2, \cdots, \alpha_n$ denote the eigenvalues of $B|E_{j_1} \oplus \cdots \oplus E_{j_k}$. Let $J = \{j_0, j_1, \cdots, j_k\}$, let $V = \bigoplus_{j=1, j \notin J}^{r} E_j$, and let $\beta_2, \cdots, \beta_m$ denote the eigenvalues of $B|V$. By making a change of basis, we may assume that $E_{j_0} = \mathrm{Sp}\{e_1, e_{n+1}\}$, $E_{j_1} \oplus \cdots \oplus E_{j_k} = \mathrm{Sp}\{e_2, \cdots, e_n\}$, and $V = \mathrm{Sp}\{e_{n+2}, \cdots, e_{n+m}\}$. We may also assume that

$$B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}$$

where $B_{11} = \mathrm{diag}\{a, D_1\}$ with D_1 $(n-1) \times (n-1)$, B_{12} is 0 except for the $(1, 1)$ entry which is equal to b, $B_{21} = -B_{12}'$, and $B_{22} = \mathrm{diag}\{a, D_2\}$ with D_2 $(m-1) \times (m-1)$. Let

$$e^{Bt} = \begin{bmatrix} C_{11}(t) & C_{12}(t) \\ C_{21}(t) & C_{22}(t) \end{bmatrix}.$$

Then $C_{11}(t) = \mathrm{diag}\{e^{at} \cos bt, e^{D_1 t}\}$, $C_{12}(t)$ is 0 except for its $(1, 1)$ entry which is $e^{at} \sin bt$, $C_{21}(t) = -C_{12}'(t)$, and $C_{22}(t) = \mathrm{diag}\{e^{at} \cos bt, e^{D_2 t}\}$.

We construct a Poincaré map at the point $S_0 = \mathrm{Sp}\{e_1, \cdots, e_n\} \in T(l)$. S_0 corresponds to the origin in the chart $Y \to \mathrm{Sp}\begin{bmatrix} I \\ Y \end{bmatrix}$ for $G^n(\mathbb{R}^{n+m})$, where $Y \in \mathbb{R}^{m \times n}$. Let W denote this chart. The τ-periodic solution $e^{Bt}(S_0)$ is contained in this chart except for those values of t for which $\cos bt = 0$. The expression for $e^{Bt}(S_0)$ in the local coordinates of this chart is $\tilde{Y}(t)$ where the $m \times n$ matrix $\tilde{Y}(t)$ is 0 except for its $(1, 1)$ entry which is $-\tan bt$.

Let U denote the subset of W consisting of the subspaces of the form $\mathrm{Sp}\begin{bmatrix} I \\ Y \end{bmatrix}$ where $y_{11} = 0$. Then U is a codimension 1 submanifold of W which intersects the periodic orbit transversally at S_0. Let $S_1 = \mathrm{Sp}\begin{bmatrix} I \\ Y \end{bmatrix} \in U$. $e^{Bt}(S_1) \in W$ iff $C_{11}(t) + C_{12}(t)Y$ is nonsingular, which is equivalent to $\cos bt$ being nonzero. If this is the case, then

$$e^{Bt}(S_1) = \mathrm{Sp}\begin{bmatrix} I \\ (C_{21}(t) + C_{22}(t)Y)(C_{11}(t) + C_{12}(t)Y)^{-1} \end{bmatrix}.$$

Then $e^{Bt} \in U$ iff the $(1, 1)$ entry of $(C_{21}(t) + C_{22}(t)Y)(C_{11}(t) + C_{12}(t)Y)^{-1}$ is 0. This is equivalent to having $\tan bt = 0$. It follows that if $t = \tau$, $e^{B\tau}(S_1) \in U$. Thus, the restriction of $e^{B\tau}$ to the submanifold U is a Poincaré map for the periodic orbit $T(l)$. Since $C_{12}(\tau) = 0$ and $C_{21}(\tau) = 0$, the Poincaré map is given in local coordinates by the map $Y \to C_{22}(\tau) Y C_{11}(\tau)^{-1}$, which is a linear map. Here Y is an $m \times n$ matrix with $y_{11} = 0$. It is easily verified that the eigenvalues of this linear map (and hence of the derivative of the Poincaré map at S_0) are as given in the following result. Note that $e^{-\alpha_1 \tau} = -e^{-a\tau}$ and $e^{\beta_1 \tau} = -e^{a\tau}$.

PROPOSITION 3. *Let $T(l)$ be a 1-dimensional invariant torus. Then (using the above notation) the $mn-1$ eigenvalues of the derivative of the associated Poincaré map are $\{e^{(\beta_j - \alpha_i)\tau}: i = 1, \cdots, n; j = 1, \cdots, m$ and $(i, j) \neq (1, 1)\}$.*

COROLLARY. *Every 1-dimensional invariant torus is a hyperbolic periodic orbit.*

Proof. It follows from Proposition 3 and Assumption A1 that none of the $mn-1$ eigenvalues of the derivative of the Poincaré map are on the unit circle. □

We can now state necessary and sufficient conditions for the ERDE to be a Morse-Smale vector field. Recall that B has p real eigenvalues and $2q$ nonreal eigenvalues.

THEOREM 8. *The* ERDE *is a Morse-Smale vector field iff either of the following two conditions is satisfied*: (i) $\min(n, m) \leqq 1$, (ii) $q \leqq 1$.

Proof. If either (i) or (ii) is satisfied, there can be no invariant tori of dimension greater than one. From Propositions 1, 2, 3, it follows that the ERDE is Morse-Smale. On the other hand, if $n \geqq 2$, $m \geqq 2$, and $q \geqq 2$, the nonwandering set of the ERDE contains at least one invariant torus of dimension greater than one and therefore cannot be Morse-Smale. □

Remark 2. In the case where B has only real eigenvalues (i.e. $q = 0$), the nonwandering set contains only equilibrium points. In this very special case, C. R. Schneider proved that the ERDE is Morse-Smale and described the phase portrait [33]. By demonstrating the existence of invariant tori of dimension greater than one, Hermann and Martin showed that the ERDE is not Morse-Smale if $\min(n, m) \geqq 2$ and $q \geqq 2$ [18]. However, Theorem 8 includes a new result, namely that the ERDE is Morse-Smale whenever there are no invariant tori of dimension greater than one.

We now consider the structural stability of the ERDE. Let ϕ_t^1 and ϕ_t^2 be two flows on a manifold M. ϕ_t^1 and ϕ_t^2 are said to be *topologically equivalent* if there exists a homeomorphism of M which sends orbits of ϕ_t^1 into orbits of ϕ_t^2 [42]. Let $\chi(M)$ denote the set of all C^r, $r > 0$, vector fields on M. $\chi(M)$ forms a vector space, and given a C^r norm, $r < \infty$, a Banach space. A vector field $X \in \chi(M)$ is said to be *structurally stable* if there is some neighborhood N of X in $\chi(M)$ such that every vector field $\tilde{X} \in N$ is topologically equivalent to X.

It is also useful to define a weaker form of equivalence. Let $\Omega(\phi_t^1)$ and $\Omega(\phi_t^2)$ denote the nonwandering sets of ϕ_t^1 and ϕ_t^2. We say [42] that ϕ_t^1 and ϕ_t^2 are *topologically equivalent on* Ω if there exists a homeomorphism of $\Omega(\phi_t^1)$ onto $\Omega(\phi_t^2)$ which maps orbits into orbits. Then $X \in \chi(M)$ is said to be *Ω-stable* if there is a neighborhood N of X in $\chi(M)$ such that every vector field $\tilde{X} \in N$ is topologically equivalent on Ω to X.

We recall that each ERDE corresponds to an infinitesimal generator $B \in \mathrm{gl}(n+m, \mathbb{R})$. ($\mathrm{gl}(n+m, \mathbb{R})$ denotes the vector space of all $(n+m) \times (n+m)$ real matrices, which is the Lie algebra of the general linear group $\mathrm{Gl}(n+m, \mathbb{R})$.) We will say that the ERDE determined by B is *structurally stable within the class of* ERDE's if there exists a neighborhood N of B in $\mathrm{gl}(n+m, \mathbb{R})$ such that the vector field determined by B is topologically equivalent to the vector field determined by every $\tilde{B} \in N$. By replacing "topologically equivalent" with "topologically equivalent on Ω", we obtain the definition of the ERDE determined by B being *Ω-stable within the class of* ERDE's.

THEOREM 9. *Suppose that* $B \in \mathrm{gl}(n+m, \mathbb{R})$ *satisfies Assumption* A1. *If* $\min(n, m) \leqq 1$ *or* $q \leqq 1$, *the associated* ERDE *is structurally stable. Otherwise, the associated* ERDE *is not Ω-stable within the class of* ERDE's.

Proof. If $\min(n, m) \leqq 1$ or $q \leqq 1$, then the associated ERDE is Morse-Smale by Theorem 8. This implies that it is structurally stable [30]. Now suppose that $\min(n, m) \geqq 2$ and $q \geqq 2$. Then the nonwandering set contains at least one invariant torus of dimension $k \geqq 2$. Let $a_1 \pm ib_1, \cdots, a_q \pm ib_q$ be the nonreal eigenvalues of B. Given any neighborhood N of B in $\mathrm{gl}(n+m, \mathbb{R})$, we can find $\tilde{B}$, $\hat{B} \in N$ such that the imaginary parts of the eigenvalues of $\tilde{B}$ ($\hat{B}$) are all commensurable (all noncommensurable). Then every invariant torus of $\tilde{B}$ ($\hat{B}$) of dimension at least 2 contains periodic (almost periodic) orbits. Thus, the ERDE's associated with $\tilde{B}$ and $\hat{B}$ are not topologically equivalent on Ω. Hence, the ERDE associated with B cannot be Ω-stable within the class of ERDE's. □

Remark 3. It was conjectured in the recent paper [9] that the ERDE is structurally stable provided that the generator B has distinct eigenvalues. However, Theorem 9

shows that this conjecture is false. The assumption of distinct eigenvalues does not even imply Ω-stability within the class of ERDE's.

We have considered structural stability of the phase portrait in the *global* sense. One can also consider structural stability in a local sense by investigating what happens to the phase portrait in a neighborhood of an equilibrium point when the differential equation is perturbed. Local structural stability of the Riccati equation is studied in [6].

Morse theory relates the global phase portrait of a gradient vector field to the topology of the underlying manifold. This was extended by Smale [43] to the class of Morse-Smale vector fields. Let X be a Morse-Smale vector field on a compact ν-dimensional manifold M with equilibria $x_1, \cdots, x_m$ and closed orbits $\gamma_1, \cdots, \gamma_n$. Let $W^s(x_i)$ and $W^s(\gamma_j)$ denote the stable manifolds of x_i and γ_j respectively. Define the index of x_i to be $\operatorname{Ind}(x_i) = \dim W^s(x_i)$, and define the index of γ_j to be $\operatorname{Ind}(\gamma_j) = \dim W^s(\gamma_j) - \dim \gamma_j = \dim W^s(\gamma_j) - 1$. Define the Morse series for the vector field X to be the polynomial

$$M_X(t) = \sum_i t^{\operatorname{Ind}(x_i)} + \sum_j (1+t) t^{\operatorname{Ind}(\gamma_j)}.$$

Let $b_i(M, K)$ denote the ith Betti number of M for the coefficient field K. Let $P_K(M; t)$ denote the Poincaré polynomial of M for the coefficient field K. In other words,

$$P_K(M; t) = \sum_{i=0}^{\nu} b_i(M, K) t^i.$$

Then it follows from Smale's result that each coefficient of $M_X(t)$ is greater than or equal to the corresponding coefficient of $P_{Z_2}(M; t)$. In other words, the coefficient of t^i in $M_X(t)$ is greater than or equal to $b_i(M, Z_2)$. Further generalizations of the theory of Morse index are described in [10].

As an application of Theorem 8, we will obtain a new calculation of the mod 2 Betti numbers of the Grassmann manifold based on the phase portrait of the Riccati differential equation. We will use the following result due to E. E. Floyd [13]:

THEOREM. *If a transformation of period* 2 *acts on a compact manifold M, and if F is its fixed point set, then*

$$\sum_i b_i(F, Z_2) \leqq \sum_i b_i(M, Z_2).$$

Using Floyd's theorem, we can obtain a lower bound on the sum of the mod 2 Betti numbers for the Grassmann manifold.

LEMMA 2. $\sum_i b_i(G^n(\mathbb{R}^{n+m}), Z_2) \geqq \binom{n+m}{n}$.

Proof. The proof is by induction on $n+m$. If $n+m=1$, the assertion is trivial. Suppose that the assertion holds for $G^k(\mathbb{R}^p)$ whenever $p < n+m$. Let $P = \operatorname{diag}\{d_1, \cdots, d_{n+m}\}$ with $d_1 = d_2 = \cdots = d_{n+m-1} = 1$ and $d_{n+m} = -1$. P induces a diffeomorphism of $G^n(\mathbb{R}^{n+m})$ (which we also denote by P) given by $S \to P(S)$, where $P(S)$ is the image of the subspace S under P. Let F denote the fixed point set of this period 2 transformation. F consists of all n-dimensional P-invariant subspaces. Let $V = \operatorname{Sp}\{e_1, \cdots, e_{n+m-1}\}$ and let $W = \operatorname{Sp}\{e_{n+m}\}$. Then $F = \{S_1 \oplus S_2 \colon S_1 \in G^n(V), S_2 \in G^0(W)\} \sqcup \{S_1 \oplus S_2 \colon S_1 \in G^{n-1}(V), S_2 \in G^1(W)\}$. Thus, F is isomorphic to $G^n(\mathbb{R}^{n+m-1}) \sqcup G^{n-1}(\mathbb{R}^{n+m-1})$. By the induction hypothesis,

$$\sum_i b_i(F, Z_2) \geqq \binom{n+m-1}{n} + \binom{n+m-1}{n-1} = \binom{n+m}{n}.$$

Applying Floyd's theorem, we obtain

$$\sum_i b_i(G^n(\mathbb{R}^{n+m}), Z_2) \geqq \binom{n+m}{n},$$

which completes the proof. □

THEOREM 10. *$b_s(G^n(\mathbb{R}^{n+m}), Z_2)$ is equal to the number of partitions of s into n parts of size less than or equal to m.*

Proof. Choose B to have distinct real eigenvalues. Then the ERDE has $\binom{n+m}{n}$ equilibrium points and no other invariant tori. Also, $\dim E_j = 1$ for all j, $r = n+m$, and $\dim M_j = j$ for all j. Let $l = (l_1, \cdots, l_{n+m})$ be such that $l_j = 0$ or 1 and $\sum_{j=1}^{n+m} l_j = n$. By Theorem 7, $\dim W^s(T(l)) = \sum_{j=1}^{n+m} l_j(j-1-\sum_{i=1}^{j-1} l_i)$. This expression can be simplified by letting $j_1 < j_2 < \cdots < j_n$ denote the elements of $\{j\colon l_j = 1\}$. Using the fact that $\sum_{i=1}^{j_\nu - 1} l_i = \nu - 1$, we obtain the formula

$$\dim W^s(T(l)) = \sum_{\nu=1}^{n} (j_\nu - \nu).$$

Thus, $T(l)$ is an equilibrium point of index s iff $\sum_{\nu=1}^{n} (j_\nu - \nu) = s$. This shows that $(j_1 - 1, j_2 - 2, \cdots, j_n - n)$ is a partition of s into n parts of size at most m (with $j_1 - 1 \leqq \cdots \leqq j_n - n$). On the other hand, suppose that $(c_1, \cdots, c_n)$ is a partition of s into n parts of size at most m with $c_1 \leqq \cdots \leqq c_n$. Define $j_\nu = c_\nu + \nu$, $\nu = 1, \cdots, n$. Then $1 \leqq j_1 < \cdots < j_n \leqq n+m$. This means that there is a unique choice of l for which $j_1, \cdots, j_n$ are the elements of $\{j\colon l_j = 1\}$. Furthermore, the index of the corresponding equilibrium point is s. Consequently, the number of equilibrium points of index s is equal to the number of partitions of s into n parts of size at most m. Let m_s denote this number. Thus, m_s is the coefficient of t^s in the Morse series for the vector field we are considering. The sum $\sum_s m_s$ must equal the number of equilibrium points $\binom{n+m}{n}$. By Theorem 8, the vector field is Morse-Smale. Hence, $m_s \geqq b_s(G^n(\mathbb{R}^{n+m}), Z_2)$ for all s. Applying Lemma 2, we have

$$\binom{n+m}{n} = \sum_s m_s \geqq \sum_s b_s(G^n(\mathbb{R}^{n+m}), Z_2) \geqq \binom{n+m}{n},$$

which implies that $b_s(G^n(\mathbb{R}^{n+m}), Z_2) = m_s$ for all s, completing the proof. □

As we have seen, the ERDE is not generally Morse-Smale, due to the existence of invariant tori of dimension greater than one in the nonwandering set. However, we propose to define a Morse series $M_B(t)$ for the ERDE with generator B as follows: For each invariant torus $T(l)$, define the index of $T(l)$ to be $\operatorname{Ind}(T(l)) = \dim W^s(T(l)) - \dim T(l)$. Then we define the Morse series to be

$$M_B(t) = \sum_l (1+t)^{\dim T(l)} t^{\operatorname{Ind}(T(l))}.$$

Note that if every invariant torus has dimension less than 2, this definition coincides with the definition of the Morse series for a Morse-Smale vector field. Also note that the factor $(1+t)^{\dim T(l)}$ is precisely the Poincaré polynomial for the invariant torus $T(l)$. Thus, our definition of $M_B(t)$ is analogous to the definition of the Morse series for the gradient vector field corresponding to a Morse-Bott function, i.e. a function with nondegenerate critical submanifolds.

Having defined a Morse series for the ERDE, it is natural to ask whether Morse-type inequalities hold. The next result shows that this is indeed the case. In fact, if Z_2 is the coefficient field, the ERDE satisfies Morse-type *equalities.*

THEOREM 11. *Suppose that* $B \in \mathrm{gl}(n+m, \mathbb{R})$ *satisfies Assumption* A1. *Then*

$$M_B(t) = P_{Z_2}(G^n(\mathbb{R}^{n+m}); t).$$

Proof. In § 3.2, we discussed how the choice of a complete flag of subspaces gives a cell decomposition of $G^n(\mathbb{R}^{n+m})$ into the union of $\binom{n+m}{n}$ cells $\{U(a)\}$. If $j_1 < j_2 < \cdots < j_n$ are the n elements of the set $\{j: a_j = 1\}$, then we noted that $\dim U(a) = \sum_{i=1}^{n} (j_i - i)$. By an argument analogous to the one used in the proof of Theorem 10, it follows that the number of cells of dimension s is equal to the number of partitions of s into n parts of size less than or equal to m. Thus, the number of cells of dimension s is equal to $b_s(G^n(\mathbb{R}^{n+m}), Z_2)$. Hence,

$$P_{Z_2}(G^n(\mathbb{R}^{n+m}); t) = \sum_{\substack{\text{all} \\ \text{cells}}} t^{\dim \text{ cell}}.$$

In the special case where the complete flag is chosen so as to refine the stable flag $M_0 \subset M_1 \subset \cdots \subset M_r$, we know from Theorem 4 that each stable manifold $W^s(T(l))$ is a disjoint union of cells. Furthermore, if $\dim T(l) = k$, then $W^s(T(l))$ is the union of 2^k cells, and $\binom{k}{\nu}$ of these cells have dimension equal to $\dim W^s(T(l)) - k + \nu = \mathrm{Ind}\, T(l) + \nu$, $\nu = 0, \cdots, k$. The Binomial Theorem then implies that

$$(1+t)^k t^{\mathrm{Ind}\,(T(l))} = \sum_{\substack{\text{all cells} \\ \text{in } W^s(T(l))}} t^{\dim \text{ cell}}.$$

If we sum this equation over all the invariant tori $T(l)$, the left-hand side gives the Morse series $M_B(t)$ while the right-hand side gives $P_{Z_2}(G^n(\mathbb{R}^{n+m}); t)$. □

4. Phase portrait of the extended symplectic Riccati differential equation. In this section, we determine the complete phase portrait for the ESRDE on the Lagrange-Grassmann manifold $\mathscr{L}(n)$. We make the following assumptions which we denote collectively as Assumption A2: (1) the $2n$ eigenvalues of the infinitesimal generator $H = \begin{bmatrix} A & -L \\ -Q & -A' \end{bmatrix}$ are distinct, (2) if λ_i and λ_j are a pair of eigenvalues with the same real part, then $\lambda_i = \bar{\lambda}_j$, and (3) H has no eigenvalues on the imaginary axis. Note that we place no individual restrictions on the matrices A, L, Q other than requiring that L and Q be symmetric. Assumptions (1) and (2) are relaxed considerably in § 6.

As discussed in § 2, the ESRDE is the differential equation on $\mathscr{L}(n)$ whose flow is given by $S(t, S_0) = e^{Ht}(S_0)$. However, it will be convenient to extend the ESRDE to a differential equation on $G^n(\mathbb{R}^{2n})$. The flow on $G^n(\mathbb{R}^{2n})$ is given by $S(t, S_0) = e^{Ht}(S_0)$, but where we now permit the initial point S_0 to be any element of $G^n(\mathbb{R}^{2n})$. Regarded as a differential equation on $G^n(\mathbb{R}^{2n})$, the ESRDE is a special case of the ERDE, so all the results of the preceding section apply to characterize the phase portrait. Since $\mathscr{L}(n)$ is an invariant manifold of the ESRDE on $G^n(\mathbb{R}^{2n})$, the phase portrait for the ESRDE on $\mathscr{L}(n)$ is the intersection of the phase portrait on $G^n(\mathbb{R}^{2n})$ with the submanifold $\mathscr{L}(n)$. For example, the nonwandering set on $\mathscr{L}(n)$ is the intersection of $\mathscr{L}(n)$ with the nonwandering set on $G^n(\mathbb{R}^{2n})$. The corresponding statement applies to the stable and unstable manifolds. Thus, from a *set-theoretic* point of view, the phase portrait of the ESRDE on $\mathscr{L}(n)$ is immediately obtained from the results in § 3. However, these results do not furnish a *geometric* description of the phase portrait. For example, we know that the nonwandering set is the intersection with $\mathscr{L}(n)$ of finitely many tori, but we do not know the topological structure of these intersections. It is the problem of describing the geometry of the phase portrait which we consider in this section.

Since the Hamiltonian matrix H belongs to the Lie algebra $\mathrm{sp}(n,\mathbb{R})$, i.e. $JH+H'J=0$—it follows that the eigenvalues of H come in pairs. If λ is an eigenvalue, then so is $-\lambda$. Let $2p$ denote the number of real eigenvalues of H, and let $4q$ denote the number of nonreal eigenvalues of H. Let $E_1,\cdots,E_r,E_{r+1},\cdots,E_{2r}$ denote the primary components of H. (Thus, when we apply results from § 3, we must use $2r$ in place of r.) We assume that the order of the indexing is determined by increasing real part of the corresponding eigenvalues. This implies that if E_j corresponds to a real eigenvalue λ, then E_{2r-j+1} corresponds to $-\lambda$. Similarly, if E_j corresponds to a conjugate pair $\lambda,\bar{\lambda}$, then E_{2r-j+1} corresponds to $-\lambda,-\bar{\lambda}$. In particular, $\dim E_j=\dim E_{2r-j+1}$. Note also that $E_1,\cdots,E_r$ correspond to eigenvalues with negative real part, while $E_{r+1},\cdots,E_{2r}$ correspond to eigenvalues with positive real part.

4.1. Nonwandering set. We begin by describing two known results. The first lemma was proved by A. C. M. Van Swieten [46] and generalizes a result of J. E. Potter [31] and K. Mårtensson [25]. (A proof is included in [35].)

LEMMA 3. *Let $\tilde{H}\in\mathrm{sp}(n,\mathbb{R})$ and let $p(s)$ denote its characteristic polynomial. Suppose that $p(s)=p_1(-s)p_2(s)$ with $p_1(-s)$ and $p_2(s)$ relatively prime. Let $S_1=\ker p_1(\tilde{H})$ and let $S_2=\ker p_2(\tilde{H})$. Then $x_2'Jx_1=0$, $\forall x_1\in S_1$, $\forall x_2\in S_2$.*

Let $L^+(H)$ denote the direct sum of the primary components of H corresponding to its eigenvalues with *negative* real part, and let $L^-(H)$ denote the direct sum of the primary components of H corresponding to its eigenvalues with *positive* real part. Since H is assumed to have no eigenvalues on the imaginary axis, $L^+(H)$ and $L^-(H)$ are each n-dimensional. It then follows from Lemma 3 that $L^+(H)\in\mathscr{L}(n)$ and $L^-(H)\in\mathscr{L}(n)$. The next result was proved by Shayman [37].

LEMMA 4. *Let $\tilde{H}\in\mathrm{sp}(n,\mathbb{R})$ and suppose that $\tilde{H}$ has no eigenvalues on the imaginary axis. Let $\tau>0$ be fixed. There exists a bijection δ of the set of n-dimensional Lagrangian $e^{\tilde{H}\tau}$-invariant subspaces onto the set of all $e^{\tilde{H}\tau}$-invariant subspaces of $L^+(\tilde{H})$, which is given by $\delta(S)=S\cap L^+(\tilde{H})$ and $\delta^{-1}(N)=N\oplus([J(N)]^\perp\cap L^-(\tilde{H}))$. Furthermore, S is $\tilde{H}$-invariant iff $\delta(S)$ is $\tilde{H}$-invariant.*

We will first characterize the nonwandering set of the ESRDE under the additional assumption that the imaginary parts of all the eigenvalues of H are commensurable. Later we will remove this assumption.

Let $l=(l_1,\cdots,l_r,l_{r+1},\cdots,l_{2r})$ be such that $\sum_{i=1}^{2r}l_i=n$ and $0\leqq l_i\leqq\dim E_i$, $\forall i$. The corresponding invariant torus for the ESRDE on $G^n(\mathbb{R}^{2n})$ is

$$T(l)=G^{l_1}(E_1)\times\cdots\times G^{l_r}(E_r)\times G^{l_{r+1}}(E_{r+1})\times\cdots\times G^{l_{2r}}(E_{2r}).$$

If $\dim T(l)=0$, then $T(l)$ consists of a single equilibrium point. Otherwise, the additional assumption we have made concerning the imaginary part of the eigenvalues of H implies that every motion on $T(l)$ is periodic with the same minimum period $\tau>0$.

Let $S\in T(l)$. For the moment, suppose that $\dim T(l)>0$, so that S generates a periodic motion of period $\tau>0$. Let $S=S_1\oplus\cdots\oplus S_r\oplus S_{r+1}\oplus\cdots\oplus S_{2r}$ with $S_j\in G^{l_j}(E_j)$. By Lemma 4, $S\in\mathscr{L}(n)$ iff $S_{r+1}\oplus\cdots\oplus S_{2r}=[J(S_1\oplus\cdots\oplus S_r)]^\perp\cap L^-(H)$. Since S is $e^{H\tau}$-invariant, it follows that $S_1\oplus\cdots\oplus S_r$ is $e^{H\tau}$-invariant. Since $JH+H'J=0$, it follows that $[J(S_1\oplus\cdots\oplus S_r)]^\perp$ is also $e^{H\tau}$-invariant. It is clear that $e^{H\tau}$ has the same primary components as does H, namely $E_1,\cdots,E_{2r}$. This implies that $[J(S_1\oplus\cdots\oplus S_r)]^\perp\cap L^-(H)=([J(S_1\oplus\cdots\oplus S_r)]^\perp\cap E_{r+1})\oplus\cdots\oplus([J(S_1\oplus\cdots\oplus S_r)]^\perp\cap E_{2r})$. It is an easy consequence of Lemma 3 that $J(E_i)\perp E_j$ provided that $j\neq 2r-i+1$. It follows that $[J(S_1\oplus\cdots\oplus S_r)]^\perp\cap L^-(H)=([J(S_r)]^\perp\cap E_{r+1})\oplus\cdots\oplus([J(S_1)]^\perp\cap E_{2r})$. We conclude that $S\in\mathscr{L}(n)$ iff $S_{2r-i+1}=[J(S_i)]^\perp\cap E_{2r-i+1}$, $i=1,\cdots,r$.

We claim that $\dim [J(S_i)]^\perp \cap E_{2r-i+1} = \dim E_{2r-i+1} - \dim S_i$. To show this, we note that the $e^{H\tau}$-invariance of S implies the $e^{H\tau}$-invariance of S_i which implies the $e^{H\tau}$-invariance of $[J(S_i)]^\perp$. It follows that

$$[J(S_i)]^\perp = \bigoplus_{j=1}^{2r} [J(S_i)]^\perp \cap E_j = \Bigg(\bigoplus_{\substack{j=1 \\ j \neq 2r-i+1}}^{2r} E_j \Bigg) \oplus ([J(S_i)]^\perp \cap E_{2r-i+1}).$$

Thus, $2n - \dim S_i = 2n - \dim E_{2r-i+1} + \dim [J(S_i)]^\perp \cap E_{2r-i+1}$ which establishes the claim. We conclude that $T(l) \cap \mathscr{L}(n)$ is empty unless $l_{2r-i+1} = \dim E_{2r-i+1} - l_i$, $i = 1, \cdots, r$. Since $\dim E_i = \dim E_{2r-i+1}$, this is equivalent to the condition that $l_i + l_{2r-i+1} = \dim E_i$, $i = 1, \cdots, r$. If l satisfies this constraint, then

$$T(l) \cap \mathscr{L}(n) = \{S_1 \oplus \cdots \oplus S_r \oplus ([J(S_r)]^\perp \cap E_{r+1}) \oplus \cdots \oplus ([J(S_1)]^\perp \cap E_{2r}): \\ S_i \in G^{l_i}(E_i),\ i = 1, \cdots, r\}.$$

This shows that $T(l) \cap \mathscr{L}(n)$ is isomorphic to $G^{l_1}(E_1) \times \cdots \times G^{l_r}(E_r)$ and is therefore a torus. Since $\dim E_i = \dim E_{2r-i+1}$ and $l_i + l_{2r-i+1} = \dim E_i$, $\dim T(l) \cap \mathscr{L}(n) = \frac{1}{2} \dim T(l)$.

The above conclusions were derived under the assumption that $\dim T(l) > 0$. If $\dim T(l) = 0$, then $T(l)$ consists of a single equilibrium point S. Since S is H-invariant, S is $e^{H\tau}$-invariant for any τ, and all of the preceding arguments remain valid. Thus $S \in \mathscr{L}(n)$ iff $l_i + l_{2r-i+1} = \dim E_i$, $i = 1, \cdots, r$. If this condition is satisfied, then $S = S_1 \oplus \cdots \oplus S_r \oplus ([J(S_r)]^\perp \cap E_{r+1}) \oplus \cdots \oplus ([J(S_1)]^\perp \cap E_{2r})$. If $l_i = 0$, then $S_i = 0$ and $[J(S_i)]^\perp \cap E_{2r-i+1} = E_{2r-i+1}$, while if $l_i = \dim E_i$, then $S_i = E_i$ and $[J(S_i)]^\perp \cap E_{2r-i+1} = 0$.

Now we will remove the additional assumption that the imaginary parts of all the eigenvalues of H are commensurable. We need the following lemma. Recall that the real canonical form for a semisimple (i.e. diagonalizable real matrix is a matrix which is zero except for 1×1 blocks and 2×2 blocks centered on the main diagonal. Each 2×2 block is of the form $\left[\begin{smallmatrix} a & b \\ -b & a \end{smallmatrix}\right]$ and corresponds to the pair of complex conjugate eigenvalues $a \pm ib$.

LEMMA 5. *Let $\tilde{H} \in \mathrm{sp}(n, \mathbb{R})$ and suppose that $\tilde{H}$ is semisimple and has no eigenvalues on the imaginary axis. Then there exists $P \in \mathrm{Sp}(n, \mathbb{R})$ such that $P^{-1}\tilde{H}P$ has the form $\left[\begin{smallmatrix} D & 0 \\ 0 & -D' \end{smallmatrix}\right]$, where (1) D is a semisimple matrix in real canonical form; (2) every eigenvalue of D has negative real part; (3) the blocks on the main diagonal of D are ordered by increasing real part of the corresponding eigenvalues.*

Proof. Clearly there exist $\tilde{P} \in \mathrm{Gl}(2n, \mathbb{R})$ and D as above such that

$$\tilde{H}\tilde{P} = \tilde{P} \begin{bmatrix} D & 0 \\ 0 & -D' \end{bmatrix}.$$

Let $\tilde{P} = [V, W]$ with V, W each $2n \times n$, and let $R = V'JW$. Then $\tilde{H}V = VD$ and $\tilde{H}W = -WD'$. Hence, $0 = V'(J\tilde{H} + \tilde{H}'J)W = -RD' + D'R$, showing that D' commutes with R.

Since $\mathrm{Sp}\ V = L^+(\tilde{H})$ and $\mathrm{Sp}\ W = L^-(\tilde{H})$, 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 implies that R is nonsingular. Define $P = [V, WR^{-1}]$. Then $P'JP = J$, so

$P \in \mathrm{Sp}(n, \mathbb{R})$. Also, it follows easily from the fact that D' commutes with R^{-1} that

$$\tilde{H}P = P\begin{bmatrix} D & 0 \\ 0 & -D' \end{bmatrix},$$

which completes the proof. □

We can now prove the following result.

PROPOSITION 4. *Let $T(l)$ be an invariant torus for the* ESRDE *on $G^n(\mathbb{R}^{2n})$. Then $T(l) \cap \mathscr{L}(n)$ is nonempty iff $l_i + l_{2r-i+1} = \dim E_i$, $i = 1, \cdots, r$. If l satisfies this condition, then* $T(l) \cap \mathscr{L}(n) = \{S_1 \oplus \cdots \oplus S_r \oplus ([J(S_r)]^{\perp} \cap E_{r+1}) \oplus \cdots \oplus ([J(S_1)]^{\perp} \cap E_{2r})$: $S_i \in G^{l_i}(E_i)$, $i = 1, \cdots, r\}$.

Proof. We have already proved this result in the special case where the imaginary parts of all the eigenvalues of H are commensurable. If H does not satisfy this condition, use Lemma 5 to express H in the form

$$H = P\begin{bmatrix} D & 0 \\ 0 & -D' \end{bmatrix}P^{-1}$$

with $P \in \mathrm{Sp}(n, \mathbb{R})$ and D as described in the statement of the lemma. Modify D by changing the imaginary parts of its complex eigenvalues so that they are all commensurable. Let $\tilde{D}$ denote the resulting matrix, and set

$$\tilde{H} = P\begin{bmatrix} \tilde{D} & 0 \\ 0 & -\tilde{D}' \end{bmatrix}P^{-1}.$$

Then $\tilde{H} \in \mathrm{sp}(n, \mathbb{R})$. If $\tilde{E}_1, \cdots, \tilde{E}_{2r}$ denote the primary components of $\tilde{H}$ ordered according to increasing real part of the corresponding eigenvalues, then $\tilde{E}_i = E_i$, $\forall i$. Thus, the invariant tori for $\tilde{H}$ are the same as those for H. Since we already know that the result holds for $\tilde{H}$, it must hold for H as well. □

Given any integers $l_1, \cdots, l_r$ satisfying $0 \leqq l_i \leqq \dim E_i$, there is obviously a unique choice for $l_{r+1}, \cdots, l_{2r}$ such that $l_i + l_{2r-i+1} = \dim E_i$, $i = 1, \cdots, r$. Furthermore, this choice of $l_{r+1}, \cdots, l_{2r}$ is such that the conditions $\sum_{i=1}^{2r} l_i = n$ and $0 \leqq l_i \leqq \dim E_i$ $(i = r+1, \cdots, 2r)$ are satisfied. Consequently, there is exactly one invariant torus for the ESRDE on $\mathscr{L}(n)$ for each choice of integers $l_1, \cdots, l_r$ which satisfy $0 \leqq l_i \leqq \dim E_i$, $i = 1, \cdots, r$. This implies that there are $(1 + \dim E_1) \cdots (1 + \dim E_r)$ invariant tori in all. This number is equal to $2^p 3^q$.

Suppose that l corresponds to an invariant torus $T(l) \cap \mathscr{L}(n)$. By Proposition 4, it is clear that $\dim T(l) \cap \mathscr{L}(n)$ is equal to the cardinality of the set $\{j: 1 \leqq j \leqq r, l_j = 1, \dim E_j = 2\}$. Thus, a k-dimensional invariant torus is specified by choosing $l_1, \cdots, l_r$ as follows: (1) Choose $l_{j_1} = 1, \cdots, l_{j_k} = 1$ where $j_1, \cdots, j_k$ are such that $\dim E_{j_1} = 2, \cdots, \dim E_{j_k} = 2$. (2) For $j \notin \{j_1, \cdots, j_k\}$, choose either $l_j = 0$ or $l_j = \dim E_j$. It follows that the number of k-dimensional invariant tori for the ESRDE on $\mathscr{L}(n)$ is $\binom{q}{k} 2^{p+q-k}$, $k = 0, \cdots, q$.

Suppose that $T(l) \cap \mathscr{L}(n)$ is a k-dimensional invariant torus $(k > 0)$, and let $j_1 < j_2 < \cdots < j_k$ denote the elements of $\{j: 1 \leqq j \leqq r, l_j = 1, \dim E_j = 2\}$. Let $\sigma_{j_\nu} \pm i\omega_{j_\nu}$ be the conjugate pair of eigenvalues corresponding to the 2-dimensional primary component E_{j_ν}, $\nu = 1, \cdots, k$. $T(l)$ is a $2k$-dimensional invariant torus for the ESRDE on $G^n(\mathbb{R}^{2n})$. Since the pair of complex conjugate eigenvalues corresponding to the primary component $E_{2r-j_\nu+1}$ is $-\sigma_{j_\nu} \pm i\omega_{j_\nu}$, it follows from Theorem 1 that each $S \in T(l)$ generates a periodic motion if the $2k$ numbers $\omega_{j_1}, \cdots, \omega_{j_k}, \omega_{j_1}, \cdots, \omega_{j_k}$ are commensurable, in which case the period is given by the least common multiple of $\pi/\omega_{j_1}, \cdots, \pi/\omega_{j_k}$. Otherwise each $S \in T(l)$ generates an almost periodic motion. However, this motion

is never dense in $T(l)$ since each imaginary part is repeated twice in the list $\omega_{j_1}, \cdots, \omega_{j_k}$, $\omega_{j_1}, \cdots, \omega_{j_k}$. It also follows from Theorem 1 that if $S, \tilde{S} \in T(l)$ and $e^{Ht}(\tilde{S}) \to e^{Ht}(S)$ as $t \to \infty$ or $t \to -\infty$, then $S = \tilde{S}$. We obtain the following theorem.

THEOREM 12. (a) *The nonwandering set of the* ESRDE *on* $\mathscr{L}(n)$ *is a union of invariant tori. There are exactly* $\binom{q}{k}2^{p+q-k}$ *tori of dimension* k, $k = 0, \cdots, q$. (b) *If* $T(l) \cap \mathscr{L}(n)$ *is a* 0*-dimensional invariant torus, then* $T(l) \cap \mathscr{L}(n)$ *is an equilibrium point.* (c) *If* $T(l) \cap \mathscr{L}(n)$ *is a* k*-dimensional invariant torus with* $k > 0$, *and if* $\{\omega_{j_\nu}: \nu = 1, \cdots, k\}$ *are commensurable, then each* $S \in T(l) \cap \mathscr{L}(n)$ *generates a periodic motion with period equal to the least common multiple of* $\{\pi/\omega_{j_\nu}: \nu = 1, \cdots, k\}$. *Otherwise, each* $S \in T(l) \cap \mathscr{L}(n)$ *generates an almost periodic motion which is dense in* $T(l) \cap \mathscr{L}(n)$ *iff no pair of the* ω_{j_ν} *are commensurable. In all cases, if* $S, \tilde{S} \in T(l) \cap \mathscr{L}(n)$ *with* $e^{Ht}(\tilde{S}) \to e^{Ht}(S)$ *as* $t \to \infty$ *or* $t \to -\infty$, *then* $S = \tilde{S}$.

Proof. (a) All that remains to be proven is that the $\binom{q}{k}2^{p+q-k}$ invariant tori exhaust the nonwandering set of the ESRDE on $\mathscr{L}(n)$. By Corollary 2 of Theorem 2, the nonwandering set of the ESRDE on $G^n(\mathbb{R}^{2n})$ is the union of the invariant tori $\{T(l)\}$. Since $T(l) \cap \mathscr{L}(n)$ is either empty or is one of the $\binom{q}{k}2^{p+q-k}$ invariant tori for the ESRDE on $\mathscr{L}(n)$, these tori exhaust the nonwandering set. (b) Already proven. (c) All of the assertions follow from the corresponding statements for the flow in $T(l)$ except the claim that each $S \in T(l) \cap \mathscr{L}(n)$ generates a dense trajectory iff no pair of the ω_{j_ν} $(\nu = 1, \cdots, k)$ are commensurable. $T(l) \cap \mathscr{L}(n)$ is a k-dimensional torus by the isomorphism $T(l) \cap \mathscr{L}(n) \cong G^1(E_{j_1}) \times \cdots \times G^1(E_{j_k})$. If $S \in T(l) \cap \mathscr{L}(n)$, then by Proposition 4, S has the form $S = S_1 \oplus \cdots \oplus S_r \oplus ([J(S_r)]^\perp \cap E_{r+1}) \oplus \cdots \oplus ([J(S_1)]^\perp \cap E_{2r})$. The isomorphism identifies S with $(S_{j_1}, \cdots, S_{j_k}) \in G^1(E_{j_1}) \times \cdots \times G^1(E_{j_k})$ and the motion $e^{Ht}(S)$ with the motion $(e^{Ht}(S_{j_1}), \cdots, e^{Ht}(S_{j_k}))$. The motion $e^{Ht}(S_{j_\nu})$ traverses the circle $G^1(E_{j_\nu})$ with period π/ω_{j_ν}. It follows immediately that $(e^{Ht}(S_{j_1}), \cdots, e^{Ht}(S_{j_k}))$ winds densely in $G^1(E_{j_1}) \times \cdots \times G^1(E_{j_k})$ iff no pair of the ω_{j_ν} are commensurable. By the isomorphism, the same conclusion applies to the motion $e^{Ht}(S)$ in $T(l) \cap \mathscr{L}(n)$. □

4.2. Stable and unstable manifolds. In § 3.2, we described how the choice of a complete flag of subspaces gives rise to a cell decomposition of the Grassmann manifold. In particular, *any* complete flag for $\mathbb{R}^{2n}$ gives a cell decomposition of $G^n(\mathbb{R}^{2n})$. However, if we wish to obtain a cell decomposition for the submanifold $\mathscr{L}(n)$ from the cell decomposition for $G^n(\mathbb{R}^{2n})$, we must choose the complete flag in a special way.

Let $V_1 \subset \cdots \subset V_{2n}$ be the complete flag for $\mathbb{R}^{2n}$ given by $V_j = \operatorname{Sp}\{e_1, e_2, \cdots, e_j\}$ for $j = 1, 2, \cdots, n$, and $V_j = \operatorname{Sp}\{e_1, e_2, \cdots, e_n, e_{2n}, e_{2n-1}, \cdots, e_{3n-j+1}\}$ for $j = n+1, n+2, \cdots, 2n$. We will call $\{V_j\}_1^{2n}$ the *standard symplectic flag for* $\mathbb{R}^{2n}$. Let $a = (a_1, \cdots, a_{2n})$ be such that $a_i = 0$ or 1 and $\sum_{i=1}^{2n} a_i = n$. Let $U(a) = \{S \in G^n(\mathbb{R}^{2n}): \dim S \cap V_j = \sum_{i=1}^{j} a_i,\ j = 1, \cdots, 2n\}$. Let $j_\nu = \min\{j: \sum_{i=1}^{j} a_i = \nu\}$, $\nu = 1, \cdots, n$. As noted previously, $U(a)$ is real-analytically isomorphic to Euclidean space of dimension $\sum_{j=1}^{2n} a_j(j - \sum_{i=1}^{j} a_i)$. This dimension is also given by the expression $\sum_{\nu=1}^{n}(j_\nu - \nu)$. Let N denote this dimension. Also define $d = \sum_{i=1}^{n} a_i$. Note that $j_d \leqq n$ and $j_{d+1} > n$.

Let $Z(a)$ denote the set of all $2n \times n$ rank n matrices X which have the following form: (1) Rows $j_1, j_2, \cdots, j_d, 3n - j_n + 1, 3n - j_{n-1} + 1, \cdots, 3n - j_{d+1} + 1$ form an $n \times n$ identity submatrix. (2) Let $J = \{j_1, j_2, \cdots, j_d, 3n - j_n + 1, 3n - j_{n-1} + 1, \cdots, 3n - j_{d+1} + 1\}$. If $j \notin J$ and $k \leqq d$, then x_{jk} is 0 if $j > j_k$ and is arbitrary otherwise. If $j \notin J$ and $k > d$, then x_{jk} is 0 if $n < j < 3n - j_{n+d-k+1} + 1$ and is arbitrary otherwise. Then the mapping $X \mapsto \operatorname{Sp} X$ maps $Z(a)$ isomorphically onto $U(a)$. Note that we are parametrizing the subspaces in $U(a)$ by associating each $S \in U(a)$ with the unique matrix X in "modified column echelon form" whose columns span S. By saying that X is in "modified column

echelon form," we mean that X would be in column echelon form if the order of rows $n+1, n+2, \cdots, 2n$ and of columns $d+1, d+2, \cdots, n$ were reversed. The reordering of the rows corresponds to the fact that the standard symplectic flag is derived from the ordered basis $\{e_1, e_2, \cdots, e_n, e_{2n}, e_{2n-1}, \cdots, e_{n+1}\}$ for $\mathbb{R}^{2n}$ rather than from the standard ordered basis for $\mathbb{R}^{2n}$.

Example 1. Let $n=5$, $a=(0\,1\,1\,0\,0\,1\,1\,0\,0\,1)$. Then $j_1=2, j_2=3, j_3=6, j_4=7, j_5=10$, and $d=2$. Then the matrices in $Z(a)$ are those which have the form

$$X=\left[\begin{array}{ccccc} x_{11} & x_{12} & x_{13} & x_{14} & x_{15} \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & x_{43} & x_{44} & x_{45} \\ 0 & 0 & x_{53} & x_{54} & x_{55} \\ \hline 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & x_{73} & 0 & 0 \\ 0 & 0 & x_{83} & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{array}\right].$$

In this case, $U(a)$ is isomorphic to $\mathbb{R}^{13}$.

We wish to determine the structure of the intersection of $U(a)$ with the Lagrange-Grassmann manifold $\mathscr{L}(n)$.

LEMMA 6. *$U(a)\cap\mathscr{L}(n)$ is empty unless $a_i+a_{2n-i+1}=1$, $i=1,\cdots,n$.*

Proof. Let $S\in U(a)$ and let X be the unique matrix in $Z(a)$ such that $S=\operatorname{Sp} X$. Partition X as $\left[\begin{smallmatrix}\alpha\\ \beta\end{smallmatrix}\right]$ with α, β each $n\times n$. Suppose that the condition $a_i+a_{2n-i+1}=1$ does not hold for all i. Then there exists an integer j with $1\leqq j\leqq n$ such that $a_j=1$ and $a_{2n-j+1}=1$.

Now, S is Lagrangian iff $X'JX=0$, which is equivalent to the condition that $\alpha'\beta$ be symmetric. Since $a_j=1$, the jth row of α is one of the rows from the $n\times n$ identity submatrix in X. Consequently, there is some $k\leqq d$ such that $\alpha_{jk}=1$. It then follows from the structure of X that $\alpha_{ik}=0$ for all $i>j$. Since $a_{2n-j+1}=1$, the jth row of β is also one of the rows from the $n\times n$ identity submatrix in X. Hence, there is some $\nu>d$ such that $\beta_{j\nu}=1$. It follows from the structure of X that $\beta_{i\nu}=0$ for all $i<j$. This implies that the (k,ν)th entry of $\alpha'\beta$ is 1. On the other hand, the (ν,k)th entry of $\alpha'\beta$ is 0 since $k\leqq d$ and the first d columns of β are identically 0. Hence, $\alpha'\beta$ cannot be symmetric. Thus, $S\notin\mathscr{L}(n)$. □

We now assume that $a=(a_1,\cdots,a_{2n})$ satisfies the additional condition that $a_i+a_{2n-i+1}=1$, $i=1,\cdots,n$. Let $S\in U(a)$, and let X be the unique matrix in $Z(a)$ such that $S=\operatorname{Sp} X$. We will determine the equations which the N free parameters in X must satisfy in order for S to be Lagrangian. Note that if $P\in\operatorname{Sp}(n,\mathbb{R})$, then $S\in\mathscr{L}(n)$ iff $P(S)\in\mathscr{L}(n)$.

It is trivial to verify that if θ belongs to the orthogonal group $O(n)$ then the matrix

$$P_1=\left[\begin{array}{c|c} \theta & 0 \\ \hline 0 & \theta \end{array}\right]$$

is symplectic. In particular, P_1 is symplectic if θ is a permutation matrix. It is also trivial to verify that for any j such that $1\leqq j\leqq n$, the matrix P_2^j which is obtained from I by replacing row j with row $n+j$ and replacing row $n+j$ with $-(\text{row } j)$ is symplectic. We can choose $P_1\in\operatorname{Sp}(n,\mathbb{R})$ such that the first d rows of P_1X are the first d rows of

the $n \times n$ identity matrix. Since $a_i + a_{2n-i+1} = 1$, $i = 1, \cdots, n$, it follows that the jth row of X belongs to the identity submatrix of X iff the $(n+j)$th row of X does *not* belong. Since

$$P_1 = \left[\begin{array}{c|c} \theta & 0 \\ \hline 0 & \theta \end{array}\right]$$

with θ a permutation matrix, it follows that the identity submatrix of P_1X consists of rows $1, 2, \cdots, d$, $n+d+1$, $n+d+2, \cdots, 2n$. Let $\tilde{X}$ be the matrix $P_2^{d+1}P_2^{d+2} \cdots P_2^n P_1 X$. Then the first n rows of $\tilde{X}$ form an identity submatrix. Rows $n+1, \cdots, n+d$ of $\tilde{X}$ are the nontrivial rows from among rows $n+1, \cdots, 2n$ of X. Rows $n+d+1, \cdots, 2n$ of $\tilde{X}$ are the nontrivial rows from among rows $1, \cdots, n$ of X, each multiplied by -1. Write $\tilde{X}$ in partitioned form as $\tilde{X} = [\begin{smallmatrix} I \\ Y \end{smallmatrix}]$ with Y $n \times n$. Then S is Lagrangian iff Y is symmetric.

Example 2. Let X be the matrix in Example 1. Then

$$\tilde{X} = \left[\begin{array}{ccccc} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ \hline 0 & 0 & x_{73} & 0 & 0 \\ 0 & 0 & x_{83} & 0 & 0 \\ -x_{11} & -x_{12} & -x_{13} & -x_{14} & -x_{15} \\ 0 & 0 & -x_{43} & -x_{44} & -x_{45} \\ 0 & 0 & -x_{53} & -x_{54} & -x_{55} \end{array}\right].$$

We see that if $S = \operatorname{Sp} X$ with X given by Example 1, then $S \in \mathscr{L}(n)$ iff $x_{11} = -x_{73}$, $x_{12} = -x_{83}$, $x_{14} = x_{43}$, $x_{15} = x_{53}$, $x_{45} = x_{54}$. Thus, if $n = 5$ and $a = (0\,1\,1\,0\,0\,1\,1\,0\,0\,1)$, then $U(a) \cap \mathscr{L}(n)$ is isomorphic to $\mathbb{R}^8$.

The preceding example actually includes all of the features of the general case which we now consider. It is not hard to see from the structure of X that the zero entries of Y occur symmetrically with respect to the main diagonal. Also, the number of zero entries on the main diagonal is equal to d. The total number of zero entries in Y is $n^2 - \dim U(a) = n^2 - N$. Thus, the number of parameters in Y which are above the main diagonal is $\frac{1}{2}n(n-1) - \frac{1}{2}(n^2 - N - d) = \frac{1}{2}(N + d - n)$. This is the number of independent linear constraints imposed by the requirement that Y be symmetric. It follows that $U(a) \cap \mathscr{L}(n)$ is isomorphic to Euclidean space of dimension $N - \frac{1}{2}(N + d - n) = \frac{1}{2}(N + n - d)$. This proves the next lemma.

LEMMA 7. *Suppose that $a_i + a_{2n-i+1} = 1$, $i = 1, \cdots, n$. Then $U(a) \cap \mathscr{L}(n)$ is isomorphic to Euclidean space of dimension*

$$\frac{1}{2}\left[n - d + \sum_{j=1}^{2n} a_j \left(j - \sum_{i=1}^{j} a_i\right)\right].$$

Lemmas 6 and 7 show that the intersections of the $\binom{2n}{n}$ cells $U(a)$ with $\mathscr{L}(n)$ gives a partition of $\mathscr{L}(n)$ into 2^n disjoint subsets $\{U(a) \cap \mathscr{L}(n)\colon a_i + a_{2n-i+1} = 1, i = 1, \cdots, n\}$, with each subset isomorphic to Euclidean space of some dimension.

We are now in a position to describe the geometric structure of the stable and unstable manifolds for the ESRDE on $\mathscr{L}(n)$. Let $l = (l_1, \cdots, l_{2r})$ with $l_1, \cdots, l_{2r}$ nonnegative integers satisfying $l_i + l_{2r-i+1} = \dim E_i$, $i = 1, \cdots, r$. (This condition implies

that $\sum_{i=1}^{2r} l_i = n$.) By Proposition 4, $T(l) \cap \mathscr{L}(n)$ is an invariant torus for the ESRDE on $\mathscr{L}(n)$ and $\dim T(l) \cap \mathscr{L}(n) = \frac{1}{2} \dim T(l)$. Let k denote the dimension of $T(l) \cap \mathscr{L}(n)$. Define the stable flag of subspaces $M_1 \subset M_2 \subset \cdots \subset M_{2r} = \mathbb{R}^{2n}$ by setting $M_j = \bigoplus_{i=1}^{j} E_i$, $j = 1, \cdots, 2r$. Define the unstable flag of subspaces $N_1 \subset N_2 \subset \cdots \subset N_{2r} = \mathbb{R}^{2n}$ by setting $N_j = \bigoplus_{i=1}^{j} E_{2r-i+1}$, $j = 1, \cdots, 2r$. As in the discussion preceding Theorem 4, we refine $\{M_j\}$ to a complete flag by inserting a subspace M_j' between M_{j-1} and M_j whenever $\dim E_j = 2$. Since $\dim T(l) = 2k$, the set $\{j\colon l_j = 1 \text{ and } \dim E_j = 2, j = 1, \cdots, 2r\}$ contains $2k$ elements, say $j_1, \cdots, j_{2k}$. Then $W^s(T(l))$ is the union of exactly 2^{2k} cells $U(a)$ which correspond to the complete flag. Each cell is given by fixing a vector $b = (b_1, \cdots, b_{2k})$ with $b_j = 0$ or 1, and setting $W^s(T(l), b) = \{S \in W^s(T(l))\colon \dim S \cap M_{j_\nu}' = \dim S \cap M_{j_\nu - 1} + b_\nu,\ \nu = 1, \cdots, 2k\}$. The analysis preceding Theorem 4 shows that the dimension of $W^s(T(l), b)$ is

$$\sum_{j=1}^{2r} l_j \left(\dim M_{j-1} - \sum_{i=1}^{j-1} l_i \right) + 2k - \sum_{\nu=1}^{2k} b_\nu.$$

It follows from Lemma 5 that by making a symplectic change of coordinates in $\mathbb{R}^{2n}$, we can require that the complete flag be the standard symplectic flag $V_1 \subset \cdots \subset V_{2n}$ defined at the beginning of this subsection. Using the fact that $\dim E_j = \dim E_{2r-j+1}$ and our assumption that $l_j + l_{2r-j+1} = \dim E_j$, it follows that $j_{k+1} = 2r - j_k + 1$, $j_{k+2} = 2r - j_{k-1} + 1, \cdots, j_{2k} = 2r - j_1 + 1$. In the discussion preceding Theorem 4, it is shown that each choice of the vector b uniquely determines a vector $a = (a_1, \cdots, a_{2n})$ such that $W^s(T(l), b) = U(a)$. By Lemmas 6 and 7, it follows that $U(a) \cap \mathscr{L}(n)$ is nonempty iff $a_i + a_{2n-i+1} = 1$, $i = 1, \cdots, n$. Using the facts that $\dim E_j = \dim E_{2r-j+1}$ and $l_j + l_{2r-j+1} = \dim E_j$, it is easily seen from the procedure for determining a from l and b that $a_i + a_{2n-i+1} = 1$, $i = 1, \cdots, n$ iff $b_{2k-\nu+1} = b_\nu$, $\nu = 1, \cdots, k$. Thus, $W^s(T(l), b) \cap \mathscr{L}(n)$ is nonempty iff $b_{2k-\nu+1} = b_\nu$, $\nu = 1, \cdots, k$. Hence, exactly 2^k of the 2^{2k} cells in $W^s(T(l))$ intersect $\mathscr{L}(n)$.

Suppose that $b_{2k-\nu+1} = b_\nu$, $\nu = 1, \cdots, k$, so that $W^s(T(l), b) \cap \mathscr{L}(n)$ is nonempty. Using the fact that $\sum_{i=1}^{n} a_i = \sum_{j=1}^{r} l_j$, it follows from Lemma 7 that $W^s(T(l), b) \cap \mathscr{L}(n)$ is isomorphic to Euclidean space of dimension $\frac{1}{2}(N + n - d)$, where N is the dimension of $W^s(T(l), b)$ and $d = \sum_{j=1}^{r} l_j$. Using the assumption that $b_{2k-\nu+1} = b_\nu$, $\nu = 1, \cdots, k$, the dimension is given by the formula

$$\frac{1}{2}\left\{ n - \sum_{j=1}^{r} l_j + \sum_{j=1}^{2r} l_j \left(\dim M_{j-1} - \sum_{i=1}^{j-1} l_i \right) \right\} + k - \sum_{\nu=1}^{k} b_\nu.$$

This proves part (a) of the next theorem. The proof of part (b) is completely analogous.

THEOREM 13. *Suppose that $T(l) \cap \mathscr{L}(n)$ is a k-dimensional invariant torus. Then*

(a) *The stable manifold $W^s(T(l)) \cap \mathscr{L}(n)$ is the disjoint union of 2^k cells. Exactly $\binom{k}{\nu}$ of these cells have dimension*

$$\frac{1}{2}\left\{ n - \sum_{j=1}^{r} l_j + \sum_{j=1}^{2r} l_j \left(\dim M_{j-1} - \sum_{i=1}^{j-1} l_i \right) \right\} + \nu, \qquad \nu = 0, \cdots, k.$$

(b) *The unstable manifold $W^u(T(l)) \cap \mathscr{L}(n)$ is the disjoint union of 2^k cells. Exactly $\binom{k}{\nu}$ of these cells have dimension*

$$\frac{1}{2}\left\{ n - \sum_{j=1}^{r} l_{2r-j+1} + \sum_{j=1}^{2r} l_{2r-j+1} \left(\dim N_{j-1} - \sum_{i=1}^{j-1} l_{2r-i+1} \right) \right\} + \nu,$$

$\nu = 0, \cdots, k$.

Suppose that $T(l) \cap \mathscr{L}(n)$ is a k-dimensional invariant torus, and let $S_1 \in T(l) \cap \mathscr{L}(n)$. Recall from § 3.1 that $W^s(S_1) = \Pi_+^{-1}(S_1) = \{S_0 \in G^n(\mathbb{R}^{2n})\colon e^{Ht}(S_0) \to$

$e^{Ht}(S_1)$ as $t\to\infty\}$, and $W^u(S_1)=\Pi_-^{-1}(S_1)=\{S_0\in G^n(\mathbb{R}^{2n})\colon e^{Ht}(S_0)\to e^{Ht}(S_1)$ as $t\to-\infty\}$. By Theorem 5, $W^s(S_1)$ and $W^u(S_1)$ are isomorphic to Euclidean space of dimensions $\sum_{j=1}^{2r} l_j(\dim M_{j-1}-\sum_{i=1}^{j-1} l_i)$ and $\sum_{j=1}^{2r} l_{2r-j+1}(\dim N_{j-1}-\sum_{i=1}^{j-1} l_{2r-i+1})$ respectively. The next result describes the topology of the sets $W^s(S_1)\cap\mathscr{L}(n)=\{S_0\in\mathscr{L}(n)\colon e^{Ht}(S_0)\to e^{Ht}(S_1)$ as $t\to\infty\}$ and $W^u(S_1)\cap\mathscr{L}(n)=\{S_0\in\mathscr{L}(n)\colon e^{Ht}(S_0)\to e^{Ht}(S_1)$ as $t\to-\infty\}$.

THEOREM 14. *Let $T(l)\cap\mathscr{L}(n)$ be a k-dimensional invariant torus, and let $S_1\in T(l)\cap\mathscr{L}(n)$. Then*

(a) *$W^s(S_1)\cap\mathscr{L}(n)$ is analytically isomorphic to Euclidean space of dimension $\frac{1}{2}\{n-\sum_{j=1}^{r} l_j+\sum_{j=1}^{2r} l_j(\dim M_{j-1}-\sum_{i=1}^{j-1} l_i)\}$.*

(b) *$W^u(S_1)\cap\mathscr{L}(n)$ is analytically isomorphic to Euclidean space of dimension $\frac{1}{2}\{n-\sum_{j=1}^{r} l_{2r-j+1}+\sum_{j=1}^{2r} l_{2r-j+1}(\dim N_{j-1}-\sum_{i=1}^{j-1} l_{2r-i+1})\}$.*

Proof. From Lemma 5, it follows that by making a symplectic change of coordinates in $\mathbb{R}^{2n}$ if necessary, we may assume that H has the form $\left[\begin{smallmatrix} D & 0 \\ 0 & -D \end{smallmatrix}\right]$ with D as described in the statement of that lemma. This means that the standard symplectic flag $V_1\subset\cdots\subset V_{2n}$ refines the stable flag $M_1\subset\cdots\subset M_{2r}$. It follows that the cell decomposition of $G^n(\mathbb{R}^{2n})$ determined by the standard symplectic flag decomposes $W^s(T(l))$ into a union of 2^{2k} cells $\{W^s(T(l), b)\}$.

Let $m_i=\dim M_i$ for $i=0,1,\cdots,2r$. (So $m_0=0$.) It is not hard to see that by making an additional symplectic change of coordinates given by a matrix of the form $\theta=\operatorname{diag}\{\theta_1,\cdots,\theta_r,\theta_1,\cdots,\theta_r\}$ where θ_i is a special orthogonal matrix of size equal to $\dim E_i$, we may also assume that S_1 is spanned by $\{e_{m_{i-1}+j}\colon 1\leqq i\leqq r$ such that $l_i\neq 0$; $j=1,\cdots,l_i\}\cup\{e_{3n-m_{i-1}+1-j}\colon r+1\leqq i\leqq 2r$ such that $l_i\neq 0$; $j=1,\cdots,l_i\}$. (Note that θ commutes with H, so H is unaffected by the additional change of coordinates.) It then follows that the standard symplectic flag can be obtained by using S_1 to refine the stable flag by the procedure described in the proof of Theorem 5. This implies that $W^s(S_1)=W^s(T(l), b)$ for the choice $b=(1,\cdots,1)$. By the proof of Theorem 13, $W^s(T(l), b)\cap\mathscr{L}(n)$ is isomorphic to Euclidean space of dimension $\frac{1}{2}\{n-\sum_{j=1}^{r} l_j+\sum_{j=1}^{2r} l_j(\dim M_{j-1}-\sum_{i=1}^{j-1} l_i)\}$, which completes the proof of (a).

The proof of (b) is completely analogous to the proof of (a). □

Theorem 6 shows that the ERDE has either an equilibrium point or a periodic orbit whose stable manifold is open and dense. The next result shows that the ESRDE always has an equilibrium point whose stable manifold is open and dense in the Lagrange-Grassmann manifold.

THEOREM 15. *The* ESRDE *has an equilibrium point whose stable manifold is open and dense in $\mathscr{L}(n)$, and an equilibrium point whose unstable manifold is open and dense in $\mathscr{L}(n)$.*

Proof. Let $l=(l_1,\cdots,l_{2r})$ with $l_j=0, j=1,\cdots,r$ and $l_j=\dim E_j, j=r+1,\cdots,2r$. Then $T(l)\cap\mathscr{L}(n)$ is a 0-dimensional invariant torus, and hence an equilibrium point. By an argument similar to that used in the proof of Theorem 6, $S\in W^s(T(l))$ iff $S\cap M_r=0$. (Recall that M_r is the n-dimensional subspace $E_1\oplus\cdots\oplus E_r$.) Thus, $W^s(T(l))\cap\mathscr{L}(n)=\{S\in\mathscr{L}(n)\colon S\cap M_r=0\}$, which is open and dense in $\mathscr{L}(n)$.

To obtain an equilibrium point whose region of attraction in backward time is open and dense in $\mathscr{L}(n)$, redefine l to be $(l_1,\cdots,l_{2r})$ with $l_j=\dim E_j, j=1,\cdots,r$ and $l_j=0, j=r+1,\cdots,2r$. Then $T(l)\cap\mathscr{L}(n)$ is an equilibrium point. (Recall that N_r is the n-dimensional subspace $E_{r+1}\oplus\cdots\oplus E_{2r}$.) Then $W^u(T(l))\cap\mathscr{L}(n)=\{S\in\mathscr{L}(n)\colon S\cap N_r=0\}$, which is open and dense in $\mathscr{L}(n)$. □

If $T(l)\cap\mathscr{L}(n)$ is a k-dimensional invariant torus, Theorem 13 describes the stable manifold $W^s(T(l))\cap\mathscr{L}(n)$ as the union of 2^k cells. We will now show that $W^s(T(l))\cap\mathscr{L}(n)$ is an embedded submanifold of $\mathscr{L}(n)$ and is a bundle over the torus $T(l)\cap\mathscr{L}(n)$. Rather than introducing additional and cumbersome notation, we analyze a concrete

example which illustrates all of the features of the general case. In fact, the proof of the general results follows step-by-step the analysis of the example.

Let $n=5$. $\mathscr{L}(5)$ is a 15-dimensional submanifold of $G^5(\mathbb{R}^{10})$, which is itself 25-dimensional. Let $r=3$, and suppose that $\dim E_1=2$, $\dim E_2=1$, $\dim E_3=2$, $\dim E_4=2$, $\dim E_5=1$, $\dim E_6=2$. By Lemma 5, there is no loss of generality in assuming that $E_1=\mathrm{Sp}\{e_1, e_2\}$, $E_2=\mathrm{Sp}\{e_3\}$, $E_3=\mathrm{Sp}\{e_4, e_5\}$, $E_6=\mathrm{Sp}\{e_6, e_7\}$, $E_5=\mathrm{Sp}\{e_8\}$, $E_4=\mathrm{Sp}\{e_9, e_{10}\}$, where $\{e_1, \cdots, e_{2n}\}$ is the standard basis for $\mathbb{R}^{2n}$.

Let $l=(1,0,1,1,1,1)$. Then $T(l)\cap\mathscr{L}(5)$ is a 2-dimensional invariant torus. By Theorem 13, $W^s(T(l))\cap\mathscr{L}(5)$ is the union of 4 cells of dimensions 8, 9, 9, 10. Refine the flag $M_1\subset M_2\subset M_3\subset M_4\subset M_5\subset M_6$ to the standard symplectic flag $V_1\subset\cdots\subset V_{10}$. Relative to this complete flag, $W^s(T(l))\cap\mathscr{L}(5)$ is the union of the 4 cells defined by

$$
\begin{aligned}
U_1=\{S\in\mathscr{L}(5)&: \dim S\cap V_1=1, \dim S\cap V_2=1, \dim S\cap V_3=1,\\
&\dim S\cap V_4=2, \dim S\cap V_5=2, \dim S\cap V_6=3,\\
&\dim S\cap V_7=3, \dim S\cap V_8=4, \dim S\cap V_9=5, \dim S\cap V_{10}=5\},\\
U_2=\{S\in\mathscr{L}(5)&: \dim S\cap V_1=1, \dim S\cap V_2=1, \dim S\cap V_3=1,\\
&\dim S\cap V_4=1, \dim S\cap V_5=2, \dim S\cap V_6=2,\\
&\dim S\cap V_7=3, \dim S\cap V_8=4, \dim S\cap V_9=5, \dim S\cap V_{10}=5\},\\
U_3=\{S\in\mathscr{L}(5)&: \dim S\cap V_1=0, \dim S\cap V_2=1, \dim S\cap V_3=1,\\
&\dim S\cap V_4=2, \dim S\cap V_5=2, \dim S\cap V_6=3,\\
&\dim S\cap V_7=3, \dim S\cap V_8=4, \dim S\cap V_9=4, \dim S\cap V_{10}=5\},\\
U_4=\{S\in\mathscr{L}(5)&: \dim S\cap V_1=0, \dim S\cap V_2=1, \dim S\cap V_3=1,\\
&\dim S\cap V_4=1, \dim S\cap V_5=2, \dim S\cap V_6=2,\\
&\dim S\cap V_7=3, \dim S\cap V_8=4, \dim S\cap V_9=4, \dim S\cap V_{10}=5\}.
\end{aligned}
$$

These cells are isomorphic to $\mathbb{R}^8$, $\mathbb{R}^9$, $\mathbb{R}^9$, $\mathbb{R}^{10}$ respectively. To parametrize each cell, we first parametrize the corresponding cell for $G^5(\mathbb{R}^{10})$ as in Example 1. Then we impose the linear constraints which must be satisfied in order for the subspaces to belong to $\mathscr{L}(5)$. The linear constraints are determined as in Example 2. We obtain

$$
\mathrm{Sp}\begin{bmatrix}
1 & 0 & 0 & 0 & 0\\
0 & -x_{93} & x_{23} & x_{24} & x_{25}\\
0 & -x_{94} & x_{24} & x_{34} & x_{35}\\
0 & 1 & 0 & 0 & 0\\
0 & 0 & x_{25} & x_{35} & x_{55}\\
\hline
0 & 0 & 0 & 0 & 0\\
0 & 0 & 1 & 0 & 0\\
0 & 0 & 0 & 1 & 0\\
0 & 0 & x_{93} & x_{94} & 0\\
0 & 0 & 0 & 0 & 1
\end{bmatrix}
\qquad U_1
$$

$$
\mathrm{Sp}\begin{bmatrix}
1 & 0 & 0 & 0 & 0\\
0 & -x_{103} & x_{23} & x_{24} & x_{25}\\
0 & -x_{104} & x_{24} & x_{34} & x_{35}\\
0 & -x_{105} & x_{25} & x_{35} & x_{45}\\
0 & 1 & 0 & 0 & 0\\
\hline
0 & 0 & 0 & 0 & 0\\
0 & 0 & 1 & 0 & 0\\
0 & 0 & 0 & 1 & 0\\
0 & 0 & 0 & 0 & 1\\
0 & 0 & x_{103} & x_{104} & x_{105}
\end{bmatrix}
\qquad U_2
$$

$$\mathrm{Sp}\left[\begin{array}{ccccc} -x_{73} & -x_{93} & x_{13} & x_{14} & x_{15} \\ 1 & 0 & 0 & 0 & 0 \\ 0 & -x_{94} & x_{14} & x_{34} & x_{35} \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & x_{15} & x_{35} & x_{55} \\ \hline 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & x_{73} & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & x_{93} & x_{94} & 0 \\ 0 & 0 & 0 & 0 & 1 \end{array}\right], \qquad \mathrm{Sp}\left[\begin{array}{ccccc} -x_{73} & -x_{103} & x_{13} & x_{14} & x_{15} \\ 1 & 0 & 0 & 0 & 0 \\ 0 & -x_{104} & x_{14} & x_{34} & x_{35} \\ 0 & -x_{105} & x_{15} & x_{35} & x_{45} \\ 0 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & x_{73} & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & x_{103} & x_{104} & x_{105} \end{array}\right].$$

U_3 U_4

These 4 cells can be modified in an obvious way to obtain 4 charts W_1, W_2, W_3, W_4 which cover $W^s(T(l)) \cap \mathscr{L}(5)$. They are given by

$$\mathrm{Sp}\left[\begin{array}{ccccc} 1 & 0 & 0 & 0 & 0 \\ -x_{63} & -x_{93} & x_{23} & x_{24} & x_{25} \\ 0 & -x_{94} & x_{24} & x_{34} & x_{35} \\ 0 & 1 & 0 & 0 & 0 \\ 0 & -x_{95} & x_{25} & x_{35} & x_{55} \\ \hline 0 & 0 & x_{63} & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & x_{93} & x_{94} & x_{95} \\ 0 & 0 & 0 & 0 & 1 \end{array}\right], \qquad \mathrm{Sp}\left[\begin{array}{ccccc} 1 & 0 & 0 & 0 & 0 \\ -x_{63} & -x_{103} & x_{23} & x_{24} & x_{25} \\ 0 & -x_{104} & x_{24} & x_{34} & x_{35} \\ 0 & -x_{105} & x_{25} & x_{35} & x_{45} \\ 0 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & x_{63} & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & x_{103} & x_{104} & x_{105} \end{array}\right],$$

W_1 W_2

$$\mathrm{Sp}\left[\begin{array}{ccccc} -x_{73} & -x_{93} & x_{13} & x_{14} & x_{15} \\ 1 & 0 & 0 & 0 & 0 \\ 0 & -x_{94} & x_{14} & x_{34} & x_{35} \\ 0 & 1 & 0 & 0 & 0 \\ 0 & -x_{95} & x_{15} & x_{35} & x_{55} \\ \hline 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & x_{73} & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & x_{93} & x_{94} & x_{95} \\ 0 & 0 & 0 & 0 & 1 \end{array}\right], \qquad \mathrm{Sp}\left[\begin{array}{ccccc} -x_{73} & -x_{103} & x_{13} & x_{14} & x_{15} \\ 1 & 0 & 0 & 0 & 0 \\ 0 & -x_{104} & x_{14} & x_{34} & x_{35} \\ 0 & -x_{105} & x_{15} & x_{35} & x_{45} \\ 0 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & x_{73} & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & x_{103} & x_{104} & x_{105} \end{array}\right].$$

W_3 W_4

Each of these charts is a submanifold chart relative to one of the standard charts for $\mathscr{L}(5)$. (The standard charts for the Lagrange-Grassmann manifold are described in the appendix.) Thus, $W^s(T(l)) \cap \mathscr{L}(5)$ is an embedded submanifold of $\mathscr{L}(5)$.

$T(l) \cap \mathscr{L}(5)$ is itself covered by 4 submanifold charts $\bar{W}_1$, $\bar{W}_2$, $\bar{W}_3$, $\bar{W}_4$ given by

$$\mathrm{Sp}\left[\begin{array}{ccccc} 1 & 0 & 0 & 0 & 0 \\ -x_{63} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & -x_{95} & 0 & 0 & 0 \\ \hline 0 & 0 & x_{63} & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & x_{95} \\ 0 & 0 & 0 & 0 & 1 \end{array}\right], \quad \bar{W}_1$$

$$\mathrm{Sp}\left[\begin{array}{ccccc} 1 & 0 & 0 & 0 & 0 \\ -x_{63} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & -x_{105} & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & x_{63} & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & x_{105} \end{array}\right], \quad \bar{W}_2$$

$$\mathrm{Sp}\left[\begin{array}{ccccc} -x_{73} & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & -x_{95} & 0 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & x_{73} & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & x_{95} \\ 0 & 0 & 0 & 0 & 1 \end{array}\right], \quad \bar{W}_3$$

$$\mathrm{Sp}\left[\begin{array}{ccccc} -x_{73} & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & -x_{105} & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & x_{73} & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & x_{105} \end{array}\right], \quad \bar{W}_4$$

Let Ω denote the nonwandering set of the ESRDE on $G^n(\mathbb{R}^{2n})$. Then $\Omega \cap \mathscr{L}(n)$ is the nonwandering set of the ESRDE on $\mathscr{L}(n)$. We have the mappings $\Pi_+ : G^n(\mathbb{R}^{2n}) \to \Omega$ and $\Pi_- : G^n(\mathbb{R}^{2n}) \to \Omega$ defined earlier. Let $\hat{\Pi}_+$ and $\hat{\Pi}_-$ denote the restrictions to $\mathscr{L}(n)$ of Π_+ and Π_- respectively. $\hat{\Pi}_+$ maps $W^s(T(l)) \cap \mathscr{L}(n)$ onto $T(l) \cap \mathscr{L}(n)$. We will also use $\hat{\Pi}_+$ to denote its restriction to $W^s(T(l)) \cap \mathscr{L}(n)$. If $S_1 \in T(l) \cap \mathscr{L}(n)$, then $\hat{\Pi}_+^{-1}(S_1) = W^s(S_1) \cap \mathscr{L}(n)$. Thus, the fiber $\hat{\Pi}_+^{-1}(S_1)$ is isomorphic to Euclidean space of dimension $\frac{1}{2}\{n - \sum_{j=1}^{r} l_j + \sum_{j=1}^{2r} l_j(\dim M_{j-1} - \sum_{i=1}^{j-1} l_i)\}$, which in this example is equal to 8. It is clear that $\hat{\Pi}_+^{-1}(\bar{W}_i) = W_i$, $i = 1, 2, 3, 4$. Furthermore, there is an obvious isomorphism $\gamma_i : W_i \to \bar{W}_i \times \mathbb{R}^8$ with the property that if $p_i : \bar{W}_i \times \mathbb{R}^8 \to \bar{W}_i$ is the natural projection, then $p_i \circ \gamma_i$ is the restriction of $\hat{\Pi}_+$ to W_i.

The analysis of the example can be applied essentially unchanged (but at the expense of introducing some rather cumbersome notation) to describe the structure of $W^s(T(l)) \cap \mathscr{L}(n)$ for an arbitrary invariant torus $T(l) \cap \mathscr{L}(n)$. If k is the dimension of $T(l) \cap \mathscr{L}(n)$, then by Theorem 13, $W^s(T(l)) \cap \mathscr{L}(n)$ is the union of 2^k cells, the largest of which has dimension $\frac{1}{2}\{n - \sum_{j=1}^{r} l_j + \sum_{j=1}^{2r} l_j(\dim M_{j-1} - \sum_{i=1}^{j-1} l_i)\} + k$. From these cells we obtain 2^k submanifold charts for $W^s(T(l)) \cap \mathscr{L}(n)$. Thus, $W^s(T(l)) \cap \mathscr{L}(n)$ is an embedded submanifold of $\mathscr{L}(n)$ of dimension $\frac{1}{2}\{n - \sum_{j=1}^{r} l_j + \sum_{j=1}^{2r} l_j(\dim M_{j-1} - \sum_{i=1}^{j-1} l_i)\} + k$. We have the projection $\hat{\Pi}_+ : W^s(T(l)) \cap \mathscr{L}(n) \to T(l) \cap \mathscr{L}(n)$. If $S_1 \in T(l) \cap \mathscr{L}(n)$, then $\hat{\Pi}_+^{-1}(S_1) = W^s(S_1) \cap \mathscr{L}(n)$ which is isomorphic to Euclidean space of dimension $d_s = \frac{1}{2}\{n - \sum_{j=1}^{r} l_j + \sum_{j=1}^{2r} l_j(\dim M_{j-1} - \sum_{i=1}^{j-1} l_i)\}$. Each chart W_i $(i = 1, \cdots, 2^k)$ for $W^s(T(l)) \cap \mathscr{L}(n)$ is the inverse image of a chart $\bar{W}_i$ for

$T(l) \cap \mathscr{L}(n)$. Also, there exist isomorphisms $\gamma_i: W_i \to \bar{W}_i \times \mathbb{R}^{d_s}$ such that $p_i \circ \gamma_i = \hat{\Pi}_+ | W_i$. Hence, $\hat{\Pi}_+: W^s(T(l)) \cap \mathscr{L}(n) \to T(l) \cap \mathscr{L}(n)$ is a locally trivial bundle with $\mathbb{R}^{d_s}$ as fiber. The transition functions for this bundle are invertible polynomial mappings of $\mathbb{R}^{d_s}$.

The next theorem summarizes these conclusions as well as the corresponding results for the unstable manifolds.

THEOREM 16. *Let $T(l) \cap \mathscr{L}(n)$ be a k-dimensional invariant torus. Then*

(a) *$W^s(T(l)) \cap \mathscr{L}(n)$ is an embedded submanifold of $\mathscr{L}(n)$ of dimension $k + d_s$, where $d_s = \frac{1}{2}\{n - \sum_{j=1}^{r} l_j + \sum_{j=1}^{2r} l_j(\dim M_{j-1} - \sum_{i=1}^{j-1} l_i)\}$.*

(b) *$\hat{\Pi}_+: W^s(T(l)) \cap \mathscr{L}(n) \to T(l) \cap \mathscr{L}(n)$ is a locally trivial bundle with fiber R^{d_s} and polynomial transition functions.*

(c) *$W^u(T(l)) \cap \mathscr{L}(n)$ is an embedded submanifold of $\mathscr{L}(n)$ of dimension $k + d_u$, where $d_u = \frac{1}{2}\{n - \sum_{j=1}^{r} l_{2r-j+1} + \sum_{j=1}^{2r} l_{2r-j+1}(\dim N_{j-1} - \sum_{i=1}^{j-1} l_{2r-i+1})\}$.*

(d) *$\hat{\Pi}_-: W^u(T(l)) \cap \mathscr{L}(n) \to T(l) \cap \mathscr{L}(n)$ is a locally trivial bundle with fiber $\mathbb{R}^{d_u}$ and polynomial transition functions.*

4.3. Morse theory and structural stability. Next we determine exactly when the ESRDE is a Morse-Smale vector field on $\mathscr{L}(n)$. We start with a lemma. Let $V_1 \subset \cdots \subset V_{2n}$ be the standard symplectic flag defined at the beginning of § 4.2. Define a second complete flag $W_1 \subset \cdots \subset W_{2n}$ with $W_j = \{e_{n+1}, e_{n+2}, \cdots, e_{n+j}\}$ for $j \leqq n$ and $W_j = \{e_{n+1}, e_{n+2}, \cdots, e_{2n}, e_n, e_{n-1}, \cdots, e_{2n-j+1}\}$ for $j > n$. Let $\alpha = (\alpha_1, \cdots, \alpha_{2n})$ and $\beta = (\beta_1, \cdots, \beta_{2n})$ be such that $\alpha_i = 0$ or 1, $\beta_i = 0$ or 1, $\alpha_i + \alpha_{2n-i+1} = 1$, $\beta_i + \beta_{2n-i+1} = 1$, $i = 1, \cdots, n$. It follows trivially that $\sum_{i=1}^{2n} \alpha_i = n = \sum_{i=1}^{2n} \beta_i$. Define $X(\alpha) = \{S \in G^n(\mathbb{R}^{2n})$: $\dim S \cap V_j = \sum_{i=1}^{j} \alpha_i, j = 1, \cdots, 2n\}$ and $Y(\beta) = \{S \in G^n(\mathbb{R}^{2n})$: $\dim S \cap W_j = \sum_{i=1}^{j} \beta_{2n-i+1}, j = 1, \cdots, 2n\}$.

LEMMA 8. *$X(\alpha) \cap \mathscr{L}(n)$ and $Y(\beta) \cap \mathscr{L}(n)$ intersect transversally as submanifolds of $\mathscr{L}(n)$.*

Proof. If $X(\alpha) \cap \mathscr{L}(n)$ and $Y(\beta) \cap \mathscr{L}(n)$ are disjoint, then the assertion holds trivially. So we may assume that $X(\alpha) \cap \mathscr{L}(n)$ and $Y(\beta) \cap \mathscr{L}(n)$ have nonempty intersection. Since $X(\alpha) \cap Y(\beta)$ is nonempty, it follows from Lemma 1 that $\sum_{i=1}^{j} \alpha_i \leqq \sum_{i=1}^{j} \beta_i, j = 1, \cdots, 2n$. Let $j_1 < \cdots < j_n$ be the elements of $\{j: \alpha_j = 1\}$, and let $l_1 < \cdots < l_n$ be the elements of $\{l: \beta_l = 1\}$. Since $\sum_{i=1}^{j} \alpha_i \leqq \sum_{i=1}^{j} \beta_i$ for all j, it follows that $l_p \leqq j_p$, $p = 1, \cdots, n$.

Let $S_0 \in (X(\alpha) \cap \mathscr{L}(n)) \cap (Y(\beta) \cap \mathscr{L}(n))$. We must show that $T_{S_0}(X(\alpha) \cap \mathscr{L}(n)) + T_{S_0}(Y(\beta) \cap \mathscr{L}(n)) = T_{S_0}(\mathscr{L}(n))$. The proof of this requires only slight modification of the proof of part (b) of Lemma 1. To each subspace S belonging to the cell $X(\alpha) \cap \mathscr{L}(n)$ is a $2n \times n$ matrix Z such that $\operatorname{Sp} Z = S$. If $j_d \leqq n$ and $j_{d+1} > n$ (i.e. $\sum_{i=1}^{n} \alpha_i = d$), then rows $j_1, j_2, \cdots, j_d, 3n - j_n + 1, 3n - j_{n-1} + 1, \cdots, 3n - j_{d+1} + 1$ form an $n \times n$ identity submatrix. (See e.g. the parametrization of the cells U_1, U_2, U_3, U_4 in the discussion prior to Theorem 16.) There is also a standard chart for $\mathscr{L}(n)$ whose elements are described by $2n \times n$ matrices with an identity submatrix in these same n rows. (The standard charts for $\mathscr{L}(n)$ are described in the appendix.) Let v be an arbitrary tangent vector to $\mathscr{L}(n)$ at S_0. Then there is a curve $S(t)$ in $\mathscr{L}(n)$ with $S(0) = S_0$ such that $d/dt|_{t=0} S(t) = v$. Furthermore, we can choose $S(t)$ so that it corresponds to a straight line in local coordinates. Analogous to the procedure used in the proof of Lemma 1, we construct curves $S_1(t)$ and $S_2(t)$ in $\mathscr{L}(n)$ with $S_1(0) = S_2(0) = S_0$ and such that $\dot{S}_1(0) + \dot{S}_2(0) = \dot{S}(0) = v$. By construction, it is clear that $S_1(t) \in X(\alpha) \cap \mathscr{L}(n)$ for all t. Thus, $\dot{S}_1(0) \in T_{S_0}(X(\alpha) \cap \mathscr{L}(n))$.

It remains only to show that $S_2(t) \in Y(\beta) \cap \mathscr{L}(n)$ for all t. Since $S_0 \in Y(\beta)$, it suffices to show that $\dim S_2(t) \cap W_j = \dim S_0 \cap W_j, j = 1, \cdots, 2n$. Let $Z_2(t)$ be the $2n \times n$ full rank matrix which corresponds to $S_2(t)$ via the chart described above. For each

$q \in \{1, \cdots, n\}$, let $Z_2^q(t)$ denote the submatrix of $Z_2(t)$ consisting of its first q rows. For each $q \in \{n+1, \cdots, 2n\}$, let $Z_2^q(t)$ denote the submatrix of $Z_2(t)$ consisting of the first n rows of $Z_2(t)$ together with the last $q-n$ rows of $Z_2(t)$. Now, $Z_2(t)$ maps $\mathbb{R}^n$ isomorphically onto $S_2(t)$. Furthermore, $Z_2(t)$ maps $\ker Z_2^q(t)$ onto $S_2(t) \cap W_{2n-q}$. It follows that $\dim S_2(t) \cap W_{2n-q} = \dim S_2(0) \cap W_{2n-q}$ iff $\operatorname{rank} Z_2^q(t) = \operatorname{rank} Z_2^q(0)$. It follows easily from the structure of $Z_2(t)$ that row operations can be used to transform $Z_2(t)$ to $Z_2(0)$. Furthermore, the row operations are such that row operations applied to the submatrix $Z_2^q(t)$ involve only rows in the submatrix. Thus, $\operatorname{rank} Z_2^q(t) = \operatorname{rank} Z_2^q(0)$ for each q. We conclude that $\dim S_2(t) \cap W_j = \dim S_0 \cap W_j$, $j = 1, \cdots, 2n$ which completes the proof. □

Using Lemma 8, we can show that stable and unstable manifolds for the ESRDE always intersect transversally.

PROPOSITION 5. *Let $T(l) \cap \mathscr{L}(n)$ and $T(l') \cap \mathscr{L}(n)$ be invariant tori for the* ESRDE *on $\mathscr{L}(n)$. Then $W^s(T(l)) \cap \mathscr{L}(n)$ and $W^u(T(l')) \cap \mathscr{L}(n)$ intersect transversally.*

Proof. By making a symplectic change of coordinates, if necessary, we may assume that H has the form $\left[\begin{smallmatrix} D & 0 \\ 0 & -D' \end{smallmatrix}\right]$ with D as described in the statement of Lemma 5. Then the complete flags $\{V_j\}_1^{2n}$ and $\{W_j\}_1^{2n}$ defined prior to Lemma 8 refine the stable and unstable flags $\{M_j\}_1^{2r}$ and $\{N_j\}_1^{2r}$ respectively. By Theorem 13, $W^s(T(l)) \cap \mathscr{L}(n)$ and $W^u(T(l')) \cap \mathscr{L}(n)$ are each disjoint unions of finitely many cells corresponding respectively to the complete flags $\{V_j\}_1^{2n}$ and $\{W_j\}_1^{2n}$. In the notation of Lemma 8, each cell for $W^s(T(l)) \cap \mathscr{L}(n)$ is of the form $X(\alpha) \cap \mathscr{L}(n)$ for some α, while each cell for $W^u(T(l')) \cap \mathscr{L}(n)$ is of the form $Y(\beta) \cap \mathscr{L}(n)$ for some β. By Lemma 8, $X(\alpha) \cap \mathscr{L}(n)$ and $Y(\beta) \cap \mathscr{L}(n)$ intersect transversally. Since $X(\alpha) \cap \mathscr{L}(n)$ and $Y(\beta) \cap \mathscr{L}(n)$ are embedded submanifolds of $W^s(T(l)) \cap \mathscr{L}(n)$ and $W^u(T(l')) \cap \mathscr{L}(n)$ respectively, it follows immediately that $W^s(T(l)) \cap \mathscr{L}(n)$ and $W^u(T(l')) \cap \mathscr{L}(n)$ intersect transversally. □

From Proposition 2, we know that if the ESRDE is considered as a differential equation on $G^n(\mathbb{R}^{2n})$, then every equilibrium point is hyperbolic. Since every equilibrium point of the ESRDE on $\mathscr{L}(n)$ is also an equilibrium point of the ESRDE on $G^n(\mathbb{R}^{2n})$, we have.

PROPOSITION 6. *Every equilibrium point of the* ESRDE *on $\mathscr{L}(n)$ is hyperbolic.*

In a recent paper [19], Hermann and Martin have discussed the Poincaré map associated with a 1-dimensional invariant torus (i.e. a periodic orbit) for the ESRDE on $\mathscr{L}(n)$. However, there is an error in the analysis which results in an incorrect conclusion regarding the eigenvalues of the derivative of the Poincaré map.

Let $T(l) \cap \mathscr{L}(n)$ be a 1-dimensional invariant torus. It follows from Proposition 4 that each element of $T(l) \cap \mathscr{L}(n)$ is of the form $E_{j_1} \oplus \cdots \oplus E_{j_{r-1}} \oplus \tilde{S} \oplus ([J(\tilde{S})]^{\perp} \cap E_{2r-j_0+1})$ where $j \in \{j_1, \cdots, j_{r-1}\}$ iff $2r-j+1 \notin \{j_1, \cdots, j_{r-1}\}$, j_0 and $2r-j_0+1$ do not belong to $\{j_1, \cdots, j_{r-1}\}$, $\dim E_{j_0} = 2$, and $\tilde{S}$ is any 1-dimensional subspace of E_{j_0}. Let $\lambda_1 = a + ib$ denote one of the pair of complex conjugate eigenvalues of $H|E_{j_0}$ chosen so that $b > 0$. Let $\lambda_2 = -a + ib$, and let $\tau = \pi/b$. Let $\lambda_3, \cdots, \lambda_n$ denote the eigenvalues of $H|E_{j_1} \oplus \cdots \oplus E_{j_{r-1}}$. The proof of Lemma 5 is easily modified to show that by making a symplectic change of basis, we may assume that $E_{j_0} = \operatorname{Sp}\{e_1, e_{n+2}\}$, $E_{2r-j_0+1} = \operatorname{Sp}\{e_2, e_{n+1}\}$, $E_{j_1} \oplus \cdots \oplus E_{j_r} = \operatorname{Sp}\{e_3, \cdots, e_n\}$, and $E_{2r-j_1+1} \oplus \cdots \oplus E_{2r-j_r+1} = \operatorname{Sp}\{e_{n+3}, \cdots, e_{2n}\}$. We may also assume that

$$H = \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix}$$

where $H_{11} = \operatorname{diag}\{a, -a, D_1\}$ with D_1 $(n-2) \times (n-2)$, H_{12} is 0 except for its $(1, 2)$ and

(2, 1) entries which are equal to b, $H_{21} = -H_{12}$, and $H_{22} = -H_{11}'$. Let

$$e^{Ht} = \begin{bmatrix} C_{11}(t) & C_{12}(t) \\ C_{21}(t) & C_{22}(t) \end{bmatrix}.$$

Then $C_{11}(t) = \operatorname{diag}\{e^{at}\cos bt, e^{-at}\cos bt, e^{D_1't}\}$, $C_{12}(t)$ is 0 except for its (1, 2) and (2, 1) entries which are $e^{at}\sin bt$ and $e^{-at}\sin bt$ respectively, $C_{21}(t) = -C_{12}'(t)$, and $C_{22}(t) = \operatorname{diag}\{e^{-at}\cos bt, e^{at}\cos bt, e^{-D_1't}\}$.

We construct a Poincaré map at the point $S_0 = \operatorname{Sp}\{e_1, \cdots, e_n\} \in T(l) \cap \mathscr{L}(n)$. S_0 corresponds to the origin in the chart $Y \to \operatorname{Sp}\begin{bmatrix} I \\ Y \end{bmatrix}$ for $\mathscr{L}(n)$, where $Y \in S(n)$. Let W denote this chart. The τ-periodic solution $e^{Ht}(S_0)$ is contained in this chart except for those values of t for which $\cos bt = 0$. The expression for $e^{Ht}(S_0)$ in the local coordinates of this chart is $\tilde{Y}(t)$ where the $n \times n$ symmetric matrix $\tilde{Y}(t)$ is 0 except for its (1, 2) and (2, 1) entries, which are $-\tan bt$.

Let U denote the subset of W consisting of the subspaces of the form $\operatorname{Sp}\begin{bmatrix} I \\ Y \end{bmatrix}$ where $y_{12} = y_{21} = 0$. Then U is a codimension 1 submanifold of W which intersects the periodic orbit transversally at S_0. Let $S_1 = \operatorname{Sp}\begin{bmatrix} I \\ Y \end{bmatrix} \in U$. $e^{Ht}(S_1) \in W$ iff $C_{11}(t) + C_{12}(t)Y$ is nonsingular, which is equivalent to $\cos^2 bt - y_{11}y_{22}\sin^2 bt$ being nonzero. If this is the case, then

$$e^{Ht}(S_1) = \operatorname{Sp}\begin{bmatrix} I \\ (C_{21}(t) + C_{22}(t)Y)(C_{11}(t) + C_{12}(t)Y)^{-1} \end{bmatrix}.$$

Then $e^{Ht}(S_1) \in U$ if the (1, 2) and (2, 1) entries of $(C_{21}(t) + C_{22}(t)Y)(C_{11}(t) + C_{12}(t)Y)^{-1}$ are 0. This is equivalent to having $\cos bt \sin bt(1 + y_{11}y_{22}) = 0$. It follows that if $t = \tau$, $e^{H\tau}(S_1) \in U$. Thus, the restriction of $e^{H\tau}$ to the submanifold U is a Poincaré map for the periodic orbit $T(l) \cap \mathscr{L}(n)$. Since $C_{12}(\tau) = C_{21}(\tau) = 0$, the Poincaré map is given in local coordinates by the map $Y \to C_{22}(\tau)YC_{11}(\tau)^{-1}$, which is a linear map. Here Y is a symmetric matrix with $y_{12} = y_{21} = 0$. It is easily verified that the eigenvalues of this linear map (and hence of the derivative of the Poincaré map at S_0) are as given in the following result. Note that $e^{-\lambda_1\tau} = -e^{-a\tau}$ and $e^{-\lambda_2\tau} = -e^{a\tau}$.

PROPOSITION 7. *Let $T(l) \cap \mathscr{L}(n)$ be a 1-dimensional invariant torus. Then (using the above notation) the $\frac{1}{2}n(n+1) - 1$ eigenvalues of the derivative of the associated Poincaré map are $\{e^{-(\lambda_i + \lambda_j)\tau}: 1 \leqq i \leqq j \leqq n$ and $(i, j) \neq (1, 2)\}$.*

COROLLARY. *Every 1-dimensional invariant torus for the* ESRDE *on $\mathscr{L}(n)$ is a hyperbolic periodic orbit.*

Proof. It follows from Proposition 7 and Assumption A2 that none of the $\frac{1}{2}n(n+1) - 1$ eigenvalues of the derivative of the Poincaré map are on the unit circle. □

We can now obtain necessary and sufficient conditions for the ESRDE to be a Morse-Smale vector field. Recall that H has $2p$ real eigenvalues and $4q$ nonreal eigenvalues.

THEOREM 17. *The* ESRDE *is a Morse-Smale vector field on $\mathscr{L}(n)$ iff $q \leqq 1$.*

Proof. If $q \leqq 1$, there can be no invariant tori of dimension greater than one. From Theorem 12 and Propositions 5, 6, 7, it follows that the ESRDE is Morse-Smale. On the other hand, if $q \geqq 2$, it follows from Theorem 12 that the nonwandering set of the ESRDE contains at least one invariant torus of dimension greater than one and therefore cannot be Morse-Smale. □

We now consider the structural stability of the ESRDE on $\mathscr{L}(n)$. Every ESRDE corresponds to an infinitesimal generator $H \in \operatorname{sp}(n, \mathbb{R})$. We will say that the ESRDE determined by H is *structurally stable within the class of* ESRDE's if there exists a neighborhood N of H in $\operatorname{sp}(n, \mathbb{R})$ such that the vector field determined by H is topologically equivalent to the vector field determined by every $\tilde{H} \in N$. By replacing

"topologically equivalent" with "topologically equivalent on Ω," we obtain the definition of the ESRDE determined by H being *Ω-stable within the class of* ESRDE's. To avoid any possible confusion, let us recall that except in these definitions, the symbol "Ω" is used to denote the nonwandering set of the ESRDE on $G^n(\mathbb{R}^{2n})$. The nonwandering set of the ESRDE on $\mathscr{L}(n)$ is therefore $\Omega \cap \mathscr{L}(n)$.

THEOREM 18. *Suppose that $H \in \mathrm{sp}(n, \mathbb{R})$ satisfies Assumption A2. If $q \leqq 1$, the associated* ESRDE *is a structurally stable vector field on $\mathscr{L}(n)$. Otherwise, the associated* ESRDE *is not Ω-stable within the class of* ESRDE's.

Proof. If $q \leqq 1$, then the associated ESRDE is Morse-Smale by Theorem 17. This implies that it is structurally stable [30]. On the other hand, if $q \geqq 2$, then the nonwandering set contains at least one invariant torus of dimension at least 2. Let $a_1 + ib_1, \cdots, a_q + ib_q$ be the nonreal eigenvalues of H in the first quadrant of the complex plane. Given any neighborhood N of H in $\mathrm{sp}(n, \mathbb{R})$, we can find $\tilde{H}$, $\hat{H} \in N$ such that the imaginary parts of the eigenvalues of $\tilde{H}$ ($\hat{H}$) in the first quadrant are all commensurable (all noncommensurable). Then every invariant torus of $\tilde{H}$ ($\hat{H}$) of dimension at least 2 contains periodic (almost periodic) orbits. Thus, the ESRDE's associated with $\tilde{H}$ and $\hat{H}$ are not topologically equivalent on Ω. Hence, the ESRDE associated with H cannot be Ω-stable within the class of ESRDE's. □

Remark 4. If $q = 1$, then by Theorems 17 and 18, the ESRDE is a Morse-Smale vector field on $\mathscr{L}(n)$ and is structurally stable. However, considered as a vector field on $G^n(\mathbb{R}^{2n})$, the ESRDE is neither Morse-Smale nor structurally stable. The reason for this is that Ω contains at least one 2-dimensional invariant torus, while $\Omega \cap \mathscr{L}(n)$ contains only invariant tori of dimensions less than or equal to 1. Note that by saying that $q = 1$ for the ESRDE, we mean that H has 2 pairs of complex conjugate eigenvalues, or a total of 4 nonreal eigenvalues. This is different from saying that $q = 1$ for the ERDE, which in § 3 we took to mean that B has a single pair of complex conjugate eigenvalues.

The Betti numbers of the Lagrange-Grassmann manifold $\mathscr{L}(n)$ were determined by A. Borel [4]. In the case where the coefficient field is Z_2, the Poincaré polynomial of $\mathscr{L}(n)$ was found to be $(1+t)(1+t^2)\cdots(1+t^n)$. As an application of Theorem 17, we will obtain a new calculation of this result based on the phase portrait of the ESRDE. We need the following lemma which gives a lower bound for the sum of the mod 2 Betti numbers of the Lagrange-Grassmann manifold.

LEMMA 9. $\sum_i b_i(\mathscr{L}(n), Z_2) \geqq 2^n$.

Proof. By induction on n. Let $n = 1$. $\mathscr{L}(1) = G^1(\mathbb{R}^2)$, which is the projective line. The projective line is homeomorphic to S^1, for which the sum of the mod 2 Betti numbers is 2. So the assertion holds for $\mathscr{L}(1)$.

Suppose that the assertion holds for $\mathscr{L}(n-1)$. Let P be the $2n \times 2n$ matrix $\mathrm{diag}\{d_1, \cdots, d_{2n}\}$ with $d_i = 1$ for $i = 1, \cdots, n-1, n+1, \cdots, 2n-1$ and $d_n = d_{2n} = -1$. P induces an isomorphism of $G^n(\mathbb{R}^{2n})$ onto itself by $S \to P(S)$. Since $P \in \mathrm{Sp}(n, \mathbb{R})$, it follows that P maps $\mathscr{L}(n)$ onto itself. Let $F = \{S \in G^n(\mathbb{R}^{2n}) : P(S) = S\}$, the fixed point set of the mapping P. Let $V = \mathrm{Sp}\{e_1, \cdots, e_{n-1}, e_{n+1}, \cdots, e_{2n-1}\}$ and let $W = \mathrm{Sp}\{e_n, e_{2n}\}$. V and W are the eigenspaces of P corresponding to the eigenvalues 1 and -1 respectively. Since F is the set of all n-dimensional invariant subspaces of P, it follows that F is the disjoint union $\{S_1 \oplus S_2 : S_1 \in G^n(V), S_2 \in G^0(W)\} \sqcup \{S_1 \oplus S_2 : S_1 \in G^{n-1}(V), \quad S_2 \in G^1(W)\} \sqcup \{S_1 \oplus S_2 : S_1 \in G^{n-2}(V), S_2 \in G^2(W)\}$, where $G^j(V)$ and $G^j(W)$ denote the Grassmann manifolds of all j-dimensional subspaces of V and W respectively.

Let $S \in F$. Write $S = S_1 \oplus S_2$ with $S_1 \subset V$, $S_2 \subset W$. Since $PJ = JP$, we have $J(S) = J(S_1) \oplus J(S_2)$ with $J(S_1) \subset V$, $J(S_2) \subset W$. Since $V \perp W$, $J(S) \perp S$ iff $J(S_1) \perp S_1$ and

$J(S_2)\perp S_2$. A necessary condition for $J(S_1)\perp S_1$ and $J(S_2)\perp S_2$ is that $\dim S_1 \leqq \frac{1}{2}\dim V$ and $\dim S_2 \leqq \frac{1}{2}\dim W$. Thus, if $S \in F \cap \mathscr{L}(n)$, then $\dim S \cap V = n-1$ and $\dim S \cap W = 1$. So $F \cap \mathscr{L}(n) = \{S_1 \oplus S_2 : S_1 \in G^{n-1}(V), S_2 \in G^1(W), J(S_1)\perp S_1, J(S_2)\perp S_2\}$, which shows that $F \cap \mathscr{L}(n)$ is isomorphic to $\mathscr{L}(n-1)\times\mathscr{L}(1)$. Applying the induction hypothesis, we conclude that the sum of the mod 2 Betti numbers for $F \cap \mathscr{L}(n)$ is at least $2^{n-1}\cdot 2 = 2^n$. Applying Floyd's theorem (see § 3.3) to the period 2 mapping $P: \mathscr{L}(n) \to \mathscr{L}(n)$, we conclude that the sum of the mod 2 Betti numbers for $\mathscr{L}(n)$ is at least 2^n, completing the proof. □

THEOREM 19. *The Poincaré polynomial for $\mathscr{L}(n)$ using Z_2 as the coefficient field is*

$$P_{Z_2}(\mathscr{L}(n); t) = (1+t)(1+t^2)\cdots(1+t^n).$$

Proof. Choose H to have distinct real eigenvalues. (Since $H \in \operatorname{sp}(n, \mathbb{R})$, this implies that 0 is not an eigenvalue.) Then the ESRDE has 2^n equilibrium points and no other invariant tori. Also, $\dim E_j = 1$ for all j, $r = n$, and $\dim M_j = j$ for all j. Let $l = (l_1, \cdots, l_{2n})$ be such that $l_j = 0$ or 1 and $l_j + l_{2n-j+1} = 1$, $j = 1, \cdots, n$. By Theorem 16, $\dim W^s(T(l)) \cap \mathscr{L}(n) = \frac{1}{2}\{n - d + \sum_{j=1}^{2n} l_j(j - 1 - \sum_{i=1}^{j-1} l_i)\}$, where $d = \sum_{j=1}^{n} l_j$. Let $j_1 < j_2 < \cdots < j_n$ denote the elements of $\{j: l_j = 1\}$. Using the fact that $\sum_{i=1}^{j_\nu - 1} l_i = \nu - 1$, we obtain

$$\dim W^s(T(l)) \cap \mathscr{L}(n) = \frac{1}{2}\left\{n - d + \sum_{\nu=1}^{n}(j_\nu - \nu)\right\}.$$

It is easy to see that the condition $l_j + l_{2n-j+1} = 1$ implies that

$$\sum_{\nu=1}^{d} j_\nu + \sum_{\nu=d+1}^{n}(2n+1-j_\nu) = \sum_{j=1}^{n} j = \frac{1}{2}n(n+1).$$

Using this equation to substitute for $\sum_{\nu=1}^{d} j_\nu$ and simplifying gives

$$\dim W^s(T(l)) \cap \mathscr{L}(n) = \sum_{\nu=d+1}^{n}(j_\nu - n),$$

which is the sum of an $(n-d)$-element subset of $\{1, 2, \cdots, n\}$. Letting l take on each of its 2^n possible values, $\dim W^s(T(l)) \cap \mathscr{L}(n)$ takes on the values of the sums of each of the 2^n subsets of $\{1, 2, \cdots, n\}$. Since $\dim W^s(T(l)) \cap \mathscr{L}(n)$ is the index of the equilibrium point $T(l) \cap \mathscr{L}(n)$, this implies that the Morse series for the vector field corresponding to H has factored form

$$M_H(t) = (1+t)(1+t^2)\cdots(1+t^n).$$

By Theorem 17, the ESRDE which corresponds to H is Morse-Smale. Let m_s denote the coefficient of t^s in the polynomial $M_H(t)$. Then the Morse-Smale inequalities give $m_s \geqq b_s(\mathscr{L}(n), Z_2)$. To show that this is actually an equality for each s, it suffices to show that $\sum_s m_s = \sum_s b_s(\mathscr{L}(n), Z_2)$. Since $\sum_s m_s = 2^n$, this follows immediately from Lemma 9. Thus, $m_s = b_s(\mathscr{L}(n), Z_2)$ for all s, which is equivalent to the conclusion that the Poincaré polynomial for $\mathscr{L}(n)$ using Z_2 as coefficients is $(1+t)(1+t^2)\cdots(1+t^n)$. □

We have seen that the ESRDE is not generally Morse-Smale, due to the existence of invariant tori of dimension greater than one in the nonwandering set. However, we can still define a Morse series $M_H(t)$ for the ESRDE corresponding to the generator H. For each invariant torus $T(l) \cap \mathscr{L}(n)$, define the index of $T(l) \cap \mathscr{L}(n)$ to be $\operatorname{Ind}(T(l) \cap \mathscr{L}(n)) = \dim W^s(T(l)) \cap \mathscr{L}(n) - \dim T(l) \cap \mathscr{L}(n)$. Then define the Morse series to be

$$M_H(t) = \sum_l (1+t)^{\dim T(l)\cap\mathscr{L}(n)} t^{\operatorname{Ind}(T(l)\cap\mathscr{L}(n))},$$

where the sum is over all $l=(l_1,\cdots,l_{2r})$ such that $l_i+l_{2r-i+1}=\dim E_i$, $i=1,\cdots,r$. The following result shows that the ESRDE satisfies Morse-type *equalities* provided that Z_2 is used as the coefficient field.

THEOREM 20. *Suppose that* $H\in \mathrm{sp}(n,\mathbb{R})$ *satisfies Assumption A2. Then*

$$M_H(t)=P_{Z_2}(\mathscr{L}(n);t).$$

Proof. Let $\{V_j\}_1^{2n}$ denote the standard symplectic flag for $\mathbb{R}^{2n}$ described at the beginning of § 4B. Corresponding to this complete flag is a cell decomposition of $\mathscr{L}(n)$ into the union of 2^n cells. The cells are given by $U(a)\cap\mathscr{L}(n)$ where $a=(a_1,\cdots,a_{2n})$ is such that $a_i+a_{2n-i+1}=1$, $i=1,\cdots,n$ and $U(a)=\{S\in G^n(\mathbb{R}^{2n})\colon \dim S\cap V_j=\sum_{i=1}^j a_i,\ j=1,\cdots,2n\}$. By Lemma 7, $\dim U(a)\cap\mathscr{L}(n)=\frac{1}{2}\{n-d+\sum_{j=1}^{2n}a_j(j-\sum_{i=1}^j a_i)\}$, where $d=\sum_{j=1}^n a_j$. This formula can be simplified by letting $j_1<j_2<\cdots<j_n$ denote the elements of $\{j\colon a_j=1\}$. By the same manipulations used in the proof of Theorem 19 (but with a_j in place of l_j), we obtain

$$\dim U(a)\cap\mathscr{L}(n)=\sum_{\nu=d+1}^{n}(j_\nu-n).$$

It follows that

$$\sum_{\substack{\text{all}\\ \text{cells}}} t^{\dim \text{cell}}=(1+t)(1+t^2)\cdots(1+t^n),$$

which is equal to $P_{Z_2}(\mathscr{L}(n);t)$.

By Lemma 5, there is no loss of generality in assuming that the complete flag $\{V_j\}_1^{2n}$ refines the stable flag $\{M_i\}_1^{2r}$ which corresponds to H. Then by Theorem 13, each stable manifold $W^s(T(l))\cap\mathscr{L}(n)$ is a disjoint union of some of the 2^n cells corresponding to the complete flag $\{V_j\}_1^{2n}$. If $\dim T(l)\cap\mathscr{L}(n)=k$, then $W^s(T(l))\cap\mathscr{L}(n)$ is the union of 2^k cells, and $\binom{k}{\nu}$ of these cells have dimension equal to $\dim W^s(T(l))\cap\mathscr{L}(n)-k+\nu=\mathrm{Ind}\,(T(l)\cap\mathscr{L}(n))+\nu$, $\nu=0,\cdots,k$. The Binomial Theorem then implies that

$$(1+t)^k t^{\mathrm{Ind}\,(T(l)\cap\mathscr{L}(n))}=\sum_{\substack{\text{all cells in}\\ W^s(T(l))\cap\mathscr{L}(n)}} t^{\dim\text{cell}}.$$

If we sum over all the invariant tori $T(l)\cap\mathscr{L}(n)$, the left-hand side gives $M_H(t)$ while the right-hand side gives $P_{Z_2}(\mathscr{L}(n);t)$. □

5. Phase portrait of the symplectic Riccati differential equation. In this section, we determine the phase portrait for the SRDE, i.e. for the differential equation

$$\dot K=-Q-A'K-KA+KLK$$

on the vector space $S(n)$ of real symmetric $n\times n$ matrices. As was described in § 2, there is a natural embedding of $S(n)$ in the Lagrange-Grassmann manifold $\mathscr{L}(n)$ given by $\phi\colon S(n)\to\mathscr{L}(n)$, where $\phi(K)=\mathrm{Sp}\left[\begin{smallmatrix}I\\K\end{smallmatrix}\right]$. The image of ϕ is the open and dense subset $\mathscr{L}_0(n)$ of $\mathscr{L}(n)$ consisting of those n-dimensional Lagrangian subspaces which are complementary to $\mathrm{Sp}\{e_{n+1},\cdots,e_{2n}\}$. Let $K(t,K_0)$ denote the solution of the SRDE with initial condition K_0. The solution $S(t,\phi(K_0))$ of the ESRDE on $\mathscr{L}(n)$ with initial condition $\phi(K_0)$ is given by $S(t,\phi(K_0))=e^{Ht}(\phi(K_0))$ where $H=\left[\begin{smallmatrix}A & -L\\ -Q & -A'\end{smallmatrix}\right]$. The solutions $K(t,K_0)$ and $S(t,\phi(K_0))$ are related by the equation

$$(*)\qquad \phi(K(t,K_0))=S(t,\phi(K_0))$$

which holds whenever $K(t,K_0)$ exists. Equivalently, the equation holds for the largest time-interval containing 0 for which $S(t,\phi(K_0))$ remains in the subset $\mathscr{L}_0(n)$.

It is important to note that the phase portrait of the SRDE is by no means the same as the phase portrait of the ESRDE, and by themselves the results in § 4 do not characterize the phase portrait of the SRDE. Firstly, the nonwandering set $\Omega \cap \mathscr{L}(n)$ of the ESRDE may intersect $\mathscr{L}(n) - \mathscr{L}_0(n)$. Thus, there may be equilibria, periodic solutions, and almost periodic solutions which are contained in $\mathscr{L}(n) - \mathscr{L}_0(n)$ and hence correspond to no solutions in the phase portrait of the SRDE. Also, there may be points in the nonwandering set of the ESRDE which are contained in $\mathscr{L}_0(n)$ but which generate periodic or almost periodic solutions which intersect $\mathscr{L}(n) - \mathscr{L}_0(n)$. The corresponding points in $S(n)$ are not nonwandering points since they generate solutions which escape in finite time. Thus, the first obstacle to recovering the phase portrait of the SRDE from that of the ESRDE is the possibility that $\Omega \cap \mathscr{L}(n)$ intersects $\mathscr{L}(n) - \mathscr{L}_0(n)$.

The second obstacle is that even if $\Omega \cap \mathscr{L}(n)$ is completely contained in $\mathscr{L}_0(n)$, there may be solutions of the ESRDE which cross the hypersurface $\mathscr{L}(n) - \mathscr{L}_0(n)$ in the process of converging to invariant tori in $\Omega \cap \mathscr{L}(n)$. The corresponding solutions of the SRDE do not converge to the corresponding invariant tori for the SRDE. Rather, the solutions escape in finite time.

It is clear that the phase portrait of the SRDE can be recovered from that of the ESRDE provided we can determine which nonwandering points of the ESRDE belong to $\mathscr{L}_0(n)$ and provided we can identify which solutions of the SRDE escape in either finite forward or backward time. This will be done in the next two subsections. It is interesting to note that it is at this final stage that system-theoretic concepts play a key role in the solution of the mathematical problem. The results for the ESRDE place no individual restrictions on L and Q other than the requirement that they be symmetric. In particular, they apply even when the SRDE is a linear Lyapunov differential equation ($L = 0$) or when the SRDE corresponds to a zero-sum differential game (L indefinite). However, when we seek to recover the phase portrait of the SRDE from that of the ESRDE it becomes important that the SRDE correspond to a control (or filtering) problem (L nonnegative definite) and not to a differential game. Controllability of the associated linear system also becomes important. No assumption on Q is required in order to obtain the phase portrait of the SRDE. Thus, the characterization we obtain is applicable to Riccati equations which correspond to control problems with conflicting objectives (Q indefinite). However, if we wish to obtain results concerning the asymptotic signature of solutions, then it becomes important that Q be nonnegative definite and that the associated linear system be observable.

5.1. Nonwandering set. By Theorem 12, the nonwandering set of the corresponding ESRDE is a union of invariant tori. Furthermore, there are exactly $\binom{q}{k} 2^{p+q-k}$ tori of dimension k, $k = 0, \cdots, q$. In particular, there are 2^{p+q} equilibrium points. However, it is by no means clear that this is the nonwandering set of the SRDE. The problem is that some or all of the points of a given invariant torus (for the ESRDE) may belong to $\mathscr{L}(n) - \mathscr{L}_0(n)$ and therefore not correspond to any points in the space $S(n)$ of symmetric matrices. For example, it is easy to construct a matrix H which satisfies Assumption A2 but for which $L = 0$. Then the corresponding SRDE is

$$\dot{K} = -Q - A'K - KA$$

which is a *linear* matrix equation. We can arrange to have all the eigenvalues of A belong to the open left half plane. It follows immediately that the SRDE has a unique equilibrium point. Since the corresponding ESRDE has 2^{p+q} equilibrium points, it follows that $2^{p+q} - 1$ of these must belong to $\mathscr{L}(n) - \mathscr{L}_0(n)$. We see that Theorem 12

cannot tell the whole story of the nonwandering set for the SRDE since it evidently fails to distinguish between linear and quadratic matrix differential equations. To say this another way, the extension of the Riccati equation from $S(n)$ to the compactification $\mathscr{L}(n)$ blurs the distinction between linear and quadratic equations.

The main result of this subsection is that Theorem 12 *does* describe the nonwandering set of the SRDE provided that a controllability condition is satisfied. Recall that the pair (A, B) is *controllable* if the matrix $[B, AB, \cdots, A^{n-1}B]$ has full rank. It was proven independently by Shayman [41], [35] and by Lancaster and Rodman [23] that if $L = BB'$ with (A, B) controllable, then every n-dimensional Lagrangian H-invariant subspace is complementary to $\mathrm{Sp}\{e_{n+1}, \cdots, e_{2n}\}$.

Thus, every equilibrium point for the ESRDE actually belongs to the subset $\mathscr{L}_0(n)$ and hence corresponds to an equilibrium point for the SRDE on $S(n)$. This result was generalized by Shayman [37]. It was proved that if $L = BB'$ with (A, B) controllable and if H has no eigenvalues on the imaginary axis, then for any $T > 0$, every n-dimensional Lagrangian e^{HT}-invariant subspace is complementary to $\mathrm{Sp}\{e_{n+1}, \cdots, e_{2n}\}$. Hence, every periodic orbit for the ESRDE belongs to $\mathscr{L}_0(n)$ and therefore corresponds to a periodic orbit for the SRDE on $S(n)$.

The following result generalizes the existing results to show that in the presence of controllability, every nonwandering point for the ESRDE is contained in $\mathscr{L}_0(n)$ and thus corresponds to a nonwandering point for the SRDE on $S(n)$. Recall that $L^+(H)$ ($L^-(H)$) denotes the invariant subspace associated with the eigenvalues of H with negative (positive) real part.

LEMMA 10. *Suppose that $L = BB'$, (A, B) is controllable, and H has no eigenvalues on the imaginary axis. Let $S \in \mathscr{L}(n)$ and suppose that S is of the form $S = S^+ \oplus S^-$ where $S^+ \subseteq L^+(H)$ and $S^- \subseteq L^-(H)$. Then $S \in \mathscr{L}_0(n)$.*

Proof. Since (A, B) is controllable and H has no imaginary axis eigenvalues, it is well known (see e.g. [47]) that the SRDE has a unique equilibrium point K^+ (K^-) which has the additional property that every eigenvalue of $A - BB'K^+$ ($A - BB'K^-$) is in the open left (right) half-plane. Furthermore, the difference $\Delta \equiv K^+ - K^-$ is positive definite. It is also well-known (see e.g. [35]) that $\phi(K^+) = L^+(H)$ and $\phi(K^-) = L^-(H)$.

Let k denote dim S^+. Then there exist $n \times k$ and $n \times (n-k)$ full rank matrices D^+, D^- such that

$$S^+ = \mathrm{Sp}\begin{bmatrix} D^+ \\ K^+D^+ \end{bmatrix} \quad \text{and} \quad S^- = \mathrm{Sp}\begin{bmatrix} D^- \\ K^-D^- \end{bmatrix}.$$

Since $S \in \mathscr{L}(n)$, we must have $J(S^+) \perp S^-$, which implies that $(D^-)'\Delta D^+ = 0$. Now,

$$S = \mathrm{Sp}\begin{bmatrix} D^+ & D^- \\ K^+D^+ & K^-D^- \end{bmatrix},$$

so $S \in \mathscr{L}_0(n)$ if and only if the $n \times n$ matrix $[D^+ \ D^-]$ is nonsingular. Suppose there exists $y \in \mathbb{R}^k$ and $z \in \mathbb{R}^{n-k}$ such that $[D^+ \ D^-]\begin{bmatrix} y \\ z \end{bmatrix} = 0$. Then $D^+y = -D^-z$. Premultiplying both sides by $z'(D^-)'\Delta$ gives $0 = -z'(D^-)'\Delta D^-z$. Since $\Delta > 0$, 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 H satisfies Assumption* A2 *and that $L = BB'$ with (A, B) controllable. Then the nonwandering set of the* ESRDE *is contained in $\mathscr{L}_0(n)$.*

The following theorem completely characterizes the nonwandering set of the SRDE.

THEOREM 21. *Suppose that H satisfies Assumption* A2 *and that* $L = BB'$ *with* (A, B) *a controllable pair. Then*

(a) *The nonwandering set of the* SRDE *on* $S(n)$ *is a union of invariant tori. There are exactly* $\binom{q}{k}2^{p+q-k}$ *tori of dimension* k, $k = 0, \cdots, q$.

(b) *Each invariant torus is given by* $\phi^{-1}(T(l) \cap \mathscr{L}(n))$, *where* $l = (l_1, \cdots, l_{2r})$ *is such that* $l_i + l_{2r-i+1} = \dim E_i$, $i = 1, \cdots, r$, *and* $T(l) \cap \mathscr{L}(n)$ *is as described by Proposition* 4.

(c) *If* $\phi^{-1}(T(l) \cap \mathscr{L}(n))$ *is a* 0*-dimensional invariant torus, then it is an equilibrium point.*

(d) *If* $\phi^{-1}(T(l) \cap \mathscr{L}(n))$ *is a* k*-dimensional invariant torus with* $k > 0$, *and if the associated imaginary parts* $\{\omega_{j_\nu} : \nu = 1, \cdots, k\}$ *are commensurable, then each* $K \in \phi^{-1}(T(l) \cap \mathscr{L}(n))$ *generates a periodic motion with period equal to the least common multiple of* $\{\pi/\omega_{j_\nu} : \nu = 1, \cdots, k\}$. *Otherwise, each* $K \in \phi^{-1}(T(l) \cap \mathscr{L}(n))$ *generates an almost periodic motion, which is dense in* $\phi^{-1}(T(l) \cap \mathscr{L}(n))$ *iff no pair of the* ω_{j_ν} *are commensurable. In all cases, if* $K_1, K_2 \in \phi^{-1}(T(l) \cap \mathscr{L}(n))$ *with* $K(t, K_2) \to K(t, K_1)$ *as* $t \to \infty$ *or* $t \to -\infty$, *then* $K_1 = K_2$.

Proof. It follows from (*) and Lemma 10 that ϕ is a real-analytic isomorphism of the nonwandering set of the SRDE onto the nonwandering set of the ESRDE. Since the restriction of ϕ^{-1} to the nonwandering set of the ESRDE (which is compact) is uniformly continuous, it follows from AP 5 (see Appendix B) that ϕ^{-1} takes almost periodic motions to almost periodic motions. All of the other assertions follow from the corresponding results for the ESRDE (Theorem 12). □

Remark 5. It is proven in [35] that if H has no imaginary axis eigenvalues, then at least 2^{p+q-1} of the equilibrium points of the ESRDE are contained in $\mathscr{L}(n) - \mathscr{L}_0(n)$. Since they do not belong to the image of ϕ, these equilibrium points do not correspond to equilibrium points of the SRDE on $S(n)$. Thus, controllability is necessary as well as sufficient for ϕ to be an isomorphism of the nonwandering set of the SRDE onto the nonwandering set of the ESRDE.

If C is a $p \times n$ matrix, the pair (C, A) is said to be observable if the pair (A', C') is controllable. The following result describes the signatures of the points on the invariant tori for the SRDE.

THEOREM 22. *Suppose that H satisfies Assumption* A2 *and that* $L = BB'$, $Q = C'C$ *with* (A, B) *controllable and* (C, A) *observable. Then if* $K \in \phi^{-1}(T(l) \cap \mathscr{L}(n))$, *then* K *is nonsingular and has exactly* $\sum_{i=1}^{r} l_i$ *positive eigenvalues.*

Proof. It is well known (see e.g. [48]) that if $L = BB'$, $Q = C'C$ with (A, B) controllable and (C, A) observable, then $K^+ > 0$ and $K^- < 0$. Let $k = \sum_{i=1}^{r} l_i$. From the proof of Lemma 10, there exist $n \times k$ and $n \times (n-k)$ full rank matrices D^+, D^- with $(D^-)'\Delta D^+ = 0$ such that

$$\operatorname{Sp}\begin{bmatrix} I \\ K \end{bmatrix} = \operatorname{Sp}\begin{bmatrix} D^+ & D^- \\ K^+D^+ & K^-D^- \end{bmatrix}.$$

Thus, $K = [K^+D^+ \;\; K^-D^-][D^+ \;\; D^-]^{-1}$. Let $M = \operatorname{Sp} D^+$. Then $\Delta^{-1}(M^\perp) = \operatorname{Sp} D^-$. Since $[D^+ \;\; D^-]$ is nonsingular, M and $\Delta^{-1}(M^\perp)$ are complementary subspaces in $\mathbb{R}^n$. It is easy to verify that if $x \in M$, then $x'Kx = x'K^+x$, while if $x \in \Delta^{-1}(M^\perp)$, then $x'Kx = x'K^-x$. The assertions of the theorem follow immediately from this together with the fact that $K^+ > 0$ and $K^- < 0$. □

Using Proposition 4, it is easy to actually compute the invariant tori for the SRDE on $S(n)$. The first step is to determine the primary components $E_1, \cdots, E_{2r}$ of H. Choose $l = (l_1, \cdots, l_{2r})$ such that $l_i + l_{2r-i+1} = \dim E_i$, $i = 1, \cdots, r$. Next, construct the elements $S \in T(l) \cap \mathscr{L}(n)$ by forming the direct sums $S = S_1 \oplus \cdots \oplus S_r \oplus$

$([J(S_r)]^{\perp} \cap E_{r+1}) \oplus \cdots \oplus ([J(S_1)]^{\perp} \cap E_{2r})$ where S_i is any l_i-dimensional subspace of E_i, $i = 1, \cdots, r$. Given any $S \in T(l) \cap \mathscr{L}(n)$, choose *any* basis for S and let $\begin{bmatrix} X \\ Y \end{bmatrix}$ be the $2n \times n$ matrix whose columns are this basis. (X and Y are each $n \times n$.) Since $S \in \mathscr{L}_0(n)$ by Lemma 10, X is automatically nonsingular. Then $\phi^{-1}(S)$ is the symmetric matrix YX^{-1}. By this method, we can construct each invariant torus $\phi^{-1}(T(l) \cap \mathscr{L}(n))$. This procedure is illustrated in detail in § 5.4. In the special case where $T(l) \cap \mathscr{L}(n)$ is 0-dimensional, this procedure is the well-known eigenvector method for constructing the equilibrium points of the SRDE [24], [31], [25].

5.2. Stable and unstable manifolds. In this subsection, we describe the stable and unstable manifolds for the SRDE on $S(n)$. The following lemma is an easy consequence of the results in [7] and [26]. Let K^+ and K^- denote the unique stabilizing and destabilizing equilibrium points as referred to in the proof of Lemma 10.

LEMMA 11. *Suppose that $L = BB'$ with (A, B) controllable and H has no eigenvalues on the imaginary axis. Then $K(t, K_0)$ has no finite escape time in forward time if and only if $K_0 \leqq K^+$, and has no finite escape time in backward time if and only if $K_0 \geqq K^-$.*

Suppose that H satisfies A2 and that $L = BB'$ with (A, B) controllable. Let $F_+ = \{K \in S(n): K^+ \geqq K\}$ and let $F_- = \{K \in S(n): K \geqq K^-\}$. Thus, $K(t, K_0)$ has a finite escape time in forward (backward) time iff $K_0 \notin F_+$ ($K_0 \notin F_-$). By Theorem 21, the nonwandering set of the SRDE is a union of tori, with one torus for each $l = (l_1, \cdots, l_{2r})$ which satisfies $l_i + l_{2r-i+1} = \dim E_i$, $i = 1, \cdots, r$. For each such l, let $R^s(l)$ ($R^u(l)$) denote the stable (unstable) manifold of the invariant torus $\phi^{-1}(T(l) \cap \mathscr{L}(n))$. In other words, $R^s(l) = \{K_0 \in S(n): K(t, K_0) \to \phi^{-1}(T(l) \cap \mathscr{L}(n))$ as $t \to \infty\}$ and $R^u(l) = \{K_0 \in S(n): K(t, K_0) \to \phi^{-1}(T(l) \cap \mathscr{L}(n))$ as $t \to -\infty\}$. By (*), it follows immediately that $R^s(l) = \phi^{-1}(W^s(T(l)) \cap \mathscr{L}(n)) \cap F_+$ and that $R^u(l) = \phi^{-1}(W^u(T(l)) \cap \mathscr{L}(n)) \cap F_-$.

As in § 4, we let Ω denote the nonwandering set of the ESRDE regarded as a differential equation on $G^n(\mathbb{R}^{2n})$. Then $\Omega \cap \mathscr{L}(n)$ is the nonwandering set of the ESRDE on $\mathscr{L}(n)$. We recall from § 4.2 the definition of the projections $\hat{\Pi}_+: \mathscr{L}(n) \to \Omega \cap \mathscr{L}(n)$ and $\hat{\Pi}_-: \mathscr{L}(n) \to \Omega \cap \mathscr{L}(n)$. Let η_j denote the projection onto the subspace E_j along the subspace $\bigoplus_{i=1, i \neq j}^{2r} E_i$, $j = 1, \cdots, 2r$. Then $\hat{\Pi}_+(S) = \eta_1(S \cap M_1) \oplus \cdots \oplus \eta_{2r}(S \cap M_{2r})$, and $\hat{\Pi}_-(S) = \eta_1(S \cap N_{2r}) \oplus \cdots \oplus \eta_{2r}(S \cap N_1)$. We can now prove the following result which describes the asymptotic behavior of every solution of the SRDE.

THEOREM 23. *Suppose that H satisfies Assumption* A2 *and that $L = BB'$ with (A, B) a controllable pair. Let $K_0 \in S(n)$ and let $l_i = \dim \phi(K_0) \cap M_i - \dim \phi(K_0) \cap M_{i-1}$, $i = 1, \cdots, 2r$, and let $l_i' = \dim \phi(K_0) \cap N_{2r-i+1} - \dim \phi(K_0) \cap N_{2r-i}$, $i = 1, \cdots, 2r$. Let $l = (l_1, \cdots, l_{2r})$ and let $l' = (l_1', \cdots, l_{2r}')$. Then*

(a) *If $K_0 \notin F_+$, then $K(t, K_0)$ has a finite escape time in forward time.*

(b) *If $K_0 \in F_+$, then $K_0 \in R^s(l)$. Furthermore, $K(t, K_0) \to K(t, \phi^{-1}(\hat{\Pi}_+(\phi(K_0))))$ as $t \to \infty$.*

(c) *If $K_0 \notin F_-$, then $K(t, K_0)$ has a finite escape time in backward time.*

(d) *If $K_0 \in F_-$, then $K_0 \in R^u(l')$. Furthermore, $K(t, K_0) \to K(t, \phi^{-1}(\hat{\Pi}_-(\phi(K_0))))$ as $t \to -\infty$.*

Proof. (a) and (c) follow immediately from Lemma 11. (b) From the definition of l, we have $\dim \phi(K_0) \cap M_j = \sum_{i=1}^{j} l_i$, $j = 1, \cdots, 2r$. By Theorem 2, this implies that $\phi(K_0) \in W^s(T(l)) \cap \mathscr{L}(n)$, so $K_0 \in R^s(l)$. Let $S_1 = \hat{\Pi}_+(\phi(K_0)) \in T(l) \cap \mathscr{L}(n)$. Then $\phi(K_0) \in W^s(S_1) \cap \mathscr{L}(n)$, so $\rho(S(t, \phi(K_0)), S(t, S_1)) \to 0$ as $t \to \infty$. (Recall that ρ denotes the gap metric.) Since $\mathscr{L}_0(n)$ is an open subset of the metric space $\mathscr{L}(n)$ and $T(l) \cap \mathscr{L}(n) \subset \mathscr{L}_0(n)$, there exists an open set U such that $T(l) \cap \mathscr{L}(n) \subset U \subset \bar{U} \subset \mathscr{L}_0(n)$. Since $S(t, S_1) \in T(l) \cap \mathscr{L}(n)$, it follows that $S(t, \phi(K_0)) \in \bar{U}$ for sufficiently large t. Since the restriction of ϕ^{-1} to the compact set $\bar{U}$ is uniformly continuous, the fact

that $\rho(S(t, \phi(K_0)), S(t, S_1)) \to 0$ as $t \to \infty$ implies that $\|\phi^{-1}(S(t, \phi(K_0))) - \phi^{-1}(S(t, S_1))\| \to 0$ as $t \to \infty$. Thus, $\|K(t, K_0) - K(t, \phi^{-1}(\hat{\Pi}_+(\phi(K_0))))\| \to 0$ as $t \to \infty$, which proves (b). The proof of (d) is analogous to the proof of (b). □

We recall from § 4.1 that the total number of invariant tori for the ESRDE on $\mathscr{L}(n)$ and hence for the SRDE on $S(n)$ is $\Pi_{j=1}^{r}(1 + \dim E_j) = 2^p 3^q$. Theorem 23 describes 2 partitions of $S(n)$ determined by the phase portrait of the SRDE. The first partition consists of the stable manifolds $\{R^s(l)\}$ together with the region $S(n) - F_+$ of points which generate motions which escape in finite forward time. The second partition consists of the unstable manifolds $\{R^u(l)\}$ together with the region $S(n) - F_-$ of points which generate motions which escape in finite backward time. Each partition contains $1 + 2^p 3^q$ sets.

Theorem 23 describes the asymptotic behavior of *every* solution of the SRDE. If $K(t, K_0)$ converges to an invariant torus as $t \to \infty$ (or as $t \to -\infty$), the theorem specifies not only which torus $\phi^{-1}(T(l) \cap \mathscr{L}(n))$ it converges to, but also which motion on the torus $K(t, K_0)$ approaches.

An interesting implication of Theorem 23 is that no trajectories of the SRDE approach infinity asymptotically. A trajectory either reaches infinity in finite time or converges to an invariant torus. This behavior is very different from that of a linear matrix differential equation. In the linear case, there can be solutions which grow exponentially and hence approach infinity in the limit as $t \to \infty$. It is the assumption of controllability which prohibits this behavior for the SRDE. The convergence of solutions when (A, B) is stabilizable (rather than controllable) is considered in [8].

A classical result in the theory of the Riccati equation is that $K(t, K_0) \to K^-$ as $t \to \infty$ iff $K_0 < K^+$, and $K(t, K_0) \to K^+$ as $t \to -\infty$ iff $K_0 > K^-$. (See e.g. [47].) It is easy to show how these results follow from Theorem 23. We noted earlier that $\phi(K^-)$ is the sum of the primary components of H which correspond to its right half-plane eigenvalues. In other words, $\phi(K^-) = E_{r+1} \oplus \cdots \oplus E_{2r}$. Thus, the equilibrium point $\phi(K^-)$ of the ESRDE is the invariant torus $T(l) \cap \mathscr{L}(n)$ for $l = (l_1, \cdots, l_{2r})$ with $l_i = 0$, $i = 1, \cdots, r$, and $l_i = \dim E_i$, $i = r+1, \cdots, 2r$. By Theorem 23

$$R^s(l) = \{K_0 \in F_+ : \dim \phi(K_0) \cap M_j = \sum_{i=1}^{j} l_i, j = 1, \cdots, 2r\}.$$

For the given l, the condition that

$$\dim \phi(K_0) \cap M_j = \sum_{i=1}^{j} l_i, \qquad j = 1, \cdots, 2r$$

is equivalent to the condition that $\phi(K_0) \cap M_r = 0$. Since $\phi(K^+)$ is the sum of the primary components of H which correspond to its left half-plane eigenvalues, we have $M_r = E_1 \oplus \cdots \oplus E_r = \phi(K^+)$. Since $\phi(K_0) \cap \phi(K^+) = 0$ iff $K^+ - K_0$ is nonsingular, we have $R^s(l) = \{K_0 \in S(n) : K^+ \geqq K_0$ and $\det(K^+ - K_0) \neq 0\} = \{K_0 \in S(n) : K_0 < K^+\}$. By an analogous argument, it follows from Theorem 23 that the unstable manifold of K^+ is $\{K_0 \in S(n) : K_0 > K^-\}$.

The following result describes the asymptotic signature of every solution of the SRDE which does not have a finite escape time. It follows immediately from Theorem 22 together with the fact that the set of nonsingular symmetric matrices of a given signature is open in $S(n)$.

THEOREM 24. *Suppose that H satisfies Assumption* A2 *and that $L = BB'$, $Q = C'C$ with (A, B) controllable and (C, A) observable. If $K_0 \in R^s(l)$ ($R^u(l)$), then for sufficiently large positive (negative) t, $K(t, K_0)$ is nonsingular and has exactly $\sum_{i=1}^{r} l_i$ positive eigenvalues.*

5.3. Genericity. Theorems 21 and 23 give a complete description of the phase portrait of the Riccati equation

$$(**) \qquad \dot{K} = -Q - A'K - KA + KBB'K$$

provided that Assumption A2 is satisfied and (A, B) is controllable. If we fix m and consider those Riccati equations for which B is $n \times m$, we can specify each such equation by a triple $(A, B, Q) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times m} \times S(n)$. It is clear that an open and dense subset of triples satisfy the controllability condition and conditions (1) and (2) of A2. However, the set of triples (A, B, Q) which violate condition (3) of A2 has nonempty interior in $\mathbb{R}^{n\times n} \times \mathbb{R}^{n\times m} \times S(n)$. To see this, note that since $H \in \mathrm{sp}(n, \mathbb{R})$, its spectrum is symmetrical with respect to both coordinate axes. Hence, if $\left[\begin{smallmatrix} A_0 & -B_0B_0' \\ -Q_0 & -A_0' \end{smallmatrix}\right]$ has an imaginary axis eigenvalue with multiplicity one (or more generally, with any odd multiplicity) the same is true for $\left[\begin{smallmatrix} A & -BB' \\ -Q & -A' \end{smallmatrix}\right]$ provided (A, B, Q) is sufficiently close to (A_0, B_0, Q_0). This is easily illustrated by the case where $n = m = 1$. Let $H = \left[\begin{smallmatrix} a & -b^2 \\ -q & -a \end{smallmatrix}\right]$. It is straightforward to check that H has imaginary axis eigenvalues iff $a^2 + b^2q \leqq 0$. This inequality defines a region with nonempty interior in $\mathbb{R} \times \mathbb{R} \times \mathbb{R}$.

The preceding analysis shows that although Theorems 21 and 23 apply to a very large class of Riccati equations of the form (**), they do not apply to an open and dense subset.

However, the Riccati equation which arises in optimal control and filtering problems is more specialized than (**). It is of the form

$$(***) \qquad \dot{K} = -C'C - A'K - KA + KBB'K.$$

If we fix m and p and consider those Riccati equations for which B is $n \times m$ and C is $p \times n$, we can specify each such equation by a triple $(A, B, C) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times m} \times \mathbb{R}^{p\times n}$. As before, it is clear that an open and dense subset of triples (A, B, C) satisfy the controllability condition and conditions (1) and (2) of Assumption A2. Wonham proved [49] that stabilizability of (A, B) and detectability of (C, A) imply the existence of a stabilizing solution to the algebraic Riccati equation. It is well known [21] that the existence of a stabilizing solution implies that H has no imaginary axis eigenvalues. Since stabilizability and detectability are generic properties, we conclude that the set of triples (A, B, C) which satisfy condition (3) of Assumption A2 contains a subset which is open and dense. Thus, we obtain

PROPOSITION 8. *The subset of* $\mathbb{R}^{n\times n} \times \mathbb{R}^{n\times m} \times \mathbb{R}^{p\times n}$ *consisting of those triples* (A, B, C) *for which* (A, B) *is controllable and Assumption* A2 *holds contains a subset which is open and dense in* $\mathbb{R}^{n\times n} \times \mathbb{R}^{n\times m} \times \mathbb{R}^{p\times n}$.

Thus, Theorems 21 and 23 give a complete description of the phase portrait for an open and dense subset of the Riccati equations of the form (***).

The following result is an interesting consequence of Proposition 8.

PROPOSITION 9. *Let* $H \in \mathrm{sp}(n, \mathbb{R})$ *with* $H = \left[\begin{smallmatrix} A & -L \\ -Q & -A' \end{smallmatrix}\right]$. *If* L *and* Q *are nonnegative definite, then every imaginary axis eigenvalue of* H *has even multiplicity.*

Proof. Suppose H has an imaginary axis eigenvalue of odd multiplicity. Then the same is true for every matrix $\tilde{H}$ in some neighborhood of H in $\mathrm{sp}(n, \mathbb{R})$. If L and Q are nonnegative definite, we can express L and Q as $L = BB'$, $Q = C'C$ for some B and C. By Proposition 8, arbitrarily small perturbations in A, B, and C can be chosen so that the resulting matrix $\tilde{H} \in \mathrm{sp}(n, \mathbb{R})$ has no imaginary axis eigenvalues, a contradiction. □

5.4. Example. In this subsection, we illustrate how our results are applied to a concrete example. Consider the SRDE

$$\dot{K} = -Q - A'K - KA + KLK$$

where

$$A=\frac{1}{9}\begin{bmatrix}-42 & 26 & -6 & -16\\ 6 & -70 & -30 & -16\\ -18 & 2 & -153 & 80\\ 24 & 24 & 24 & -45\end{bmatrix},\qquad L=\frac{1}{9}\begin{bmatrix}28 & -8 & 4 & 0\\ -8 & 28 & 4 & 0\\ 4 & 4 & 34 & 0\\ 0 & 0 & 0 & 18\end{bmatrix},$$

$$Q=\frac{1}{9}\begin{bmatrix}-56 & 22 & -110 & 60\\ 22 & -168 & -68 & -4\\ -110 & -68 & -704 & 420\\ 60 & -4 & 420 & -292\end{bmatrix}.$$

By direct calculation, the eigenvalues of H are determined to be $-2\pm 2i$, $-1\pm 4i$, $1\pm 4i$, $2\pm 2i$. Thus, Assumption A2 is satisfied. Furthermore, it is easily verified that $L>0$. If we set $B=L^{1/2}$, then $BB'=L$ and (A, B) is trivially controllable since B is nonsingular.

Since $p=0$ and $q=2$, we conclude from Theorem 21(a) that the nonwandering set consists of 4 0-dimensional invariant tori, 4 1-dimensional invariant tori, and 1 2-dimensional invariant torus. The 0-dimensional invariant tori are equilibrium points, and they correspond to $l=(2,2,0,0)$, $l=(2,0,2,0)$, $l=(0,2,0,2)$, and $l=(0,0,2,2)$. The 1-dimensional invariant tori are isolated periodic orbits, and they correspond to $l=(2,1,1,0)$, $l=(1,2,0,1)$, $l=(1,0,2,1)$, and $l=(0,1,1,2)$. By Theorem 21(d), they have periods $\pi/4$, $\pi/2$, $\pi/2$ and $\pi/4$ respectively. The 2-dimensional invariant torus corresponds to $l=(1,1,1,1)$. Since the imaginary parts $\{4, 2\}$ are commensurable, it follows from Theorem 21(d) that every motion on this torus is periodic with period $\pi/2$. Thus, the Riccati equation has uncountably many periodic orbits.

Using the procedure described in Lemma 5, we can find a symplectic matrix P such that

$$PHP^{-1}=\left[\begin{array}{cccc|cccc}-2 & 2 & 0 & 0 & 0 & 0 & 0 & 0\\ -2 & -2 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & -1 & 4 & 0 & 0 & 0 & 0\\ 0 & 0 & -4 & -1 & 0 & 0 & 0 & 0\\ \hline 0 & 0 & 0 & 0 & 2 & 2 & 0 & 0\\ 0 & 0 & 0 & 0 & -2 & 2 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 4\\ 0 & 0 & 0 & 0 & 0 & 0 & -4 & 1\end{array}\right].$$

One such P is given by

$$P=\frac{1}{3}\left[\begin{array}{cccc|cccc}-4 & 3 & -10 & 4 & -2 & 1 & -2 & 0\\ 2 & -6 & -10 & 4 & 1 & -2 & -2 & 0\\ -4 & -6 & 5 & -2 & -2 & -2 & 1 & 0\\ 0 & 0 & -6 & 9 & 0 & 0 & 0 & 3\\ \hline -2 & 2 & -8 & 4 & -2 & 1 & -2 & 0\\ 1 & -4 & -8 & 4 & 1 & -2 & -2 & 0\\ -2 & -4 & 4 & -2 & -2 & -2 & 1 & 0\\ 0 & 0 & -6 & 6 & 0 & 0 & 0 & 3\end{array}\right].$$

which has as its inverse the matrix

$$P^{-1}=\frac{1}{3}\left[\begin{array}{cccc|cccc} -2 & 1 & -2 & 0 & 2 & -1 & 2 & 0 \\ 1 & -2 & -2 & 0 & -1 & 2 & 2 & 0 \\ -2 & -2 & 1 & 0 & 2 & 2 & -1 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 & 0 & -3 \\ \hline 2 & -1 & 2 & 0 & -4 & 2 & -4 & 0 \\ -2 & 4 & 4 & 0 & 3 & -6 & -6 & 0 \\ 8 & 8 & -4 & 6 & -10 & -10 & 5 & -6 \\ -4 & -4 & 2 & -6 & 4 & 4 & -2 & 9 \end{array}\right].$$

From P^{-1} it follows that the primary components of H are

$$\underset{E_1}{\mathrm{Sp}\begin{bmatrix} -2 & 1 \\ 1 & -2 \\ -2 & -2 \\ 0 & 0 \\ 2 & -1 \\ -2 & 4 \\ 8 & 8 \\ -4 & -4 \end{bmatrix}}, \quad \underset{E_2}{\mathrm{Sp}\begin{bmatrix} -2 & 0 \\ -2 & 0 \\ 1 & 0 \\ 0 & 3 \\ 2 & 0 \\ 4 & 0 \\ -4 & 6 \\ 2 & -6 \end{bmatrix}}, \quad \underset{E_3}{\mathrm{Sp}\begin{bmatrix} 2 & 0 \\ 2 & 0 \\ -1 & 0 \\ 0 & -3 \\ -4 & 0 \\ -6 & 0 \\ 5 & -6 \\ -2 & 9 \end{bmatrix}}, \quad \underset{E_4}{\mathrm{Sp}\begin{bmatrix} 2 & -1 \\ -1 & 2 \\ 2 & 2 \\ 0 & 0 \\ -4 & 2 \\ 3 & -6 \\ -10 & -10 \\ 4 & 4 \end{bmatrix}}.$$

We have $K^+=\phi^{-1}(T((2,2,0,0))\cap\mathscr{L}(4))=\phi^{-1}(E_1\oplus E_2)$. Using the bases for E_1 and E_2 given above, we can write $E_1\oplus E_2$ as $\mathrm{Sp}\left[\begin{smallmatrix}X\\Y\end{smallmatrix}\right]$ with X and Y each 4×4. By Lemma 10, $E_1\oplus E_2$ is complementary to $\mathrm{Sp}\{e_5, e_6, e_7, e_8\}$, so X is nonsingular. Using the formula $K^+=YX^{-1}$, we obtain

$$K^+=\begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & -2 & 0 & 0 \\ 0 & 0 & -4 & 2 \\ 0 & 0 & 2 & -2 \end{bmatrix}.$$

We have $K^-=\phi^{-1}(T((0,0,2,2))\cap\mathscr{L}(4))=\phi^{-1}(E_3\oplus E_4)$. Let $K_1=\phi^{-1}(T((2,0,2,0))\cap\mathscr{L}(4))=\phi^{-1}(E_1\oplus E_3)$, and $K_2=\phi^{-1}(T(0,2,0,2)\cap\mathscr{L}(4))=\phi^{-1}(E_2\oplus E_4)$. Using the same method as was used to compute K^+, we obtain

$$K^-=\begin{bmatrix} -2 & 0 & 0 & 0 \\ 0 & -3 & 0 & 0 \\ 0 & 0 & -5 & 2 \\ 0 & 0 & 2 & -3 \end{bmatrix},$$

$$K_1=\frac{1}{9}\begin{bmatrix} -13 & -4 & 2 & 0 \\ -4 & -22 & 2 & 0 \\ 2 & 2 & -37 & 18 \\ 0 & 0 & 18 & -27 \end{bmatrix}, \qquad K_2=\frac{1}{9}\begin{bmatrix} -14 & 4 & -2 & 0 \\ 4 & -23 & -2 & 0 \\ -2 & -2 & -44 & 18 \\ 0 & 0 & 18 & -18 \end{bmatrix}.$$

Next, we consider the 2-dimensional invariant torus, $\phi^{-1}(T((1,1,1,1)) \cap \mathscr{L}(4))$. By Proposition 4, $T((1,1,1,1)) \cap \mathscr{L}(4)$ consists of those 4-dimensional subspaces of $\mathbb{R}^8$ which are of the form $S_1 \oplus S_2 \oplus ([J(S_2)]^{\perp} \cap E_3) \oplus ([J(S_1)]^{\perp} \cap E_4)$ where S_1 and S_2 are 1-dimensional subspaces of E_1 and E_2 respectively. These subspaces are precisely those of the form

$$S(\theta_1, \theta_2) = \operatorname{Sp} P^{-1} \left[\begin{array}{cccc} \cos\theta_1 & 0 & 0 & 0 \\ -\sin\theta_1 & 0 & 0 & 0 \\ 0 & \cos\theta_2 & 0 & 0 \\ 0 & -\sin\theta_2 & 0 & 0 \\ \hline 0 & 0 & \sin\theta_1 & 0 \\ 0 & 0 & \cos\theta_1 & 0 \\ 0 & 0 & 0 & \sin\theta_2 \\ 0 & 0 & 0 & \cos\theta_2 \end{array}\right]$$

$$= \operatorname{Sp} \left[\begin{array}{cc} -2\cos\theta_1 - \sin\theta_1 & -2\cos\theta_2 \\ \cos\theta_1 + 2\sin\theta_1 & -2\cos\theta_2 \\ -2\cos\theta_1 + 2\sin\theta_1 & \cos\theta_2 \\ 0 & -3\sin\theta_2 \\ \hline 2\cos\theta_1 + \sin\theta_1 & 2\cos\theta_2 \\ -2\cos\theta_1 - 4\sin\theta_1 & 4\cos\theta_2 \\ 8\cos\theta_1 - 8\sin\theta_1 & -4\cos\theta_2 - 6\sin\theta_2 \\ -4\cos\theta_1 + 4\sin\theta_1 & 2\cos\theta_2 + 6\sin\theta_2 \end{array}\right.$$

$$\left.\begin{array}{cc} 2\sin\theta_1 - \cos\theta_1 & 2\sin\theta_2 \\ -\sin\theta_1 + 2\cos\theta_1 & 2\sin\theta_2 \\ 2\sin\theta_1 + 2\cos\theta_1 & -\sin\theta_2 \\ 0 & -3\cos\theta_2 \\ \hline -4\sin\theta_1 + 2\cos\theta_1 & -4\sin\theta_2 \\ 3\sin\theta_1 - 6\cos\theta_1 & -6\sin\theta_2 \\ -10\sin\theta_1 - 10\cos\theta_1 & 5\sin\theta_2 - 6\cos\theta_2 \\ 4\sin\theta_1 + 4\cos\theta_1 & -2\sin\theta_2 + 9\cos\theta_2 \end{array}\right].$$

Let $X(\theta_1, \theta_2)$ and $Y(\theta_1, \theta_2)$ denote the upper and lower submatrices respectively in the indicated basis matrix for $S(\theta_1, \theta_2)$. By Lemma 10, $S(\theta_1, \theta_2)$ is complementary to $\operatorname{Sp}\{e_5, e_6, e_7, e_8\}$, so $X(\theta_1, \theta_2)$ is nonsingular for all θ_1, θ_2. Since $\phi^{-1}(S(\theta_1, \theta_2)) = Y(\theta_1, \theta_2)X(\theta_1, \theta_2)^{-1}$, the 2-dimensional invariant torus for the SRDE is given by $\{Y(\theta_1, \theta_2)X(\theta_1, \theta_2)^{-1} : 0 \leqq \theta_1, \theta_2 < \pi\}$.

Next, we consider the 1-dimensional invariant tori. We could use Proposition 4 to describe each of the 4 1-dimensional invariant tori in a manner analogous to the description of the 2-dimensional invariant torus given above. Each torus would be described in the form $\{Y(\theta)X(\theta)^{-1} : 0 \leqq \theta < \pi\}$ where $X(\theta)$ and $Y(\theta)$ are each 4×4

matrix functions of θ, and $X(\theta)$ is invertible for every θ. However, there is an alternate method for computing the 1-dimensional invariant tori which has the advantage of not requiring the inversion of a matrix function of θ. The procedure is described in our paper [37].

In [37], the procedure is used to compute every periodic orbit under the additional assumption that if λ_j and λ_k are eigenvalues of H such that $\operatorname{Re}\lambda_j > 0$, $\operatorname{Im}\lambda_j > 0$, $\operatorname{Re}\lambda_k > 0$, $\operatorname{Im}\lambda_k > 0$, then $\operatorname{Im}\lambda_j$ and $\operatorname{Im}\lambda_k$ are noncommensurable. This assumption implies that the motion on every invariant torus of dimension greater than 1 is almost periodic. Consequently, the only periodic orbits are the 1-dimensional invariant tori. Without the additional assumption, the procedure can still be used to compute every 1-dimensional invariant torus. However, it cannot be used to compute periodic orbits which lie on invariant tori of dimension greater than 1.

The general idea is that each 1-dimensional invariant torus is bounded by a pair of equilibrium points, and these equilibrium points can be used to determine the torus in a computationally attractive way. To use this method, we find those pairs of equilibrium points K_α, K_β satisfying $K_\alpha \leqq K_\beta$, $\dim\ker(K_\beta - K_\alpha) = n-2$, and such that there are no other equilibrium points K which satisfy $K_\alpha \leqq K \leqq K_\beta$. There are $q2^{p+q-1}$ such pairs K_α, K_β. Given one such pair, we find an orthogonal matrix Z such that $Z'(K_\beta - K_\alpha)Z = \operatorname{diag}\{\gamma, \delta, 0, \cdots, 0\}$ with $0 < \delta \leqq \gamma$. Let

$$X(\theta) = \begin{bmatrix} \frac{1}{2}\gamma + \frac{1}{2}\gamma\cos\theta & \frac{1}{2}\sqrt{\gamma\delta}\sin\theta \\ \frac{1}{2}\sqrt{\gamma\delta}\sin\theta & \frac{1}{2}\delta - \frac{1}{2}\delta\cos\theta \end{bmatrix},$$

and let $\hat{X}(\theta)$ be the $n \times n$ matrix which has $X(\theta)$ as its upper left-hand 2×2 submatrix and has zeros for all its remaining entries. Then the unique 1-dimensional invariant torus $\{K(\theta): 0 \leqq \theta < 2\pi\}$ satisfying $K_\alpha \leqq K(\theta) \leqq K_\beta$ for all θ is given by $K(\theta) = K_\alpha + Z\hat{X}(\theta)Z'$.

It is easily checked that the pairs (K_1, K^+), (K_2, K^+), (K^-, K_1), and (K^-, K_2) each satisfy the 3 conditions on the pair (K_α, K_β) described above. Thus, the 4 1-dimensional invariant tori can be computed from these 4 pairs of equilibria. We will use this procedure to compute the 1-dimensional invariant torus $\phi^{-1}(T(2,1,1,0) \cap \mathscr{L}(4))$. It follows from the analysis in [37] that this periodic orbit is bounded by the equilibrium points $\phi^{-1}(T((2,0,2,0)) \cap \mathscr{L}(4))$ and $\phi^{-1}(T((2,2,0,0)) \cap \mathscr{L}(4))$, i.e. by K_1 and K^+. Thus, it is this pair of equilibrium points which is used to compute $\phi^{-1}(T((2,1,1,0)) \cap \mathscr{L}(4))$. We obtain

$$K(\theta) = \frac{1}{18}\begin{bmatrix} -22-4\cos\theta & -4-4\cos\theta & 2+2\cos\theta & 6\sin\theta \\ -4-4\cos\theta & -40-4\cos\theta & 2+2\cos\theta & 6\sin\theta \\ 2+2\cos\theta & 2+2\cos\theta & -73-\cos\theta & 36-3\sin\theta \\ 6\sin\theta & 6\sin\theta & 36-3\sin\theta & -45+9\cos\theta \end{bmatrix}.$$

The 3 other 1-dimensional invariant tori can be computed in a similar way.

Next, we illustrate the convergence results. Let K_0 be the matrix

$$\begin{bmatrix} -2 & 0 & 0 & 0 \\ 0 & -2 & 0 & 0 \\ 0 & 0 & -4 & 2 \\ 0 & 0 & 2 & -2 \end{bmatrix},$$

and suppose we want to determine the behavior of the solution $K(t, K_0)$ as $t \to -\infty$. Since $K_0 - K^-$ is nonnegative definite, $K_0 \in F_-$. In other words, $K(t, K_0)$ has no finite

escape time as t decreases from 0. Consequently, $K(t, K_0)$ must approach one of the 9 invariant tori as $t \to -\infty$. To determine which torus the solution approaches, we use the unstable flag $N_1 \subset N_2 \subset N_3 \subset N_4$, where $N_1 = E_4$, $N_2 = E_3 \oplus E_4$, $N_3 = E_2 \oplus E_3 \oplus E_4$, and $N_4 = E_1 \oplus E_2 \oplus E_3 \oplus E_4 = \mathbb{R}^8$. By straightforward linear algebra, one finds that $\dim \phi(K_0) \cap N_1 = 0$, $\dim \phi(K_0) \cap N_2 = 1$, $\dim \phi(K_0) \cap N_3 = 2$, and $\dim \phi(K_0) \cap N_4 = 4$. Applying Theorem 23(d), we conclude that K_0 belongs to the unstable manifold of $\phi^{-1}(T((2,1,1,0)) \cap \mathscr{L}(4))$, which is the 1-dimensional invariant torus described in detail above.

At this point, we know that $K(t, K_0)$ approaches the invariant torus $\phi^{-1}(T((2,1,1,0)) \cap \mathscr{L}(4))$ as $t \to -\infty$. However, we can actually determine exactly which motion on the torus it approaches. To do this, we first compute the subspaces $\phi(K_0) \cap N_1$, $\phi(K_0) \cap N_2$, $\phi(K_0) \cap N_3$, $\phi(K_0) \cap N_4$, and then their images under the projections η_4, η_3, η_2, η_1 respectively. From this, we obtain

$$\hat{\Pi}_-(\phi(K_0)) = \mathrm{Sp}\left[\begin{array}{rrrr} -4 & -1 & 0 & 2 \\ 5 & -1 & 0 & 2 \\ 2 & -4 & 0 & -1 \\ 0 & 0 & 1 & 0 \\ \hline 4 & 1 & 0 & -4 \\ -10 & 2 & 0 & -6 \\ -8 & 16 & 2 & 5 \\ 4 & -8 & -2 & -2 \end{array}\right].$$

Let X and Y denote the upper and lower submatrices of the indicated matrix. Then

$$\phi^{-1}(\hat{\Pi}_-(\phi(K_0)) = YX^{-1} = \frac{1}{9}\begin{bmatrix} -13 & -4 & 2 & 0 \\ -4 & -22 & 2 & 0 \\ 2 & 2 & -37 & 18 \\ 0 & 0 & 18 & -18 \end{bmatrix}.$$

We see that this matrix is in fact $K(\theta)$ for $\theta = 0$. By Theorem 23(d), the solution $K(t, K_0)$ converges to the periodic solution $K(t, \phi^{-1}(\hat{\Pi}_-(\phi(K_0))))$ as $t \to -\infty$.

6. Generalizations. In this section, we generalize the results in §§ 3, 4, and 5 to permit the infinitesimal generator (B or H) to have multiple eigenvalues. The results extend naturally provided we assume that the generator is semisimple (diagonalizable).

6.1. Extended Riccati differential equation. In this subsection, we generalize the results in § 3 concerning the phase portrait of the ERDE on $G^n(\mathbb{R}^{n+m})$. In place of Assumption A1, we assume only that the $(n+m) \times (n+m)$ matrix B is semisimple.

Let $\mu_1 < \mu_2 < \cdots < \mu_r$ denote the distinct real parts of the eigenvalues of B. Redefine E_i to be the direct sum of the primary components of B corresponding to all those eigenvalues of B with real part μ_i, and let n_i denote the dimension of E_i. Let $l = (l_1, \cdots, l_r)$ and $T(l)$ be defined as in § 3.1 (but using the new definition of E_i). $T(l)$ is no longer a torus. Instead, $T(l)$ is isomorphic to the product of Grassmann manifolds $G^{l_1}(\mathbb{R}^{n_1}) \times \cdots \times G^{l_r}(\mathbb{R}^{n_r})$ which has dimension equal to $\sum_{i=1}^{r} l_i(n_i - l_i)$.

It is clear that $T(l)$ is both positively and negatively invariant with respect to the flow of the ERDE. The argument given to establish the almost periodicity assertion in Theorem 1 is easily adapted to show that every motion on $T(l)$ is at worst almost periodic. It should be noted, however, that in contrast to the situation when Assumption

A1 is in force, a given $T(l)$ may contain equilibria, periodic motions of different periods, and nonperiodic almost periodic motions. Thus, the generalization of Theorem 1 is the following:

THEOREM 1′. *$T(l)$ is a product of Grassmann manifolds $G^{l_1}(\mathbb{R}^{n_1})\times\cdots\times G^{l_r}(\mathbb{R}^{n_r})$ of dimension $\sum_{j=1}^{r} l_j(n_j-l_j)$ which is both positively and negatively invariant. Every motion on $T(l)$ is almost periodic.*

Let $W^s(T(l))$, $W^u(T(l))$, $\{M_j\}_1^r$, Π_+, Π_-, $W^s(S_1)$, $W^u(S_1)$ be defined as in § 3.1. (Of course, the new definition of E_j must be used.) Trivial changes in the proof of Theorem 2 show that it holds in the new setting with no change in wording. Furthermore, $S(t, S_0)\to S(t, \Pi_+(S_0))$ as $t\to\infty$, and $S(t, S_0)\to S(t, \Pi_-(S_0))$ as $t\to-\infty$. No changes are required in either the statement or proof of Theorem 3.

To decompose the stable manifolds into cells, refine the stable flag $\{M_i\}_1^r$ to a complete flag

$$M_{11}\subset M_{12}\subset\cdots\subset M_{1n_1}\subset M_{21}\subset M_{22}\subset\cdots\subset M_{2n_2}\subset\cdots\subset M_{r1}\subset M_{r2}\subset\cdots\subset M_{rn_r}$$

where $M_{in_i}=M_i$. Let $l=(l_1,\cdots,l_r)$ be fixed. Let $b\equiv\{b_{ij}: j=1,\cdots,n_i;\ i=1,\cdots,r\}$ be such that $b_{ij}=0$ or 1 and $\sum_{j=1}^{n_i} b_{ij}=l_i$, $i=1,\cdots,r$. Let $W^s(T(l), b)\equiv\{S\in G^n(\mathbb{R}^{n+m}): \dim S\cap M_{ik}=\sum_{j=0}^{i-1} l_j+\sum_{j=1}^{k} b_{ij}, \forall i, k\}$ where we define $l_0=0$ for convenience. Then $W^s(T(l), b)$ is a subset of $W^s(T(l))$, and is a Schubert cell with respect to the complete flag $\{M_{ik}\}$. It is straightforward to show that

$$\dim W^s(T(l), b)=\sum_{i=1}^{r}\sum_{j=1}^{n_i} b_{ij}\left(\sum_{k=0}^{i-1}(n_k-l_k)+\sum_{k=1}^{j}(1-b_{ik})\right),$$

where we define $n_0=0$ for convenience. It follows immediately from this formula that $W^s(T(l))$ contains a unique cell of lowest dimension, and this dimension is equal to $\sum_{i=1}^{r} l_i\sum_{k=0}^{i-1}(n_k-l_k)$.

Recall from § 3.3 that $b_i(M, K)$ denotes the ith Betti number of the manifold M for the coefficient field K, and $b_s(G^n(\mathbb{R}^{n+m}), Z_2)$ is equal to the number of partitions of s into n parts of size less than or equal to m. From the formula above for $\dim W^s(T(l), b)$ together with some combinatorial analysis, we obtain the following generalization of Theorem 4. (The result for $W^u(T(l))$ is obtained in a completely analogous manner.)

THEOREM 4′. *$W^s(T(l))$ ($W^u(T(l))$) is the disjoint union of $\binom{n_1}{l_1}\cdots\binom{n_r}{l_r}$ cells. The number of cells of dimension*

$$\sum_{i=1}^{r} l_i\sum_{k=0}^{i-1}(n_k-l_k)+\nu \qquad \left(\text{dimension } \sum_{i=1}^{r} l_{r-i+1}\sum_{k=0}^{i-1}(n_{r-k+1}-l_{r-k+1})+\nu\right),$$

$$\nu=0, 1, \cdots, \sum_{i=1}^{r} l_i(n_i-l_i),$$

is equal to

$$\sum_{(\nu_1,\cdots,\nu_r)}\prod_{i=1}^{r} b_{\nu_i}(G^{l_i}(\mathbb{R}^{n_i}), Z_2),$$

where the summation is over all $(\nu_1,\cdots,\nu_r)$ satisfying $\nu_1+\cdots+\nu_r=\nu$ and $0\leqq\nu_i\leqq l_i(n_i-l_i)$, $i=1,\cdots,r$.

The generalization of Theorem 5 is as follows:

THEOREM 5′. *Let $S_1\in T(l)$. Then*

(a) *$W^s(S_1)$ is analytically isomorphic to Euclidean space of dimension*

$$\sum_{i=1}^{r} l_i\sum_{k=0}^{i-1}(n_k-l_k).$$

(b) *$W^u(S_1)$ is analytically isomorphic to Euclidean space of dimension*

$$\sum_{i=1}^{r} l_{r-i+1} \sum_{k=0}^{i-1} (n_{r-k+1} - l_{r-k+1}).$$

Proof. To prove (a), refine $\{M_j\}_1^r$ by first inserting the subspace $M_{j-1} \oplus (S_1 \cap E_j)$ between M_{j-1} and M_j, $j = 1, \cdots, r$. Then refine the resulting flag to a complete flag in any way. It follows from Theorem 3 and the proof of Theorem 4′ that $W^s(S_1)$ coincides with the unique lowest-dimensional cell in the decomposition of $W^s(T(l))$ determined by the complete flag. The assertion then follows immediately from Theorem 4′. (b) is proved analogously. □

It is no longer true that the ERDE has either an equilibrium point or a periodic orbit whose stable (unstable) manifold is open and dense. Instead, there is an invariant Grassmann manifold whose stable (unstable) manifold is open and dense. The proof of Theorem 6 can be modified in an obvious way to obtain

THEOREM 6′. (a) *Suppose that $\sum_{j=\nu+2}^{r} n_j < n \leqq \sum_{j=\nu+1}^{r} n_j$, let $k = n - \sum_{j=\nu+2}^{r} n_j$, and let $l_j = 0$ for $j \leqq \nu$, $l_{\nu+1} = k$, $l_j = n_j$ for $j \geqq \nu+2$. Then $T(l)$ is isomorphic to $G^k(\mathbb{R}^{n_{\nu+1}})$ and $W^s(T(l))$ is open and dense.*

(b) *Suppose that $\sum_{j=1}^{\nu-1} n_j < n \leqq \sum_{j=1}^{\nu} n_j$, let $k = n - \sum_{j=1}^{\nu-1} n_j$, and let $l_j = n_j$ for $j \leqq \nu-1$, $l_\nu = k$, $l_j = 0$ for $j \geqq \nu+1$. Then $T(l)$ is isomorphic to $G^k(\mathbb{R}^{n_\nu})$ and $W^u(T(l))$ is open and dense.*

Let $l = (l_1, \cdots, l_r)$ be fixed. From Theorem 4′, $W^s(T(l))$ is the union of $\binom{n_1}{l_1} \cdots \binom{n_r}{l_r}$ cells, with a unique highest dimensional cell which has dimension $\sum_{i=1}^{r} l_i \sum_{k=0}^{i} (n_k - l_k)$. By an obvious generalization of the argument preceding Theorem 7, the other $\binom{n_1}{l_1} \cdots \binom{n_r}{l_r} - 1$ cells can be "thickened" to give $\binom{n_1}{l_1} \cdots \binom{n_r}{l_r}$ submanifold charts for $W^s(T(l))$ (relative to the standard charts for $G^n(\mathbb{R}^{n+m})$). Thus, $W^s(T(l))$ is an embedded submanifold of $G^n(\mathbb{R}^{n+m})$ of dimension $\sum_{i=1}^{r} l_i \sum_{k=0}^{i} (n_k - l_k)$. Π_+ maps $W^s(T(l))$ onto $T(l)$, and if $S_1 \in T(l)$, then by Theorem 5′, $\Pi_+^{-1}(S_1)$ is isomorphic to Euclidean space of dimension $d_s \equiv \sum_{i=1}^{r} l_i \sum_{k=0}^{i-1} (n_k - l_k)$. Each of the $\binom{n_1}{l_1} \cdots \binom{n_r}{l_r}$ charts W_i for $W^s(T(l))$ is the inverse image of a corresponding chart $\bar{W}_i$ for $T(l) \cong G^{l_1}(\mathbb{R}^{n_1}) \times \cdots \times G^{l_r}(\mathbb{R}^{n_r})$. Furthermore, there exist real analytic isomorphisms $\gamma_i : W_i \to \bar{W}_i \times \mathbb{R}^{d_s}$ such that $p_i \circ \gamma_i = \Pi_+ | W_i$, where $p_i : \bar{W}_i \times \mathbb{R}^{d_s} \to \bar{W}_i$ is the natural projection. Hence, $\Pi_+ : W^s(T(l)) \to T(l)$ is a locally trivial bundle with $\mathbb{R}^{d_s}$ as fiber. The transition functions are invertible polynomial mappings of $\mathbb{R}^{d_s}$. Thus, Theorem 7 generalizes to give

THEOREM 7′.

(a) *$W^s(T(l))$ is an embedded submanifold of $G^n(\mathbb{R}^{n+m})$ of dimension $\sum_{i=1}^{r} l_i \sum_{k=0}^{i} (n_k - l_k)$.*

(b) *$\Pi_+ : W^s(T(l)) \to T(l)$ is a locally trivial bundle with fiber $\mathbb{R}^{d_s}$ and polynomial transition functions, where $d_s = \sum_{i=1}^{r} l_i \sum_{k=0}^{i-1} (n_k - l_k)$.*

(c) *$W^u(T(l))$ is an embedded submanifold of $G^n(\mathbb{R}^{n+m})$ of dimension $\sum_{i=1}^{r} l_{r-i+1} \sum_{k=0}^{i} (n_{r-k+1} - l_{r-k+1})$.*

(d) *$\Pi_- : W^u(T(l)) \to T(l)$ is a locally trivial bundle with fiber $\mathbb{R}^{d_u}$ and polynomial transition functions, where $d_u = \sum_{i=1}^{r} l_{r-i+1} \sum_{k=0}^{i-1} (n_{r-k+1} - l_{r-k+1})$.*

Lemma 1 is unaffected by the relaxation of the assumptions. Consequently, Proposition 1 generalizes to give

PROPOSITION 1′. *For any pair $l = (l_1, \cdots, l_r)$ and $l' = (l'_1, \cdots, l'_r)$, $W^s(T(l))$ and $W^u(T(l'))$ intersect transversally.*

In contrast to the situation when Assumption A1 is in force, it is no longer true that every equilibrium point is hyperbolic. Let S_0 be an equilibrium point, let $\{\alpha_i\}_1^n$ denote the eigenvalues of the restriction $B|S_0$, and let $\{\beta_j\}_1^m$ denote the remaining eigenvalues of B. It is easily shown that even if B is nondiagonalizable, the eigenvalues

of the linearization of the ERDE at S_0 are $\{\beta_j - \alpha_i : i = 1, \cdots, n; j = 1, \cdots, m\}$. It follows immediately that S_0 is hyperbolic if and only if $\{S_0\} = T(l)$ for some l. Thus, the correct generalization of Proposition 2 (which is valid even if B is not semisimple) is

PROPOSITION 2'. *The hyperbolic equilibrium points of the* ERDE *are* $\{T(l) : \dim T(l) = 0\}$.

If $\dim T(l) = 1$, then there exists $j_0 \in \{1, \cdots, r\}$ such that $l_{j_0} = 1$, $n_{j_0} = 2$, and $l_i = 0$ or $l_i = n_i$ for $i \neq j_0$. Topologically, $T(l)$ is a circle. There are 2 cases to consider depending on whether $B|E_{j_0}$ has a real eigenvalue of multiplicity 2 or a pair of complex conjugate eigenvalues. In the first case, every point on $T(l)$ is B-invariant and hence an equilibrium point. In the second case, $T(l)$ is an isolated periodic orbit. The construction of a Poincaré map for $T(l)$ and the calculation of the eigenvalues of its derivative are identical to the construction and calculation which establish Proposition 3. Consequently, in place of the corollary to that result we have

PROPOSITION 3'. *If* $\dim T(l) = 1$, *then either every point on* $T(l)$ *is an equilibrium point, or* $T(l)$ *is a hyperbolic periodic orbit.*

Under the assumption that B is semisimple, it follows from Theorem 1', Proposition 1', Proposition 2', and Proposition 3' that the ERDE is Morse-Smale if and only if each $T(l)$ has dimension at most 1 and each 1-dimensional $T(l)$ is a periodic orbit rather than a circle of equilibria. It is clear that excepting the trivial cases where $n = 0$ or $m = 0$, a necessary condition for the ERDE to be Morse-Smale is that Assumption A1 be satisfied.

By analogy to the definition of the Morse series for a Morse-Bott function (i.e. a function with nondegenerate critical submanifolds), we define the Morse series for the ERDE with semisimple generator B to be

$$M_B(t) = \sum_l P_{Z_2}(T(l); t)\, t^{\mathrm{Ind}\,(T(l))},$$

where $P_{Z_2}(T(l); t)$ is the Poincaré polynomial of $T(l)$ for the coefficient field Z_2 and $\mathrm{Ind}\,(T(l)) = \dim W^s(T(l)) - \dim T(l)$. By Theorem 1', $P_{Z_2}(T(l); t) = \Pi_{i=1}^r P_{Z_2}(G^{l_i}(\mathbb{R}^{n_i}); t)$. In the special case when $T(l)$ is a k-dimensional torus, this simplifies to $(1+t)^k$. Thus, the definition of $M_B(t)$ given here reduces to the definition given in § 3.3 if Assumption A1 is satisfied.

The next result generalizes Theorem 11 to show that the ERDE satisfies Morse-Bott type *equalities.*

THEOREM 11'. *Suppose that* $B \in \mathrm{gl}\,(n+m, \mathbb{R})$ *is semisimple. Then*

$$M_B(t) = P_{Z_2}(G^n(\mathbb{R}^{n+m}); t).$$

Proof. By refining the stable flag to a complete flag, we obtain a decomposition of $G^n(\mathbb{R}^{n+m})$ into $\binom{n+m}{n}$ cells. As in the proof of Theorem 11, we have

$$P_{Z_2}(G^n(\mathbb{R}^{n+m}); t) = \sum_{\substack{\text{all} \\ \text{cells}}} t^{\dim \text{cell}}.$$

It follows from Theorem 4' that

$$t^{\mathrm{Ind}\,T(l)} \prod_{i=1}^r P_{Z_2}(G^{l_i}(\mathbb{R}^{n_i}); t) = \sum_{\substack{\text{all cells} \\ \text{in } W^s(T(l))}} t^{\dim \text{cell}}.$$

Summing this equation over all choices for l, the left-hand side gives $M_B(t)$ while the right-hand side gives $P_{Z_2}(G^n(\mathbb{R}^{n+m}); t)$. □

6.2. Extended symplectic Riccati differential equation. In this subsection, we generalize the results in § 4 concerning the phase portrait of the ESRDE on $\mathscr{L}(n)$. In place of Assumption A2, we assume only that the $2n \times 2n$ matrix H is semisimple and has no eigenvalues on the imaginary axis.

Let $\mu_1 < \mu_2 < \cdots < \mu_{2r}$ denote the distinct real parts of the eigenvalues of H. Redefine E_i to be the direct sum of the primary components of H corresponding to all those eigenvalues of H with real part μ_i, and let n_i denote the dimension of E_i. Since $H \in \mathrm{sp}(n, \mathbb{R})$, $\mu_i = -\mu_{2r-i+1}$ and $n_i = n_{2r-i+1}$.

Let $l = (l_1, \cdots, l_{2r})$ and $T(l)$ be defined as in § 4.1. $T(l)$ is isomorphic to the product of Grassmann manifolds $G^{l_1}(\mathbb{R}^{n_1}) \times \cdots \times G^{l_{2r}}(\mathbb{R}^{n_{2r}})$ which has dimension equal to $\sum_{i=1}^{2r} l_i(n_i - l_i)$. Essentially the same argument as that given in § 4.1 shows that Proposition 4 holds with no change in wording. Thus, $T(l) \cap \mathscr{L}(n)$ is nonempty if and only if $l_i + l_{2r-i+1} = n_i$, $i = 1, \cdots, r$, in which case $T(l) \cap \mathscr{L}(n)$ is isomorphic to $G^{l_1}(\mathbb{R}^{n_1}) \times \cdots \times G^{l_r}(\mathbb{R}^{n_r})$. In particular, $\dim T(l) \cap \mathscr{L}(n) = \frac{1}{2} \dim T(l)$. We obtain the following generalization of Theorem 12.

THEOREM 12′. *The nonwandering set of the* ESRDE *has* $(n_1+1) \cdots (n_r+1)$ *connected components,* $\{T(l) \cap \mathscr{L}(n) : l_i + l_{2r-i+1} = n_i,\ i = 1, \cdots, r\}$. $T(l) \cap \mathscr{L}(n)$ *is isomorphic to* $G^{l_1}(\mathbb{R}^{n_1}) \times \cdots \times G^{l_r}(\mathbb{R}^{n_r})$. *Every motion on* $T(l) \cap \mathscr{L}(n)$ *is almost periodic.*

Next we consider the structure of the stable manifold $W^s(T(l)) \cap \mathscr{L}(n)$ and the unstable manifold $W^u(T(l) \cap \mathscr{L}(n))$ of $T(l) \cap \mathscr{L}(n)$, where $l_i + l_{2r-i+1} = n_i$, $i = 1, \cdots, r$. Refine the stable flag $\{M_i\}_1^{2r}$ to a complete flag $\{M_{ij} : j = 1, \cdots, n_i;\ i = 1, \cdots, 2r\}$. By making a symplectic change of basis if necessary, we may assume $\{M_{ij}\}$ is the standard symplectic flag for $\mathbb{R}^{2n}$. From § 6.1, we know that $W^s(T(l))$ is the union of

$$\left[\binom{n_1}{l_1} \cdots \binom{n_r}{l_r}\right]^2$$

cells $\{W^s(T(l), b)\}$. It follows from Lemmas 6 and 7 that $W^s(T(l), b) \cap \mathscr{L}(n)$ is nonempty if and only if $b_{ij} + b_{2r-i+1, n_i - j+1} = 1$, all i, j. Thus, exactly $\binom{n_1}{l_1} \cdots \binom{n_r}{l_r}$ cells have nonempty intersection with $\mathscr{L}(n)$. Suppose that $b_{ij} + b_{2r-i+1, n_i-j+1} = 1$, all i, j. Using the formula for $\dim W^s(T(l), b)$ from § 6.1 together with Lemma 7 gives

$$\dim W^s(T(l), b) \cap \mathscr{L}(n) = \frac{1}{2}\left[n - \sum_{i=1}^{r} l_i + \sum_{i=1}^{2r} \sum_{j=1}^{n_i} b_{ij}\left(\sum_{k=0}^{i-1} (n_k - l_k) + \sum_{k=1}^{j} (1 - b_{ik})\right)\right].$$

The following generalization of Theorem 13 is an easy consequence.

THEOREM 13′. *Suppose that* $l_i + l_{2r-i+1} = n_i$, $i = 1, \cdots, r$. $W^s(T(l)) \cap \mathscr{L}(n)$ $(W^u(T(l)) \cap \mathscr{L}(n))$ *is the disjoint union of*

$$\binom{n_1}{l_1} \cdots \binom{n_r}{l_r} \textit{ cells.}$$

The number of cells of dimension

$$\frac{1}{2}\left[n - \sum_{i=1}^{r} l_i + \sum_{i=1}^{2r} l_i \sum_{k=0}^{i-1} (n_k - l_k)\right] + \nu$$

$$\left(\textit{dimension } \frac{1}{2}\left[n - \sum_{i=1}^{r} l_{2r-i+1} + \sum_{i=1}^{2r} l_{2r-i+1} \sum_{k=0}^{i-1} (n_{2r-k+1} - l_{2r-k+1})\right] + \nu\right),$$

$$\nu = 0, 1, \cdots, \sum_{i=1}^{r} l_i(n_i - l_i),$$

is equal to

$$\sum_{(\nu_1,\cdots,\nu_r)} \prod_{i=1}^{r} b_{\nu_i}(G^{l_i}(\mathbb{R}^{n_i}), Z_2),$$

where the summation is over all $(\nu_1, \cdots, \nu_r)$ *satisfying* $\nu_1 + \cdots + \nu_r = \nu$ *and* $0 \leqq \nu_i \leqq l_i(n_i - l_i)$, $i = 1, \cdots, r$.

The generalization of Theorem 14 is as follows:

THEOREM 14′. *Suppose that* $l_i + l_{2r-i+1} = n_i$, $i = 1, \cdots, r$. *Let* $S_1 \in T(l) \cap \mathscr{L}(n)$. *Then*

(a) $W^s(S_1) \cap \mathscr{L}(n)$ *is analytically isomorphic to Euclidean space of dimension*

$$\frac{1}{2}\left[n - \sum_{i=1}^{r} l_i + \sum_{i=1}^{2r} l_i \sum_{k=0}^{i-1} (n_k - l_k)\right].$$

(b) $W^u(S_1) \cap \mathscr{L}(n)$ *is analytically isomorphic to Euclidean space of dimension*

$$\frac{1}{2}\left[n - \sum_{i=1}^{r} l_{2r-i+1} + \sum_{i=1}^{2r} l_{2r-i+1} \sum_{k=0}^{i-1} (n_{2r-k+1} - l_{2r-k+1})\right].$$

Proof. (a) As in the proof of Theorem 14, it is clear that by making a symplectic change of coordinates, we may assume that the standard symplectic flag refines the stable flag $\{M_j\}_1^{2r}$ and that S_1 is spanned by $\{e_{m_{i-1}+j} : 1 \leqq i \leqq r$ such that $l_i \neq 0; j = 1, \cdots, l_i\} \cup \{e_{3n-m_{i-1}+1-j} : r+1 \leqq i \leqq 2r$ such that $l_i \neq 0$; $j = 1, \cdots, l_i\}$ where $m_i = \dim M_i$. Then (as in the proof of Theorem 5′) $W^s(S_1)$ coincides with the lowest dimensional cell in the decomposition of $W^s(T(l))$ induced by the standard symplectic flag. In other words, $W^s(S_1) = W^s(T(l), \mathrm{b})$ where $b_{ij} = 1$ for $j = 1, \cdots, l_i$ and $b_{ij} = 0$ for $j = l_i + 1, \cdots, n_i$. Then $W^s(S_1) \cap \mathscr{L}(n) = W^s(T(l), b) \cap \mathscr{L}(n)$ coincides with the lowest dimensional cell in the corresponding decomposition of $W^s(T(l)) \cap \mathscr{L}(n)$, and is therefore isomorphic to Euclidean space. The dimension formula follows immediately from Theorem 13′.

The proof of (b) is analogous to the proof of (a). □

Theorem 15 remains true under the weaker assumptions, and its proof is unchanged. Thus, the ESRDE has an equilibrium point whose stable manifold is open and dense in $\mathscr{L}(n)$, and an equilibrium point whose unstable manifold is open and dense in $\mathscr{L}(n)$.

Theorem 16 generalizes to give the following result in a manner which is analogous to the generalization of Theorem 7 to give Theorem 7′.

THEOREM 16′. *Suppose that* $l_i + l_{2r-i+1} = n_i$, $i = 1, \cdots, r$. *Then*

(a) $W^s(T(l)) \cap \mathscr{L}(n)$ *is an embedded submanifold of* $\mathscr{L}(n)$ *of dimension*

$$\frac{1}{2}\left[n - \sum_{i=1}^{r} l_i + \sum_{i=1}^{2r} l_i \sum_{k=0}^{i} (n_k - l_k)\right].$$

(b) $\hat{\Pi}_+ : W^s(T(l)) \cap \mathscr{L}(n) \to T(l) \cap \mathscr{L}(n)$ *is a locally trivial bundle with fiber* $\mathbb{R}^{d_s}$ *and polynomial transition functions, where*

$$d_s = \frac{1}{2}\left[n - \sum_{i=1}^{r} l_i + \sum_{i=1}^{2r} l_i \sum_{k=0}^{i-1} (n_k - l_k)\right].$$

(c) $W^u(T(l)) \cap \mathscr{L}(n)$ *is an embedded submanifold of* $\mathscr{L}(n)$ *of dimension*

$$\frac{1}{2}\left[n - \sum_{i=1}^{r} l_{2r-i+1} + \sum_{i=1}^{2r} l_{2r-i+1} \sum_{k=0}^{i} (n_{2r-k+1} - l_{2r-k+1})\right].$$

(d) $\hat{\Pi}_-: W^u(T(l)) \cap \mathscr{L}(n) \to T(l) \cap \mathscr{L}(n)$ *is a locally trivial bundle with fiber* $\mathbb{R}^{d_u}$ *and polynomial transition functions, where*

$$d_u = \frac{1}{2}\left[n - \sum_{i=1}^{r} l_{2r-i+1} + \sum_{i=1}^{2r} l_{2r-i+1} \sum_{k=0}^{i-1} (n_{2r-k+1} - l_{2r-k+1}) \right].$$

Lemma 8 is unaffected by the weakened assumptions. Consequently, Proposition 5 holds with no changes required in the proof other than the replacement of Theorem 13 by Theorem 13′. Hence, $W^s(T(l)) \cap \mathscr{L}(n)$ and $W^u(T(l')) \cap \mathscr{L}(n)$ intersect only transversally.

It is no longer true that every equilibrium point of the ESRDE is hyperbolic. Let S_0 be an equilibrium point and let $\{\lambda_i\}_1^n$ denote the eigenvalues of the restriction $H|S_0$. It is easily shown that even if H is nondiagonalizable, the eigenvalues of the linearization of the ESRDE at S_0 are $\{-\lambda_j - \lambda_i : 1 \leqq i \leqq j \leqq n\}$. It follows immediately that the correct generalization of Proposition 6 (which is valid even if H is not semisimple) is

PROPOSITION 6′. *The hyperbolic equilibrium points of the* ESRDE *are* $\{T(l)$: $l_j = 0$ *or* $l_j = n_j$ *and* $l_j + l_{2r-j+1} = n_j$, $j = 1, \cdots, r\}$.

Thus, the ESRDE has exactly 2^r hyperbolic equilibria.

If $\dim T(l) \cap \mathscr{L}(n) = 1$, then each element of $T(l) \cap \mathscr{L}(n)$ is of the form $E_{j_1} \oplus \cdots \oplus E_{j_{r-1}} \oplus \tilde{S} \oplus ([J(\tilde{S})]^{\perp} \cap E_{2r-j_0+1})$ where $j \in \{j_1, \cdots, j_{r-1}\}$ iff $2r-j+1 \notin \{j_1, \cdots, j_{r-1}\}$, j_0 and $2r-j_0+1$ do not belong to $\{j_1, \cdots, j_{r-1}\}$, $\dim E_{j_0} = 2$, and $\tilde{S}$ is any 1-dimensional subspace of E_{j_0}. Topologically, $T(l) \cap \mathscr{L}(n)$ is a circle. There are 2 cases to consider depending on whether $H|E_{j_0}$ has a real eigenvalue of multiplicity 2 or a pair of complex conjugate eigenvalues. In the first case, every point on $T(l) \cap \mathscr{L}(n)$ is H-invariant and hence an equilibrium point. In the second case, $T(l) \cap \mathscr{L}(n)$ is an isolated periodic orbit. The construction of a Poincaré map for $T(l) \cap \mathscr{L}(n)$ and the calculation of the eigenvalues of its derivative are identical to the construction and calculation which establish Proposition 7. Consequently, in place of the corollary to that result we have

PROPOSITION 7′. *If* $\dim T(l) \cap \mathscr{L}(n) = 1$, *then either every point on* $T(l) \cap \mathscr{L}(n)$ *is an equilibrium point, or* $T(l) \cap \mathscr{L}(n)$ *is a hyperbolic periodic orbit.*

Under the assumption that H is semisimple and has no eigenvalues on the imaginary axis, it follows from Theorem 12′ and Propositions 5, 6′, 7′ that the ESRDE is Morse-Smale if and only if $T(l) \cap \mathscr{L}(n)$ has dimension at most 1 and each 1-dimensional $T(l) \cap \mathscr{L}(n)$ is a periodic orbit rather than a circle of equilibria. It is clear that a necessary condition for the ESRDE to be Morse-Smale is that Assumption A2 be satisfied.

We define the Morse series for the ESRDE with semisimple generator H to be

$$M_H(t) = \sum_l P_{Z_2}(T(l) \cap \mathscr{L}(n); t) t^{\operatorname{Ind}(T(l) \cap \mathscr{L}(n))}$$

where $P_{Z_2}(T(l) \cap \mathscr{L}(n); t)$ is the Poincaré polynomial of $T(l) \cap \mathscr{L}(n)$ for the coefficient field Z_2, $\operatorname{Ind}(T(l) \cap \mathscr{L}(n)) = \dim W^s(T(l)) \cap \mathscr{L}(n) - \dim T(l) \cap \mathscr{L}(n)$, and the summation is over those l satisfying $l_i + l_{2r-i+1} = n_i$, $i = 1, \cdots, r$. By Theorem 12′, $P_{Z_2}(T(l) \cap \mathscr{L}(n); t) = \Pi_{i=1}^{r} P_{Z_2}(G^{l_i}(\mathbb{R}^{n_i}); t)$. The definition of $M_H(t)$ specializes to that given in § 4.3 if Assumption A2 is satisfied.

The next result generalizes Theorem 20 to show that the ESRDE satisfies Morse-Bott-type *equalities*. It is proven by using Theorem 13′ in place of Theorem 13 in the proof of Theorem 20.

THEOREM 20′. *Suppose that* $H \in \operatorname{sp}(n, \mathbb{R})$ *is semisimple and has no eigenvalues on the imaginary axis. Then*

$$M_H(t) = P_{Z_2}(\mathscr{L}(n); t).$$

6.3. Symplectic Riccati differential equation. In this subsection, we generalize the results in § 5 concerning the phase portrait of the SRDE on the space $S(n)$ of symmetric matrices. From Lemma 10 and Theorem 12′, we immediately obtain the following generalization of Theorem 21.

THEOREM 21′. *Suppose that H is semisimple and has no eigenvalues on the imaginary axis, and that $L = BB'$ with (A, B) controllable. The nonwandering set of the* SRDE *has $(n_1+1)\cdots(n_r+1)$ connected components, $\{\phi^{-1}(T(l)\cap\mathscr{L}(n))\colon l_i + l_{2r-i+1} = n_i,\ i = 1,\cdots,r\}$. $\phi^{-1}(T(l)\cap\mathscr{L}(n))$ is isomorphic to $G^{l_1}(\mathbb{R}^{n_1})\times\cdots\times G^{l_r}(\mathbb{R}^{n_r})$. Every motion on $\phi^{-1}(T(l)\cap\mathscr{L}(n))$ is almost periodic.*

No change in proof is required to show that Theorem 22 remains true under the weakened assumptions that H is semisimple and $L = BB'$, $Q = C'C$ with (A, B) controllable and (C, A) observable.

Let F_+, F_-, $R^s(l)$, $R^u(l)$ be as defined in § 5.2. No essential changes in the proof of Theorem 23 are required to show that it holds under the weakened assumptions that H is semisimple with no eigenvalues on the imaginary axis and $L = BB'$ with (A, B) controllable.

Since Theorem 22 holds under the weakened assumptions stated above, it follows immediately that Theorem 24 holds under the assumptions that H is semisimple and $L = BB'$, $Q = C'C$ with (A, B) controllable and (C, A) observable.

Theorems 21′ and 23 describe the nonwandering set and the asymptotic behavior of every solution of the SRDE assuming only that H is semisimple with no eigenvalues on the imaginary axis and that $L = BB'$ with (A, B) controllable. If in addition $Q = C'C$ with (C, A) observable, then Theorems 22 and 24 describe the (constant) signature of every nonwandering solution as well as the asymptotic signature of every solution.

While Theorem 21′ shows that every nonwandering solution is almost periodic, it does not specify which of these solutions are actually constant or periodic. However, prior results are available which do this. Since they apply even when H is not semisimple, their description is deferred to the next subsection.

6.4. Nondiagonalizable case. The phase portraits of the ERDE, ESRDE, and SRDE are considerably more complicated when the infinitesimal generators (B for the ERDE, H for the ESRDE and SRDE) are not semisimple.

First we consider the ERDE in the case where the $(n+m)\times(n+m)$ matrix B is nondiagonalizable. The sets $\{T(l)\}$ can be defined as in § 6.1. As before, $T(l)$ is isomorphic to the product $G^{l_1}(\mathbb{R}^{n_1})\times\cdots\times G^{l_r}(\mathbb{R}^{n_r})$ and is both positively and negatively invariant under the flow of the ERDE.

In contrast to the case where B is semisimple, it is generally not true that every motion on $T(l)$ is almost periodic. For example, if $n = m = 1$ and

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

then the ERDE is a differential equation on $G^1(\mathbb{R}^2)$, which is topologically the circle S^1. The only choice for l is $l = (1)$ for which $T(l) \cong G^1(\mathbb{R}^2)$. It is clear that $T(l)$ contains exactly 1 equilibrium point,

$$\operatorname{Sp}\begin{bmatrix} 1 \\ 0 \end{bmatrix},$$

and that every motion converges to this point as $t \to \pm\infty$.

In general, when B is not diagonalizable, the motion on $T(l)$ can be quite complicated. For example, the set of equilibria is the subvariety of $G^n(\mathbb{R}^{n+m})$ consisting of all n-dimensional B-invariant subspaces. The geometric structure of this subvariety

is still not completely understood although a number of results have been obtained in recent years. (See e.g. [11], [45], [44], [20], [39]). Similarly, given $T>0$, the set of points on nontrivial T-periodic orbits consists of those n-dimensional subspaces of $\mathbb{R}^{n+m}$ which are e^{BT}-invariant but not B-invariant.

We next consider the ESRDE assuming only that the Hamiltonian matrix H has no eigenvalues with zero real part. Proposition 4 remains valid as stated with only minor modification of the proof required. Thus, the ESRDE has $(n_1+1)\cdots(n_r+1)$ invariant manifolds given by $\{T(l)\cap\mathscr{L}(n)\colon l_i+l_{2r-i+1}=n_i,\ i=1,\cdots,r\}$, with $T(l)\cap\mathscr{L}(n)$ isomorphic to $G^{l_1}(\mathbb{R}^{n_1})\times\cdots\times G^{l_r}(\mathbb{R}^{n_r})$. Of course, $T(l)\cap\mathscr{L}(n)$ now may contain wandering points.

Since the equilibrium set of the ESRDE consists of $\{S\in\mathscr{L}(n)\colon H(S)\subseteq S\}$, it is clear that the equilibrium set is contained in the union of the $\{T(l)\cap\mathscr{L}(n)\}$. By Lemma 4, the mapping $S\to S\cap L^+(H)$ is a bijection of the equilibrium set onto the set of all invariant subspaces of the restriction $H|L^+(H)$. It is shown in [36] that this bijection is actually an isomorphism of projective varieties. Hence, the problem of describing the structure of the equilibrium set reduces to that of describing the variety of invariant subspaces of $H|L^+(H)$, and the results in the papers referred to above can again be applied.

If $S\in\mathscr{L}(n)$ generates a nontrivial periodic orbit of period $T>0$, then S is e^{HT}-invariant, but not H-invariant. It is clear that all such S belong to the union of the $\{T(l)\cap\mathscr{L}(n)\}$. By Lemma 4, the mapping $S\to S\cap L^+(H)$ gives a bijection of the set of points which generate nontrivial T-periodic orbits and the set of all subspaces of $L^+(H)$ which are e^{HT}-invariant but not H-invariant.

Finally, we consider the SRDE assuming only that $L=BB'$ with (A, B) controllable and that H has no eigenvalues on the imaginary axis. By Lemma 10, each invariant manifold $T(l)\cap\mathscr{L}(n)$ for the ESRDE is completely contained in $\mathscr{L}_0(n)$ and hence corresponds to an invariant manifold $\phi^{-1}(T(l)\cap\mathscr{L}(n))$ for the SRDE. $\phi^{-1}(T(l)\cap\mathscr{L}(n))$ is isomorphic to $G^{l_1}(\mathbb{R}^{n_1})\times\cdots\times G^{l_r}(\mathbb{R}^{n_r})$, and the union of the manifolds $\{\phi^{-1}(T(l)\cap\mathscr{L}(n))\}$ contains all of the equilibria and periodic orbits.

Let $A^+=A-BB'K^+$. A well-known result of J. C. Willems [47] states that there is a bijection of the set of all A^+-invariant subspaces of $\mathbb{R}^n$ onto the equilibrium set of the SRDE. This bijection is given by $M\to K^+P(M)+K^-(I-P(M))$, where $P(M)$ is the projection onto the A^+-invariant subspace M along $\Delta^{-1}(M^\perp)$. (Recall that $\Delta=K^+-K^-$.) It is shown in [36] that by identifying the equilibrium set with its image under the embedding ϕ, the bijection is an isomorphism of projective varieties. Consequently, the geometric structure of the equilibrium set is the same as that of the variety of invariant subspaces of A^+. This fact is used in [36] to obtain a description of some of the geometric properties of the equilibrium set.

In [37], we have generalized Willems' result to classify the periodic orbits as well as the equilibrium points. Given any $T>0$, the mapping $M\to K^+P(\mathrm{M})+K^-(I-P(\mathrm{M}))$ is a bijection of the set of e^{A^+T}-invariant subspaces of $\mathbb{R}^n$ onto the set of points which generate T-periodic motions. Of course, the subspaces which are e^{A^+T}-invariant but not A^+-invariant correspond to the points which generate nonconstant T-periodic motions.

More generally, let M be an arbitrary subspace of $\mathbb{R}^n$, let $P(M)$ denote the projection onto M along $\Delta^{-1}(\mathrm{M}^\perp)$, and let $K=K^+P(\mathrm{M})+K^-(I-P(\mathrm{M}))$. Then it follows easily that

$$\mathrm{Sp}\begin{bmatrix} I \\ K \end{bmatrix} = S^+\oplus S^-$$

where

$$S^+ = \mathrm{Sp}\begin{bmatrix} I \\ K^+ \end{bmatrix} P(\mathrm{M}) \quad \text{and} \quad S^- = \mathrm{Sp}\begin{bmatrix} I \\ K^- \end{bmatrix} (I - P(M)).$$

From the definition of $P(M)$ it follows that $P(\mathrm{M})'\Delta(I - P(\mathrm{M})) = 0$, which implies that $S^- \subseteq [J(S^+)]^\perp \cap L^-(H)$. Define a new Hamiltonian matrix $\tilde{H}$ by setting $\tilde{H}x = -x$, $\forall x \in L^+(H)$ and $\tilde{H}x = x$, $\forall x \in L^-(H)$. The $\tilde{H}$-invariance of S^+ implies the $\tilde{H}$-invariance of $[J(S^+)]^\perp$. Consequently, $[J(S^+)]^\perp = ([J(S^+)]^\perp \cap L^+(\tilde{H})) \oplus ([J(S^+)]^\perp \cap L^-(\tilde{H})) = L^+(H) \oplus ([J(S^+)]^\perp \cap L^-(H))$, which implies that $\dim [J(S^+)]^\perp \cap L^-(\mathrm{H}) = n - \dim S^+ = \dim S^-$. Hence, $S^- = [J(S^+)]^\perp \cap L^-(\mathrm{H})$. Thus, if K is of the form $K^+P(\mathrm{M}) + K^-(I - P(\mathrm{M}))$ with M an arbitrary subspace of $\mathbb{R}^n$, then $\phi(\mathrm{K})$ is of the form $S^+ \oplus ([J(S^+)]^\perp \cap L^-(\mathrm{H}))$ with S^+ an arbitrary subspace of $L^+(\mathrm{H})$. In other words, ϕ gives a one-to-one correspondence between the set of symmetric matrices of the form $K^+P(\mathrm{M}) + K^-(I - P(\mathrm{M}))$ and the subset of $\mathscr{L}(n)$ consisting of subspaces of the form $S^+ \oplus ([J(\mathrm{S}^+)]^\perp \cap L^-(\mathrm{H}))$.

Now, $S^+ \oplus ([J(S^+)]^\perp \cap L^-(H)) \in T(l) \cap \mathscr{L}(n)$ if and only if S^+ is of the form $S^+ = S_1 \oplus \cdots \oplus S_r$ with $S_i \in G^{l_i}(E_i)$. It is well-known that A^+ is the matrix for $H|L^+(H)$ with respect to the basis given by the columns of $\left[\begin{smallmatrix} I \\ K^+ \end{smallmatrix}\right]$. (See e.g. [35].) Let F_i denote the A^+-invariant subspace of $\mathbb{R}^n$ corresponding to the eigenvalues of $H|E_i$. In other words,

$$\begin{bmatrix} I \\ K^+ \end{bmatrix} (F_i) = E_i, \quad i = 1, \cdots, r.$$

Since

$$S^+ = \begin{bmatrix} I \\ K^+ \end{bmatrix} (M),$$

it follows that $\phi(K) \in T(l) \cap \mathscr{L}(n)$ if and only if M is of the form $W_1 \oplus \cdots \oplus W_r$ with $W_i \in G^{l_i}(F_i)$, $i = 1, \cdots, r$. We obtain the following result, which extends to the invariant manifolds $\{\phi^{-1}(T(l) \cap \mathscr{L}(n))\}$ the classification theorem in [47] for the equilibria and in [37] for the periodic orbits.

THEOREM 25. *Suppose that $L = BB'$ with (A, B) controllable and that H has no imaginary axis eigenvalues. Then $K \in \phi^{-1}(T(l) \cap \mathscr{L}(n))$ if and only if K is of the form*

$$K = K^+P(M) + K^-(I - P(M)),$$

where M is of the form $M = \bigoplus_{i=1}^r W_i$ with $W_i \in G^{l_i}(F_i)$ and $P(M)$ is the projection onto M along $\Delta^{-1}(M^\perp)$.

The proof of Theorem 22 remains valid if the assumptions are weakened to include only that $L = BB'$ with (A, B) controllable and $Q = C'C$ with (C, A) observable. Thus, under these assumptions, every $K \in \phi^{-1}(T(l) \cap \mathscr{L}(n))$ is nonsingular and has exactly $\sum_{i=1}^r l_i$ positive eigenvalues. This result is also an easy consequence of Theorem 25 together with the fact that $K^+ > 0$ and $K^- < 0$.

7. Conclusion. In this paper, we have given a complete description for the phase portraits of the extended Riccati differential equation (ERDE) on the Grassmann manifold, the extended symplectic Riccati differential equation (ESRDE) on the Lagrange-Grassmann manifold, and the symplectic Riccati differential equation (SRDE) on the space of real symmetric matrices. The results are true under very general conditions. In particular, our characterization of the phase portrait of the SRDE applies to an open and dense subset of the Riccati equations which arise from the optimal control and filtering problems.

Let us briefly recapitulate the principal features of the three phase portraits. Under Assumption A1, the nonwandering set of the ERDE is a union of finitely many invariant tori of various dimensions. The 0-dimensional invariant tori are the equilibrium points. The 1-dimensional invariant tori are isolated periodic orbits. The motion on an invariant torus of dimension greater than one is either periodic or almost periodic, depending on whether or not the imaginary parts of the associated eigenvalues are commensurable.

The key to describing the stable and unstable manifolds of the invariant tori is to associate with the ERDE a stable and unstable flag of subspaces. Each flag induces a partition of the Grassmann manifold into disjoint subsets. In the case of the stable (unstable) flag, the induced partition is precisely the partition of the phase space into the stable (unstable) manifolds of the invariant tori. By refining the stable (unstable) flag to a complete flag, a Schubert cell decomposition of the Grassmann manifold is obtained. The stable (unstable) manifold of a k-dimensional invariant torus is decomposed into the union of 2^k cells.

Although the ERDE is generally *not* a Morse-Smale vector field, it has several properties which bear striking resemblance to important properties of Morse-Smale vector fields. The nonwandering set of a Morse-Smale vector field is by definition the union of finitely many equilibrium points (0-dimensional invariant tori) and periodic orbits (1-dimensional invariant tori). As noted above, the nonwandering set of the ERDE is a union of finitely many invariant tori, but whose dimension may be greater than one. The stable and unstable manifolds of a Morse-Smale vector field are embedded submanifolds. We have proven that this is the case for the ERDE. For a Morse-Smale vector field, each stable (unstable) manifold is a bundle over the corresponding connected component of the nonwandering set (point or closed orbit) with Euclidean space as fiber. We have proven that each stable (unstable) manifold for the ERDE is a bundle over the corresponding invariant torus with Euclidean space as fiber. By definition, the stable and unstable manifolds of a Morse-Smale vector field always intersect transversally. We have shown that the same is true for the stable and unstable manifolds of the ERDE. The global phase portrait of a Morse-Smale vector field is related to the topology of the underlying manifold by the Morse inequalities for a dynamical system. We have seen that the ERDE satisfies a natural generalization of these inequalities. In fact, for the ERDE, they are actually *equalities.*

The ESRDE is a differential equation on the Lagrange-Grassmann manifold $\mathscr{L}(n)$. But it is useful to extend it to a differential equation on the Grassmann manifold $G^n(\mathbb{R}^{2n})$. When viewed as a differential equation on $G^n(\mathbb{R}^{2n})$, the ESRDE is a special case of the ERDE. Thus, the results for the ERDE give a complete description of the phase portrait for the ESRDE on $G^n(\mathbb{R}^{2n})$. To obtain the phase portrait for the ESRDE on $\mathscr{L}(n)$, we make use of the fact that the submanifold $\mathscr{L}(n)$ is both positively and negatively invariant under the flow of the ESRDE on $G^n(\mathbb{R}^{2n})$. Consequently, the nonwandering set of the ESRDE on $\mathscr{L}(n)$ is the intersection with $\mathscr{L}(n)$ of the nonwandering set of the ESRDE on $G^n(\mathbb{R}^{2n})$. Under Assumption A2, we have shown that a given invariant torus for the ESRDE on $G^n(\mathbb{R}^{2n})$ either does not intersect $\mathscr{L}(n)$ or intersects $\mathscr{L}(n)$ in a torus of one-half its dimension. Thus, the nonwandering set of the ESRDE on $\mathscr{L}(n)$ is a union of invariant tori. In fact, if $2p$ ($4q$) denotes the number of real (nonreal) eigenvalues of the associated Hamiltonian matrix H, then there are $\binom{q}{k}2^{p+q-k}$ invariant tori of dimension k, $k=0, 1, \cdots, q$.

It follows from the invariance of $\mathscr{L}(n)$ under the flow that the stable (unstable) manifolds of the invariant tori for the ESRDE on $\mathscr{L}(n)$ are obtained by intersecting with $\mathscr{L}(n)$ the stable (unstable) manifolds of the corresponding invariant tori for the ESRDE on $G^n(\mathbb{R}^{2n})$. If the stable (unstable) flag is refined to a complete flag in a way

which respects the symplectic structure, then the stable (unstable) manifold of each k-dimensional invariant torus (for the ESRDE on $\mathscr{L}(n)$) is the union of 2^k cells in the decomposition of $\mathscr{L}(n)$ determined by the complete flag.

The ESRDE is generally not a Morse-Smale vector field. However, like the ERDE, it exhibits properties which resemble those of a Morse-Smale vector field. The stable and unstable manifolds are embedded submanifolds of $\mathscr{L}(n)$, and are bundles over the corresponding tori with Euclidean spaces as fibers. The stable and unstable manifolds always intersect transversally. In addition, the phase portrait of the ESRDE is related to the topology of $\mathscr{L}(n)$ by generalized Morse inequalities. In fact, for the ESRDE, they are actually *equalities.*

There is one interesting difference between the phase portraits of the ERDE and the ESRDE. The ERDE has either an equilibrium point or a periodic orbit whose stable (unstable) manifold is open and dense in the phase space. However, the ESRDE always has an equilibrium point whose stable (unstable) manifold is open and dense.

Finally, we turn to the symplectic Riccati differential equation (SRDE) on the space $S(n)$ of real symmetric $n \times n$ matrices. It is this differential equation which is crucial to the optimal control and filtering applications. We have given a complete description of the phase portrait. We have shown that if (A, B) is controllable, the nonwandering set of the SRDE is analytically isomorphic to the nonwandering set of the corresponding ESRDE. Thus, under Assumption A2, there are $\binom{q}{k}2^{p+q-k}$ k-dimensional invariant tori, $k = 0, 1, \cdots, q$. Altogether, there are $2^p 3^q$ invariant tori. The asymptotic behaviour of a given motion at $t \to \infty$ ($t \to -\infty$) can be determined from the asymptotic behavior of the corresponding motion of the ESRDE, provided the former motion does not escape in finite positive (negative) time. Either the motion converges to a motion on one of the $2^p 3^q$ invariant tori as $t \to \infty$ ($t \to -\infty$), or it escapes in finite positive (negative) time. By combining our characterization of the stable and unstable manifolds for the ESRDE with a known result concerning finite escape times, we have described the asymptotic behavior of every solution of the SRDE.

We have a comment regarding the computational problems involved in determining the phase portrait for a given Riccati equation. To determine the number of invariant tori only requires determining how many of the eigenvalues of the Hamiltonian matrix H are real. To actually construct the invariant tori and their stable and unstable manifolds, it is necessary to find the eigenvectors of H. (See § 5.4 for an example.) Since H is assumed to have distinct eigenvalues, both of these tasks are reasonable from a numerical standpoint. In contrast, it is *not* reasonable to ask for the number of periodic orbits. There are either finitely many (in fact, $q2^{p+q-1}$) or uncountably many periodic orbits. Unless $q \leqq 1$, there exist arbitrarily small perturbations of H which produce either situation. Thus, it is not numerically feasible to determine whether or not there are finitely many periodic orbits.

The results for the ERDE extend in a natural way to the more general case where Assumption A1 is replaced by the assumption that the generator B is semisimple. The invariant manifolds in the nonwandering set are now products of Grassmannians instead of tori, and there is an invariant Grassmann submanifold which has a stable (unstable) manifold which is open and dense in $G^n(\mathbb{R}^{n+m})$. The stable and unstable manifolds remain unions of Schubert cells. If a Morse series is defined by analogy to the Morse series for a Morse-Bott function, the ERDE satisfies Morse inequalities—in fact as equalities.

The results for the ESRDE extend naturally to the case where Assumption A2 is replaced by the weaker assumption that the generator H is semisimple and has no eigenvalues on the imaginary axis. The connected components of the nonwandering set are now products of Grassmann manifolds, but as before, there is an equilibrium

point whose stable (unstable) manifold is open and dense in $\mathscr{L}(n)$. The stable and unstable manifolds are again unions of cells, and Morse inequalities are satisfied as equalities.

The results for the SRDE are generalized to the case where H is semisimple (diagonalizable) with no imaginary axis eigenvalues, and $L = BB'$ with (A, B) controllable. In this case, the connected components of the nonwandering set are products of Grassmann manifolds. A given solution either escapes in finite forward (backward) time or converges to an almost periodic solution on one of these products as $t \to \infty$ ($t \to -\infty$). For a convergent solution, our results give an explicit formula for the limiting motion. If $Q = C'C$ with (C, A) observable, then our results also describe the constant signature of every nonwandering solution and the asymptotic signature of every nonescaping solution.

When the generator matrices are nondiagonalizable, the phase portraits of the ERDE, ESRDE, and SRDE are considerably more complicated. There are still invariant manifolds which are products of Grassmannians and which contain all of the equilibrium points and periodic orbits. However, in contrast to the semisimple case, these invariant manifolds may contain wandering points. An indication of the degree of complexity resulting from nondiagonalizability is given by the complicated geometric structure of the variety of invariant subspaces of a nondiagonalizable matrix.

In this paper, we have not dealt explicitly with the RDE defined on the space $\mathbb{R}^{m\times n}$ of real $m \times n$ matrices. However, it should be clear that our description of the phase portrait of the ERDE will yield the complete phase portrait of the RDE provided two results become available: (1) A description of which nonwandering points for the ERDE belong to $G^n(\mathbb{R}^{n+m}) - G_0^n(\mathbb{R}^{n+m})$ (the hypersurface at infinity) or a sufficient condition on the generator matrix B for the nonwandering set of the ERDE to be completely contained in $G_0^n(\mathbb{R}^{n+m})$ (and hence correspond to the nonwandering set of the RDE); (2) A description of exactly which initial points K_0 generate solutions which escape in finite forward or backward time. Some work in the direction of these problems is described by J. Medanic [28].

Appendix A. Standard charts for the Grassmann and Lagrange–Grassmann manifolds. First we describe the so-called *standard charts* for the Grassmann manifold $G^n(\mathbb{R}^{n+m})$ of all n-dimensional subspaces of $\mathbb{R}^{n+m}$. Let $\alpha = (\alpha_1, \cdots, \alpha_n)$ be a multi-index. (I.e. $\alpha_1, \cdots, \alpha_n$ are integers satisfying $1 \leqq \alpha_1 < \alpha_2 < \cdots < \alpha_n \leqq n+m$). Let $e_1, \cdots, e_{n+m}$ denote the standard basis vectors for $\mathbb{R}^{n+m}$, and let L_α denote the m-dimensional subspace $\operatorname{Sp}\{e_j : j \notin \{\alpha_1, \cdots, \alpha_n\}\}$. Let W_α denote the subset of $G^n(\mathbb{R}^{n+m})$ consisting of those subspaces which are complementary to L_α. In other words, $W_\alpha = \{S \in G^n(\mathbb{R}^{n+m}) : S \cap L_\alpha = 0\}$.

Let $S \in W_\alpha$, and let X be any $(n+m) \times n$ full rank matrix such that $\operatorname{Sp} X = S$. In other words, the columns of X form a basis for S. Since S is complementary to L_α, rows $\alpha_1, \cdots, \alpha_n$ of X form a nonsingular $n \times n$ submatrix. It follows that by performing column operations on X, we can obtain a new basis matrix for S, call it $\tilde{X}$, such that rows $\alpha_1, \cdots, \alpha_n$ form an identity submatrix.

Let $\tilde{W}_\alpha$ denote the set of all $(n+m) \times n$ matrices such that rows $\alpha_1, \cdots, \alpha_n$ form an identity submatrix. We have seen that each $S \in W_\alpha$ is spanned by some $\tilde{X} \in \tilde{W}_\alpha$. Furthermore, it is clear that no two matrices in $\tilde{W}_\alpha$ can span the same subspace. Thus, there is a bijection $\phi_\alpha : W_\alpha \to \tilde{W}_\alpha$ with $\phi_\alpha(S)$ the unique matrix in $\tilde{W}_\alpha$ whose columns span S. Since $\tilde{W}_\alpha$ can be identified with $\mathbb{R}^{mn}$ in a natural way, we can regard ϕ_α as a bijection of W_α onto $\mathbb{R}^{mn}$. It can be shown [3] that $\phi_\alpha : W_\alpha \to \mathbb{R}^{mn}$ is actually a homeomorphism. The pair (W_α, ϕ_α) is called a *standard chart* for $G^n(\mathbb{R}^{n+m})$. By associating each $S \in W_\alpha$ with a point in $\mathbb{R}^{mn}$, the mapping ϕ_α parametrizes the subset

W_α of $G^n(\mathbb{R}^{n+m})$. Since any $S \in G^n(\mathbb{R}^{n+m})$ must be complementary to L_α for some α, the $\binom{n+m}{n}$ open subsets $\{W_\alpha\}$ cover $G^n(\mathbb{R}^{n+m})$. Thus, the standard charts parametrize all of $G^n(\mathbb{R}^{n+m})$.

As an example, suppose that $n=2$ and $m=3$. There are $\binom{5}{2}=10$ standard charts for $G^2(\mathbb{R}^5)$. If $\alpha=(2,5)$, then W_α consists of those $S \in G^2(\mathbb{R}^5)$ which are complementary to the subspace $L_\alpha = \mathrm{Sp}\{e_1, e_3, e_4\}$. The mapping ϕ_α^{-1} is given by

$$\begin{bmatrix} x_{11} & x_{12} \\ x_{31} & x_{32} \\ x_{41} & x_{42} \end{bmatrix} \longrightarrow \mathrm{Sp} \begin{bmatrix} x_{11} & x_{12} \\ 1 & 0 \\ x_{31} & x_{32} \\ x_{41} & x_{42} \\ 0 & 1 \end{bmatrix}.$$

We now describe the standard charts for the Lagrange-Grassmann manifold $\mathscr{L}(n)$, which is an embedded submanifold of $G^n(\mathbb{R}^{2n})$. $G^n(\mathbb{R}^{2n})$ is covered by $\binom{2n}{n}$ charts $\{W_\alpha\}$ as described above. As in § 4.2, we define the standard symplectic flag $\{V_j\}_{j=1}^{2n}$ by $V_j = \mathrm{Sp}\{e_1, e_2, \cdots, e_j\}$ for $j=1,2,\cdots,n$, and $V_j = \mathrm{Sp}\{e_1, e_2, \cdots, e_n, e_{2n}, e_{2n-1}, \cdots, e_{3n-j+1}\}$ for $j=n+1, n+2, \cdots, 2n$. This complete flag determines a cell decomposition of $G^n(\mathbb{R}^{2n})$ into $\binom{2n}{n}$ cells. There is a cell $U(a)$ for every $a=(a_1,\cdots,a_{2n})$ such that $a_i=0$ or 1 for each i and $\sum_{i=1}^{2n} a_i = n$. Specifically, $U(a)=\{S \in G^n(\mathbb{R}^{2n})$: $\dim S \cap V_j = \sum_{i=1}^{j} a_i, j=1,\cdots,2n\}$. Let $j_1 < \cdots < j_n$ denote the elements of the set $\{j: a_j = 1\}$. Let $d = \sum_{i=1}^{n} a_i$. Set $\alpha_i = j_i$ for $i=1,\cdots,d$ and $\alpha_i = 3n+1-j_{n+d+1-i}$ for $i=d+1,\cdots,n$. It is straightforward to show that if $S \in U(a)$, then S is complementary to L_α and hence an element of W_α. Thus, $U(a) \subset W_\alpha$.

Lemma 6 shows that $U(a) \cap \mathscr{L}(n)$ is empty unless $a_i + a_{2n-i+1} = 1$, $i=1,\cdots,n$. This condition on $a=(a_1,\cdots,a_{2n})$ is equivalent to the requirement that $\alpha_1,\cdots,\alpha_d$, $\alpha_{d+1}-n,\cdots,\alpha_n-n$ are all distinct. Since $G^n(\mathbb{R}^{2n})$ is covered by the cells $\{U(a)\}$, it follows trivially that $\mathscr{L}(n)$ is covered by the cells $\{U(a) \cap \mathscr{L}(n)$: $a_i + a_{2n-i+1} = 1, i = 1,\cdots,n\}$. Since $U(a) \subset W_\alpha$, $\mathscr{L}(n)$ is covered by the open sets $\{W_\alpha \cap \mathscr{L}(n)$: $\alpha_1,\cdots,\alpha_d, \alpha_{d+1}-n,\cdots,\alpha_n-n$ are all distinct$\}$, where we define $d = \max\{j: \alpha_j \leqq n\}$. It is clear that there are 2^n choices for $\alpha=(\alpha_1,\cdots,\alpha_n)$ which satisfy the additional requirement that $\alpha_1,\cdots,\alpha_d, \alpha_{d+1}-n,\cdots,\alpha_n-n$ be distinct. Thus, there are 2^n open sets $W_\alpha \cap \mathscr{L}(n)$ in the covering of $\mathscr{L}(n)$.

Fix $\alpha=(\alpha_1,\cdots,\alpha_n)$ such that $\alpha_1,\cdots,\alpha_d, \alpha_{d+1}-n,\cdots,\alpha_n-n$ are distinct. Let $S \in W_\alpha$, and let X be the unique matrix in $\tilde{W}_\alpha$ such that $S = \mathrm{Sp}\, X$. By a procedure analogous to the one described immediately prior to Example 2 in § 4.2, we can find $P \in \mathrm{Sp}(n,\mathbb{R})$ such that the matrix $\tilde{X} = PX$ has the following structure: The first n rows of $\tilde{X}$ form an identity submatrix. Rows $n+1,\cdots,n+d$ of $\tilde{X}$ are the nontrivial rows from among rows $n+1,\cdots,2n$ of X. Rows $n+d+1,\cdots,2n$ of $\tilde{X}$ are the nontrivial rows from among rows $1,\cdots,n$ of X, each multiplied by -1.

For example, suppose that $n=3$ and that $\alpha=(1,2,6)$. Note that $d=2$ and that α_1, α_2, and α_3-3 are distinct. $X \in \tilde{W}_\alpha$ iff X is of the form

$$X = \left[\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1 & 0 \\ x_{31} & x_{32} & x_{33} \\ \hline x_{41} & x_{42} & x_{43} \\ x_{51} & x_{52} & x_{53} \\ 0 & 0 & 1 \end{array}\right].$$

Then

$$\tilde{X} = \left[\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \hline x_{41} & x_{42} & x_{43} \\ x_{51} & x_{52} & x_{53} \\ -x_{31} & -x_{32} & -x_{33} \end{array}\right].$$

Since $P \in \mathrm{Sp}(n, \mathbb{R})$, S is Lagrangian iff $P(S)$ is Lagrangian. If we write X in partitioned form as $\tilde{X} = \left[\begin{smallmatrix} I \\ Y \end{smallmatrix}\right]$, then $P(S)$ is Lagrangian iff Y is symmetric. Thus, the condition that $Y = Y'$ gives the conditions which the n^2 coordinates in the matrix X must satisfy in order for the subspace S to belong to $\mathscr{L}(n)$. In the example, it follows that $S \in \mathscr{L}(3)$ iff $x_{51} = x_{42}$, $x_{31} = -x_{43}$, and $x_{32} = -x_{53}$. Thus, the subset $W_\alpha \cap \mathscr{L}(3)$ is parametrized by the matrices of the form

$$X = \left[\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -x_{43} & -x_{53} & x_{33} \\ \hline x_{41} & x_{42} & x_{43} \\ x_{42} & x_{52} & x_{53} \\ 0 & 0 & 1 \end{array}\right]$$

which may be regarded as points in $\mathbb{R}^6$. In exactly this way, 2^n charts which cover $\mathscr{L}(n)$ are obtained by intersecting with $\mathscr{L}(n)$ those standard charts W_α for $G^n(\mathbb{R}^{2n})$ for which α satisfies the additional requirement that $\alpha_1, \cdots, \alpha_d, \alpha_{d+1} - n, \cdots, \alpha_n - n$ be distinct. The image of the restriction $\phi_\alpha | W_\alpha \cap \mathscr{L}(n)$ may be identified with $\mathbb{R}^{n(n+1)/2}$. It is the 2^n submanifold charts $\{(W_\alpha \cap \mathscr{L}(n), \phi_\alpha | W_\alpha \cap \mathscr{L}(n))\}$ which we refer to as the standard charts for $\mathscr{L}(n)$.

Appendix B. Properties of almost periodic functions. 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 [12, Chaps. 1 and 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] S. BATTERSON, *Structurally stable Grassmann transformations*, Trans. Amer. Math. Soc., 231 (1977), pp. 385-404.

[2] G. D. BIRKHOFF, *Dynamical Systems*, AMS Colloquium Publications Vol. 9, American Mathematical Society, Providence, RI, 1927.

[3] W. M. BOOTHBY, *An Introduction to Differentiable Manifolds and Riemannian Geometry*, Academic Press, New York, 1975.

[4] A. BOREL, *Sur la cohomologie des espaces filtrés principaux et des espaces homogènes des groupes de Lie compacts*, Ann. Math., 57 (1953), pp. 115-176.

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

[6] R. S. BUCY, *Structural stability for the Riccati equation*, this Journal, 13 (1975), pp. 749-753.

[7] R. S. BUCY AND J. RODRIGUEZ-CANABAL, *A negative definite equilibrium and its induced cone of global existence for the Riccati equation*, SIAM J. Math. Anal., 3 (1972), pp. 644-646.

[8] F. M. CALLIER AND J. L. WILLEMS, *Criterion for the convergence of the solution of the Riccati differential equation*, IEEE Trans. Aut. Control, AC-26 (1981), pp. 1232-1242.

[9] J. CASTI, *The linear-quadratic control problem: some recent results and outstanding problems*, SIAM Rev., 22 (1980), pp. 459-485.

[10] C. CONLEY, *Isolated Invariant Sets and the Morse Index*, CBMS Regional Conference Series in Applied Mathematics 38, American Mathematical Society, Providence, RI, 1978.

[11] R. G. DOUGLAS AND C. PEARCY, *On a topology for invariant subspaces*, J. Funct. Anal., 2 (1968), pp. 323-341.

[12] A. M. FINK, *Almost Periodic Differential Equations*, Springer-Verlag, Berlin, 1974.

[13] E. E. FLOYD, *Periodic maps via Smith theory*, in Seminar on Transformation Groups, A. Borel, ed., Annals of Math. Studies No. 46, Princeton Univ. Press, Princeton, NJ, 1960.

[14] I. GOHBERG, P. LANCASTER AND L. RODMAN, *Matrix Polynomials*, Academic Press, New York, 1982.

[15] P. GRIFFITHS AND J. HARRIS, *Principles of Algebraic Geometry*, John Wiley, New York, 1978.

[16] R. HERMANN, *Cartanian Geometry, Nonlinear Waves, and Control Theory, Part A*, Interdisciplinary Mathematics, Vol. XX, Math. Sci. Press, Brookline, MA, 1979.

[17] R. HERMANN AND C. MARTIN, *Lie and Morse theory for periodic orbits of vector fields and matrix Riccati equations*, I: *General Lie-theoretic methods*, Math. Systems Theory, 15 (1982), pp. 277-284.

[18] ———, *Lie and Morse theory for periodic orbits of vector fields and matrix Riccati equations*, II, Math. Systems Theory, 16 (1983), pp. 297-306.

[19] ———, *Periodic solutions of the Riccati equation*, in Proc. 19th IEEE Conference on Decision and Control, 1980, pp. 645-648.

[20] R. HOTTA AND N. SHIMOMURA, *The fixed point subvarieties of unipotent transformations on generalized flag varieties and the green functions*, Math. Ann., 241 (1979), pp. 193-208.

[21] V. KUČERA, *A review of the matrix Riccati equation*, Kybernetika, 9 (1973), pp. 42-61.

[22] N. H. KUIPER, *Topological conjugacy of real projective transformations*, Topology, 15 (1976), pp. 13-22.

[23] P. LANCASTER AND L. RODMAN, *Existence and uniqueness theorems for the algebraic Riccati equation*, Int. J. Control, 32 (1980), pp. 285-309.

[24] A. G. J. MACFARLANE, *An eigenvector solution of the optimal linear regulator problem*, J. Electron. Contr., 14 (1963), pp. 496-501.

[25] K. MÅRTENSSON, *On the matrix Riccati equation*, Inform. Sci., 3 (1971), pp. 17-49.

[26] C. MARTIN, *Finite escape time for Riccati differential equations*, Systems and Control Letters, 1 (1981), pp. 127-131.

[27] ———, *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.

[28] J. MEDANIC, *Geometric properties and invariant manifolds of the Riccati equation*, IEEE Trans. Automat. Contr., AC-27 (1982), pp. 670-677.

[29] T. NISHIMURA AND H. KANO, *Periodic oscillations of matrix Riccati equations in time-invariant systems*, IEEE Trans. Automat. Contr., AC-25 (1980), pp. 749-755.
[30] J. PALIS AND S. SMALE, *Structural stability theorems*, in Global Analysis, Proc. Symposia in Pure Mathematics, American Mathematical Society Providence, RI, 1970.
[31] J. E. POTTER, *Matrix quadratic solutions*, SIAM J. Appl. Math., 14 (1966), pp. 496-501.
[32] W. T. REID, *Riccati differential equations*, Academic Press, New York, 1972.
[33] C. R. SCHNEIDER, *Global aspects of the matrix Riccati equation*, Math. Systems Theory, 7 (1973), pp. 281-286.
[34] M. A. SHAYMAN, *A Symmetry Group for the Matrix Riccati Equation*, Systems and Control Letters, 2 (1982), pp. 17-24.
[35] ———, *Geometry of the algebraic Riccati equation*, I, this Journal, 21 (1983), pp. 375-394.
[36] ———, *Geometry of the algebraic Riccati equation*, II, this Journal, 21 (1983), pp. 395-409.
[37] ———, *On the periodic solutions of the matrix Riccati equation*, Math. Systems Theory, 16 (1983), pp. 267-287.
[38] ———, *On the phase portrait of the matrix Riccati equation arising from the periodic control problem*, this Journal, 23 (1985), pp. 717-751.
[39] ———, *On the variety of invariant subspaces of a finite-dimensional linear operator*, Trans. Amer. Math. Soc., 274 (1982), pp. 721-747.
[40] ———, *Phase portrait of the Riccati equation from the periodic control problem*, Proc. 1984 American Control Conference, San Diego, CA, June, 1984, pp. 250-257.
[41] ———, *Varieties of invariant subspaces and the algebraic Riccati equation*, Ph.D. Thesis, Harvard Univ., Cambridge, MA, 1980.
[42] S. SMALE, *Differentiable dynamical systems*, Bull. Amer. Math. Soc., 73 (1967), pp. 747-817.
[43] ———, *Morse inequalities for a dynamical system*, Bull. Amer. Math. Soc., 66 (1960), pp. 43-49.
[44] T. A. SPRINGER, *A construction of representations of Weyl groups*, Inventiones Math., 44 (1978), pp. 279-293.
[45] R. STEINBERG, *On the desingularization of the unipotent variety*, Inventiones Math., 36 (1976), pp. 209-224.
[46] A. C. M. VAN SWIETEN, *Qualitative behavior of dynamical games with feedback strategies*, Ph.D. Thesis, Univ. Groningen, the Netherlands, 1977.
[47] J. C. WILLEMS, *Least squares stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Automat Control, AC 16 (1971), pp. 621-634.
[48] H. K. WIMMER, *On the algebraic Riccati equation*, Bull. Austral. Math. Soc., 14 (1976), pp. 457-461.
[49] W. M. WONHAM, *On a matrix Riccati equation of stochastic control*, this Journal, 6 (1968), pp. 681-697.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 1, January 1986

AMBUSH STRATEGIES IN SEARCH GAMES ON GRAPHS*

STEVE ALPERN† AND MIROSLAV ASIC‡

Abstract. A blind searcher and a blind hider move at below unit speed along a finite length graph Q known to both, until the first time T when they meet. A two person zero-sum game arises if the searcher pays the hider T units. We consider circumstances under which it may be optimal for the searcher to "lie in wait" at a node of Q, hoping the hider will come to him. We also explicitly define a notion of "equilibrium in distribution" for such games, which has been implicit in the literature. We show that for the graph consisting of two nodes connected by three arcs of equal length there are optimal ambush strategies but there is no equilibrium in distribution.

Key words. search, game, graph, zero-sum

1. Introduction. Let Q be a graph with known arc lengths and a distinguished point q_0. A search (or "hide and seek" or "Princess and Monster" [8]) game may be played on Q as follows. The searcher starts at q_0 at time 0 and moves about Q at unit speed until the first (capture) time T that he coincides with the hider. Meanwhile the hider may move in any continuous way about Q. Both players know Q and q_0 but they cannot see each other. This scheme defines a two person zero-sum game with the capture time T (or an increasing function $U(T)$) as the payoff to the maximizing hider. The existence of the minimax value $V = V(Q, q_0)$, and of optimal searcher strategies and ε-optimal hider strategies, has been recently established by S. Gal [7, Appendix 1] using Ky Fan's minimax theorem. The existence question is also dealt with in [3], [4].

Search games have been analyzed for several specific graphs but up to now solutions (optimal strategies and V) have been obtained only by imposing an additional "nonloitering" assumption which prevents the searcher from waiting at any node. The "nonloitering" assumption is pragmatic in that it seems to make solutions easier, indeed possible. A better justification is that in some cases [5], [6] it gives an accurate model when graphs are used to analyze search games in multidimensional regions, where there is no analogue of waiting at a node. However it cannot be denied that in analyzing search games on graphs per se, an essential question is at which nodes, when, and how long the searcher should wait "in ambush." This is a problem which should not be assumed away.

In this paper we present the first analysis of a search game on a graph which obtains a solution *without* the "nonloitering" assumption. The graph we study consists of two nodes (one is q_0) joined by three unit length arcs (Fig. 1).

Our analysis of the search game on this graph has another feature which also distinguishes it from all previously studied search games. For those games [7, p. 51] "... there exists a function $P(t)$, which decreases exponentially in t, such that for all t both the searcher and the hider can keep the probability of capture after time t around $P(t)$. ... we use this property to show that optimal strategies for $U(T) = T$ are still optimal for more general cost functions $U(T)$." This important property, which we formalize in § 2 as an "equilibrium in distribution" does *not* hold for the search game on the "three arc" graph and hence the strategies we find optimal there when

* Received by the editors October 17, 1983, and in revised form July 31, 1984.
† London School of Economics, Houghton Street, London WC2A 2AE, England.
‡ University of Belgrade, Belgrade, Yugoslavia.

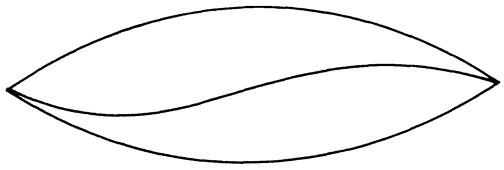


FIG. 1. *"Three arc" graph.*

$U(T) = T$ are not uniformly optimal for all U. It is probably the lack of an equilibrium in distribution for this game that makes our solution so technically difficult.

The questions dealt with here were inspired by suggestions in Gal's book [7] and in fact our initial results were obtained in the course of writing a review [2] of that book. Some of our results were first announced in that review and the details have appeared in the working paper [3].

The paper is organized as follows: Section 2 gives the formal definition of a search game on a graph and includes our new definition of an "equilibrium in distribution." Section 3 gives an analysis of ambush strategies on the k-arc graph. Section 4 gives a complete solution of the search game on three arcs.

2. Formal definition of a search game. In this section we describe the formal game theoretic background we will need to solve the "k-arc" games. We closely follow the presentation and notation of Gal's book, "Search Games" [7]. A search game Γ will be described by specifying a search space Q, pure strategy sets $\mathcal{S}$ and $\mathcal{H}$, and a nondecreasing utility function U. In this paper we will always take Q to be a graph with its shortest path metric d, although other metric spaces may be used. The pure searcher ($\mathcal{S}$) and hider ($\mathcal{H}$) strategy sets are given subsets of the space Cont(Q) consisting of all continuous functions from the nonnegative real numbers into Q. We endow $\mathcal{S}$ and $\mathcal{H}$ with the relative topology coming from the topology of "uniform convergence on compact subsets" on Cont(Q). It is customary to assume that the searcher has unit maximal speed, or that $\mathcal{S} \subset \mathcal{L}$ where

$$\mathcal{L} = \{f \in \text{Cont}(Q): d(f(t_1), f(t_2)) \leqq |t_1 - t_2| \text{ for all } t_1, t_2\}.$$

To each pair of pure strategies S in $\mathcal{S}$ and H in $\mathcal{H}$ (capital letters will denote pure strategies) there corresponds a unique "capture time" $T = T(S, H)$ defined as the least t such that $S(t) = H(t)$. If no such t exists then we define T to be infinity. We then define Γ to be the two person zero-sum game with the payoff to the maximizing hider given by $C(S, H) = U(T(S, H))$. More generally we allow the players to choose mixed strategies. Let $\mathcal{S}^*$ and $\mathcal{H}^*$ denote respectively the spaces of all regular Borel probability measures on $\mathcal{S}$ and $\mathcal{H}$. If the players choose "mixed strategies" s in $\mathcal{S}^*$ and h in $\mathcal{H}^*$ (lower case for mixed strategies) then the expected payoff to the maximizing hider is

$$c(s, h) = \int C(S, H)\, d(s \times h).$$

If

$$\inf_{s \in \mathcal{S}^*} \sup_{h \in \mathcal{H}^*} c(s, h) = \sup_{h \in \mathcal{H}^*} \inf_{s \in \mathcal{S}^*} c(s, h),$$

then we call this common number the value of the game and denote it by V, or by $V(U)$ if we wish to emphasize its dependence on U for fixed Q, $\mathcal{S}$ and $\mathcal{H}$. Gal [4] has applied Ky Fan's Minimax Theorem to establish sufficient conditions for the existence of a value for search games.

GAL'S THEOREM. *Suppose* $\Gamma = \Gamma(Q, \mathcal{S}, \mathcal{H}, U)$ *is a search game satisfying*

(i) $\mathcal{S}$ *is compact and*

(ii) U *is left continuous.*

Then Γ *has a value.*

Condition (i) will be satisfied in practice by taking $\mathcal{S}$ to be a closed subset of $\mathcal{L}$, since any such set is compact by the Arzela–Ascoli theorem. In order to find explicit solutions to certain search games on graphs, it has been necessary to restrict the searcher to pure strategies which are "nonloitering." Formally S is nonloitering if $\{t: S(t)$ is a node of Q (of degree $> 2)\}$ has Lebesgue measure zero. The anomaly is that this assumption, which in practice helps us to find the value, at the same time removes the condition (compactness of $\mathcal{S}$) which theoretically enables us to assert the existence of the value. Specifically the nonloitering strategies in $\mathcal{L}$ do not form a closed subset of $\mathcal{L}$.

Condition (ii) above is our replacement for Gal's requirement that $C(S, H)$ be lower semicontinuous in both variables. However, our condition implies his because $T(S, H)$ is lower semicontinuous and a nondecreasing left continuous function of a lower semicontinuous function is lower semicontinuous.

Gal's theorem allows for the possibility that the value is infinite although for combinatorially finite graphs the value is always finite (see [4, § 3]).

In estimating the function $c(s, h)$ for specific graphs it is helpful to use the formulae

$$c(s, h) = \int_0^\infty U(t)\, dF(s, h)(t) \quad \text{and}$$

$$c(s, h) = \int_0^\infty (1 - F(s, h)(t))\, dt \quad \text{when } U = I$$

$(I(t) \equiv t)$ where $F(s, h)$ is the cumulative distribution of capture times corresponding to the mixed strategies s and h,

$$F(s, h)(t) = \Pr(T \leqq t) = (s \times h)\{(S, H): T(S, H) \leqq t\}.$$

For any cumulative probability distribution F we write $U(F) = \int_0^\infty U(t)\, dF(t)$, which is the payoff corresponding to that distribution. For s in $\mathcal{S}^*$ and h in $\mathcal{H}^*$ we define

$$[F^-(s)](t) = \inf_h F(s, h)(t) \quad \text{and}$$

$$[F^+(h)](t) = \sup_s F(s, h)(t).$$

Intuitively $F^-(s)$ is the distribution the searcher can guarantee by playing s, and similarly for $F^+(h)$. An estimate on the distribution of T yields an estimate on the expected value of $U(T)$ so that if the value exists then we always have:

$$\sup_{h \in \mathcal{H}^*} U(F^+(h)) \leqq V(U) \leqq \inf_{s \in \mathcal{S}^*} U(F^-(s)).$$

If the left and right side of this inequality are equal, then we will say the game $\Gamma(Q, \mathcal{S}, \mathcal{H}, U)$ has an "equilibrium in distribution," which is a strictly stronger property than merely having a value. This notion has appeared in the search game literature as early as [1] but we make it precise for the first time here. All the search games which have been explicitly solved prior to this paper have equilibria in distribution. However the search game on three arcs solved in § 4 does not have an equilibrium in distribution and for that game, $\sup_h U(F^+(h)) < V(U) = \inf_s U(F^-(s))$ in the case that $U = I$.

Associated with *every* game $\Gamma(Q, \mathcal{S}, \mathcal{H}, U)$ where $\mathcal{S}$ is compact is a distribution E of capture times which we call the "equilibrium distribution." For all t this distribution satisfies $\sup_s [F^-(s)](t) = E(t) = \inf_h [F^+(h)](t)$. The existence of each number $E(t)$ follows from applying Gal's theorem to the game $\Gamma(Q, \mathcal{S}, \mathcal{H}, U_t)$, where U_t is the characteristic function of (t, ∞), and setting $E(t)$ equal to the value. We emphasize that E does not depend on the utility function U and that its existence is more general than that of an equilibrium in distribution for the game.

Suppose we can find mixed strategies s in $\mathcal{S}^*$ and h_m in $\mathcal{H}^*$, $m = 1, 2, 3, \ldots$, such that $F^-(s) = E$ and $F^+(h_m)$ converges in distribution to E. (The latter assumption means that $[F^+(h_m)](t)$ converges to $E(t)$ at all continuity points t for E.) It then follows for all continuous U for which $U(E)$ is finite that

$$U(F^-(s)) = U(E) = \lim_{m \to \infty} U(F^+(h_m)),$$

or that $\Gamma(Q, \mathcal{S}, \mathcal{H}, U)$ has an equilibrium in distribution and value $U(E) = \int_0^\infty U\, dE(t)$. We call the situation described in this paragraph a "uniform equilibrium in distribution."

To illustrate the above ideas on equilibria in distribution, we consider the simple class of games Γ_a played on the closed unit interval $[0, 1]$ with the searcher starting at the point a. By symmetry we may assume that $0 \leqq a \leqq \frac{1}{2}$. Every hider pure strategy is dominated by one of the immobile strategies H_i, $i = 0, 1$, which stay at i forever. So we restrict our attention to the mixtures $h_p = pH_0 + (1-p)H_1$ for $0 \leqq p \leqq 1$. Similarly let S_i, $i = 0, 1$, be the pure strategy of going at unit speed first to end i and then to end $1-i$, and let $s_q = qS_0 + (1-q)S_1$.

In the symmetric case $a = \frac{1}{2}$ we have $F^-(s_{1/2}) = F^+(h_{1/2}) = E$, where $E(t)$ is respectively 0, $\frac{1}{2}$ and 1 on the intervals $[0, \frac{1}{2})$ $[\frac{1}{2}, \frac{3}{2})$, and $[\frac{3}{2}, \infty)$. Hence $\Gamma_{1/2}$ has a uniform equilibrium in distribution. In fact for any utility U we have $U(F^-(s_{1/2})) = U(E) = \frac{1}{2}U(\frac{1}{2}) + \frac{1}{2}U(\frac{3}{2}) = U(F^+(h_{1/2}))$.

If $0 < a < \frac{1}{2}$ then for $0 \leqq t < 1 - a$ the hider can ensure $E(t) = 0$ by using H_1 (waiting at the further end). For $1 - a \leqq t < 1 + a$ the strategies $s_{1/2}$ and $h_{1/2}$ ensure that $E(t) = \frac{1}{2}$. For $t \geqq 1 + a$ the searcher can ensure that $E(t) = 1$ by using S_0. The thing to observe is that for both players the strategies needed to optimize $E(t)$ depend on t, so there is not a uniform equilibrium in distribution. If the utility function U has zero variation outside one of the three intervals determined by $1 - a$ and $1 + a$, then the strategies given above for that interval are optimal and produce an equilibrium in distribution. Suppose however that this is not true of U, for example if $U(T) = 0$ for $T < \frac{1}{2}$, $U(T) = 1$ for $\frac{1}{2} \leqq T < 1$, and $U(T) = 2$ for $1 \leqq T$. This corresponds to the hider getting one utility unit if uncaught at time $\frac{1}{2}$ and another unit if still uncaught at time 1. The payoff matrix in this case is given by

	H_0	H_1
S_0	0	2
S_1	2	1

.

It is easily computed that the value V is $\frac{4}{3}$ and that the unique optimal strategies are $s_{1/3}$ and $h_{1/3}$. To check that this game does not have an equilibrium in distribution, consider the conservative distributions $F^+(h_p)$, $0 \leqq p \leqq 1$. Observe that $(F^+(h_p))(\frac{1}{2}) = p$ with the supremum achieved by S_0, and $(F^+(h_p))(1) = \max[p, 1-p]$ with the supremum achieved by $S_{1-[2p]}$. Since for this U we have $U(F^+(h)) = (1 - (F^+(h))(\frac{1}{2})) +$

$(1-(F^+(h))(1))$ it follows that $U(F^+(h_p))=(1-p)+(1-\max[p,1-p])$. Therefore $\sup_h U(F^+(h))=\sup_p (2-p-\max[p,1-p])=1$ which is strictly less than the value $V(U)=\frac{4}{3}$, and $\Gamma_a(U)$ does not have an equilibrium in distribution. For the sake of completeness we conclude by observing that in the trivial end-starting game Γ_0 there is a uniform equilibrium in distribution and $V(U)=U(1)$.

3. Searcher ambush strategies for the k-arc game. Let Γ_k denote the search game on the graph consisting of k unit length arcs between two nodes A (searcher starting point) and B. From this point onwards the utility function U is identical with the capture time T. S. Gal [5], [6] has shown that the nonloitering version of this game has a uniform equilibrium in distribution with $E(t)=1-((k-1)/k)^{[t]}$ and hence a nonloitering value V^* equal to k. For this nonloitering game he established that the searcher "random oscillation" strategy s^* of going back and forth between A and B at unit speed with equiprobable and independent choices among the k arcs satisfies $F^-(s^*)=E$ and is therefore optimal. An ε-optimal hider mixed strategy $h^*(\varepsilon)$ with conservative distribution $F^+(h^*(\varepsilon))$ approaching (as $\varepsilon\to 0$) E is to wait at B until time $1-\varepsilon$ and then use a random oscillation.

In this section we implicitly (using Gal's Minimax Theorem) define an ambush mixed strategy $\hat{s}=\hat{s}_k$ for the searcher which waits at A for $0\leqq t\leqq 2$ with positive probability. Our analysis (Lemma 1) of $\hat{s}$ shows that for $k>3$ (more than three arcs) the searcher can improve by loitering and that the (unrestricted) value V_k satisfies $V_k<V_k^*=k$, where V^* denotes nonloitering value. For $k=3$ our analysis gives that $V_3\leqq 3$ which shows that there are optimal search strategies which loiter but leaves open the question (settled negatively in section four) of whether the searcher can strictly improve his performance in the three arc game by loitering.

LEMMA 1. *The (unrestricted) value V_k of the search game on k arcs satisfies $V_k\leqq(2k+p_k-1)/(2-2p_k)$ where $p_k=(k^2-k+1)^{-1}$. In particular $V_3\leqq 3$ and $V_k<k$ (k is the nonloitering value) for $k>3$.*

Proof. Consider the following searcher mixed strategy $\hat{s}$ (which of course depends on k): In the first two units of time either wait at the starting point A, with probability p_k; or randomly pick two distinct arcs and take the first to B and the second back to A, each at unit speed, with probability $1-p_k$. In either case $\hat{s}(2)=A$. Starting at time 2, pretend the game is just beginning and play any fixed optimal strategy—which exists by Gal's Theorem.

To estimate $c(\hat{s},H)$ we first partition $\mathcal{H}=\mathcal{H}_1\cup\mathcal{H}_2$ where $H\in\mathcal{H}_1$ if and only if $H(t)=A$ for some t with $0\leqq t\leqq 2$. Observe that for any $H\in\mathcal{H}$

$$1-F(\hat{s},H)(1)=\text{Prob}\,(T>1)\leqq(1)(p_k)+((k-1)/k)(1-p_k).$$

Now suppose $H_1\in\mathcal{H}_1$ and consider Prob $(T>2)$, that is, $1-F(\hat{s},H_1)(2)$. If the searcher was waiting at A, then Prob $(T>2)=0$. On the other hand, if the searcher searched distinct arcs, then the best case for the hider occurs when he is on distinct arcs at $t=1$ and $t=2$. In this case the probability that $T>2$ is given by $(k^2-3k+3)/(k^2-k)$ which is the probability that two independently chosen, random distinct pairs of numbers from $\{1,\cdots,k\}$ disagree in both coordinates. Therefore for all $H_1\in\mathcal{H}_1$ we have

$$1-F(\hat{s},H_1)(2)=\text{Prob}\,(T>2)\leqq(0)(p_k)+((k^2-3k+3)/(k^2-k))(1-p_k).$$

If $H_2\in\mathcal{H}_2$ then

$$1-F(\hat{s},H_2)(2)=\text{Prob}\,(T>2)\leqq(1)(p_k)+((k-2)/k)(1-p_k).$$

We have chosen p_k so that our estimates on Prob $(T>2)$ are the same for $\mathcal{H}_1$ and $\mathcal{H}_2$.

For any $H \in \mathcal{H}$ we may estimate $c(\hat{s}, H)$ by the formula

$$c(\hat{s}, H) \leqq 1 + 1(1 - F(\hat{s}, H)(1)) + V_k(1 - F(\hat{s}, H)(2))$$

because if $T > 2$ then since the searcher is playing optimally from time 2 the expected value of $T-2$ is no more than V_k. Thus by separately estimating $c(\hat{s}, H)$ for $H \in \mathcal{H}_1$ and $H \in \mathcal{H}_2$ by our above formulae, and obtaining the same result in each case (by choice of p_k), we obtain for all H in $\mathcal{H}$ that

$$c(\hat{s}, H) \leqq 1 + [p_k + ((k-1)/k)(1-p_k)] + [(p_k + (1-p_k)((k-2)/k))V_k].$$

It now follows from the definition of V_k that

$$V_k \leqq 1 + [p_k + ((k-1)/k)(1-p_k)] + [(p_k + (1-p_k)((k-2)/k))V_k].$$

The desired estimate on V_k is now obtained by solving for V_k in the above inequality.

4. Counter-ambush hider strategies in the 3-arc game. In this section we give a complete solution, without employing the nonloitering assumption, to the search game on the three arc graph. We have already indicated (in § 3) two distinct searcher strategies, one nonloitering (s^*) and one ambush ($\hat{s}$), which ensure that the expected capture time will not exceed 3 (on average). The hider strategy $h^*(\varepsilon)$ which Gal showed to be effective (guarantees $T \geqq 3$) against all nonloitering strategies gives only $T = 2 - \varepsilon$ against the (admittedly bad) strategy S_A of waiting at A. However (Theorem 1) the hider still has strategies which guarantee that the expected capture time is not much less than 3, and hence $V = 3$, even if the searcher is allowed to loiter.

The basic idea for the hider in countering ambush strategies is simple. The random oscillation strategy $h^*(\varepsilon)$ sometimes (actually one third of the time) involves going to and leaving a node by the same arc. But this is foolish since now the searcher could be waiting at that node, so that nodes should be entered only when required to change arcs. So it is preferable for the hider to modify $h^*(\varepsilon)$ by choosing the sequence of arcs at say time zero and, when the same arc occurs consecutively, waiting a small distance δ from the intervening vertex for a period of length 2δ. At the end of that waiting period he should resume the oscillation strategy. For technical reasons we have not been able to establish directly that this modified strategy $h^*(\varepsilon, \delta)$ guarantees that $c(s, h^*(\varepsilon, \delta)) \geqq 3 - \varepsilon$ for all s. The problem is that once ε and δ (or their distributions) are known, it is perhaps possible for the searcher to search points δ-near the vertex in the hope of finding a waiting hider.

To get around this problem, we use Gal's Minimax Theorem and show that against any given searcher strategy s (in particular an optimal one) there is a hider mixed strategy similar to what we have described above as $h^*(\varepsilon, \delta)$ in which the distributions of ε and δ depend on the searcher strategy s. The dependence of ε on s is described in the measure theoretic Lemma 3. This and the other technical result, Lemma 2, are perhaps best left until after Theorem 1, which will motivate their use.

Our final result, Corollary 1, shows that while both the nonloitering and unrestricted version of the 3-arc game both have value 3, only the former has an equilibrium in distribution.

LEMMA 2. *Let* $X = \{x = (x_1, x_2, \cdots) : x_1 = 0$ *and for all* i, $x_i = 0$ *or* 1 *and* $x_i x_{i+1} = 0\}$. *Let* $\phi : X \to R$ *be defined by* $\phi(x) = \sum_{n=1}^{\infty} \prod_{i=1}^{n} r_i(x)$, *where* $r_i(x) = (\frac{1}{3})^{x_i}(\frac{2}{3})^{1-x_i-x_{i+1}}$. *Then for all* x *in* X, $\phi(x) = 2$.

Proof. We will define vectors $e(q)$ in X, $0 \leqq q \leqq \infty$, such that $\phi(e(q)) = 2$ for all q, and such that for all x in X, $\phi(x) = \phi(e(q))$ for some q.

If $q=0,1,2,\cdots$, let $e=e(q)$ be defined by $e_1=e_2=\cdots=e_q=0$ and for $m=1,2,\cdots$; $e_{q+2m}=1$ and $e_{q+2m-1}=0$. For example $e(3)=(0,0,0,0,1,0,1,0,1,\cdots)$. We define $e(\infty)=(0,0,\cdots)$.

Fix any finite q and let $r_i=r_i(e(q))$. Then $r_i=\frac{2}{3}$ for $1\leqq i\leqq q$ and for $m=1,2,\cdots$ we have $r_{q+2m}=\frac{1}{3}$ and $r_{q+2m-1}=1$. It follows that $\prod_{i=1}^{n} r_i=(\frac{2}{3})^n$ for $n=1,\cdots,q$ and that for $m=0,1,2,\cdots$ we have

$$\prod_{i=1}^{q+2m} r_i=(\tfrac{2}{3})^q(\tfrac{1}{3})^m=\prod_{i=1}^{q+2m+1} r_i.$$

Therefore

$$\begin{aligned}\sum_{n=1}^{\infty}\prod_{i=1}^{n} r_i&=\sum_{n=1}^{q-1}\prod_{i=1}^{n} r_i+\sum_{m=0}^{\infty}\left(\prod_{i=1}^{q+2m} r_i+\prod_{i=1}^{q+2m+1} r_i\right)\\&=\sum_{n=1}^{q-1}(\tfrac{2}{3})^n+2\sum_{m=0}^{\infty}(\tfrac{2}{3})^q(\tfrac{1}{3})^m\\&=\frac{(\frac{2}{3})-(\frac{2}{3})^q}{(\frac{1}{3})}+2(\tfrac{2}{3})^q/(\tfrac{2}{3})\\&=2-3(\tfrac{2}{3})^q+3(\tfrac{2}{3})^q=2.\end{aligned}$$

Thus $\phi(e(q))=2$ for $q<\infty$. For $q=\infty$ we have

$$\phi(e(\infty))=\phi((0,0,\cdots))=\sum_{n=1}^{\infty}\prod_{i=1}^{n}(\tfrac{2}{3})=2.$$

To show that for all x in X, $\phi(x)=\phi(e(q))$ for some q, we introduce a map $g\colon X\to X$ such that ϕ is constant on orbits of g and such that each orbit converges to some $e(q)$, with respect to the metric $\rho(x,y)=\sum 2^{-i}|x_i-y_i|$ on X. Since the function $\phi\colon X\to R$ is continuous with respect to ρ, this will imply that $\phi(x)=\lim\phi(g^j(x))=\phi(e(q))=2$.

Define the map $g=X\to X$ by $\bar{x}=g(x)$, where $\bar{x}$ is determined as follows. If $x=e(q)$ for some q, set $\bar{x}=x$. Otherwise there is a least m such that $x_m=1$, $x_{m+1}=0$ and $x_{m+2}=0$. Define $\bar{x}$ by $\bar{x}_m=0$, $\bar{x}_{m+1}=1$ and $\bar{x}_i=x_i$ for all other i. For example $g((0,0,0,1,0,1,0,0,\cdots))=(0,0,0,1,0,0,1,0,\cdots)$. To establish that $\phi(g(x))=\phi(x)$, fix any x in X and write $\bar{x}=g(x)$. Let $r_i=r_i(x)$ and $\bar{r}_i=r_i(\bar{x})$. Observe that for i not equal to $m-1$, m, or $m+1$ we have $\bar{r}_i=r_i$, because r_i depends only on x_i and x_{i+1}. For the remaining three indices we have $r_{m-1}=1$, $r_m=\frac{1}{3}$, $r_{m+1}=\frac{2}{3}$, $\bar{r}_{m-1}=\frac{2}{3}$, $\bar{r}_m=1$ and $\bar{r}_{m+1}=\frac{1}{3}$. Observe that $r_{m-1}r_mr_{m+1}=\bar{r}_{m-1}\bar{r}_m\bar{r}_{m+1}$ so that $\prod_{i=1}^{n} r_i=\prod_{i=1}^{n}\bar{r}_i$ for n not equal to $m-1$ or m. Next compute

$$\prod_{i=1}^{m-1} r_i+\prod_{i=1}^{m} r_i=\left(\prod_{i=1}^{m-2} r_i\right)(r_{m-1}+r_{m-1}r_m),$$

$$\begin{aligned}\prod_{i=1}^{m-1}\bar{r}_i+\prod_{i=1}^{m}\bar{r}_i&=\left(\prod_{i=1}^{m-2}\bar{r}_i\right)(\bar{r}_{m-1}+\bar{r}_{m-1}\bar{r}_m)\\&=\left(\prod_{i=1}^{m-2} r_i\right)(\bar{r}_{m-1}+\bar{r}_{m-1}\bar{r}_m).\end{aligned}$$

Since $r_{m-1}+r_{m-1}r_m=\bar{r}_{m-1}+\bar{r}_{m-1}\bar{r}_m$ it follows from the above calculations that

$$\sum_{n=m-1}^{m}\prod_{i=1}^{n} r_i=\sum_{n=m-1}^{m}\prod_{i=1}^{n}\bar{r}_i.$$

Hence

$$\phi(g(x)) = \sum_{n=1}^{\infty} \prod_{i=1}^{n} \bar{r}_i = \left(\sum_{n=1}^{m-2} + \sum_{n=m-1}^{m} + \sum_{n=m+1}^{\infty}\right)\left(\prod_{i=1}^{n} \bar{r}_i\right)$$

$$= \left(\sum_{n=1}^{m-2} + \sum_{n=m-1}^{m} + \sum_{n=m+1}^{\infty}\right)\left(\prod_{i=1}^{n} r_i\right)$$

$$= \phi(x).$$

Finally, we must show that $\lim g^j(x)$ is always one of the $e(q)$. Let $q_n(x)$ be the number of ones in the first $2n$ coordinates of x. Clearly $n - q_n(x)$ is a nondecreasing sequence of natural numbers with a limit (possibly ∞) which we denote $q(x)$. If $q = q(x)$ is finite, then the sequence $g^j(x)$ is eventually equal to $e(\bar{q})$ where $\bar{q} = 2q$ or $2q-1$. If $q(x)$ is infinite then $g^j(x)$ converges to $e(\infty) = (0, 0, \cdots)$. Hence for all x in X,

$$\phi(x) = \phi\left(\lim_{j\to\infty} g^j(x)\right) = \phi(e(q(x))) = 2.$$

LEMMA 3. *Given any mixed strategy s in $\mathcal{S}^*$ there exist arbitrarily small positive numbers ε such that the following condition holds for s-a.e. pure strategies S in $\mathcal{S}$:*

(4.1) *If for some positive integer m, $S(m-\varepsilon)$ is a vertex, then for all δ with $0 < |\delta| \leqq 1$, $d(S(m-\varepsilon), S(m-\varepsilon+\delta)) < |\delta|$.*

Proof. For any S in $\mathcal{S}$ and any positive integers m and n, let $J(S, m, n)$ be the subset of $[0, 1]$ consisting of all ε such that $S(m-\varepsilon)$ is a vertex (A or B) and $d(S(m-\varepsilon), S(m-\varepsilon \pm 1/n)) = 1/n$. Clearly the distance between distinct points of $J(S, m, n)$ is at least $2/n$ so certainly $J(S, m, n)$ is finite. Consequently the set $J(S) = \cup_{m,n} J(S, m, n)$ is countable and so $\mu(J(S)) = 0$, where μ denotes Lebesgue measure on $[0, 1]$. Let F be the subset of $[0, 1] \times \mathcal{S}$ defined by letting (ε, S) belong to F if and only if ε belongs to $J(S)$. Observe that (ε, S) belongs to F exactly for those pairs which fail to satisfy condition (4.1). Applying Fubini's theorem to the characteristic function f of the set F yields

$$\int_0^1 \int_{\mathcal{S}} f(\varepsilon, S)\, ds\, d\mu = \int_{\mathcal{S}} \int_0^1 f(\varepsilon, S)\, d\mu\, ds = \int_{\mathcal{S}} \mu(J(S))\, ds = 0.$$

Hence for μ-almost all ε in $[0, 1]$ we must have $\int_{\mathcal{S}} f(\varepsilon, S)\, ds = 0$, or $s(\{S: \varepsilon \text{ belongs to } J(S)\}) = 0$ completing the proof.

THEOREM 1. *For the search game on three arcs (without the nonloitering assumption) the value is* 3.

Proof. The existence of a value V and an optimal searcher mixed strategy $\bar{s}$ is guaranteed by Gal's theorem [7, Appendix 1]. It was shown in Lemma 1 that $V \leqq 3$. We will now show that given any ε^* we can find a hider mixed strategy $\bar{h} = \bar{h}(\bar{s}, \varepsilon^*)$ such that $c(\bar{s}, \bar{h}) \geqq 3 - \varepsilon^*$. Since $\bar{s}$ is assumed optimal, we will then have $V \geqq c(\bar{s}, \bar{h})$, or $V \geqq 3 - \varepsilon^*$.

The hider strategy $\bar{h}$ is defined as follows. Choose a small positive ε which satisfies condition (4.1) for $\bar{s}$-almost all searcher pure strategies S. Lemma 3 ensures that this choice can always be made with arbitrarily small ε. Denote $t_i = i - \varepsilon$ for all integers i and let Z_i denote either A or B respectively, depending on whether i is odd or even. Let $\delta_1 = 0$ and for $i = 2, 3, \cdots$ choose $\delta_i > 0$ sufficiently small so that

$$\bar{s}(\{S: 0 < d(S(t_i), Z_i) \leqq 2\delta_i\}) < \varepsilon 2^{-i}. \tag{4.2}$$

This is always possible since for fixed i the sets with say $\delta_i = 1/n$ are decreasing and have empty intersection. To choose a pure hider strategy $\bar{H}$ according to $\bar{h}$, proceed

as follows (this procedure defines $\bar{h}$). First pick a pure strategy H using the oscillation strategy $h^*(\varepsilon)$ of staying at B until time $1-\varepsilon$ and then oscillating randomly as described in § 3. Let $t_i^- = t_i - \delta_i$ and $t_i^+ = t_i + \delta_i$. Define $\bar{H}$, given H, by $\bar{H}(t) = H(t_i^-) = H(t_i^+)$ for $t_i^- \leqq t \leqq t_i^+$ if $H(t_i^-) = H(t_i^+)$ and $\bar{H}(t) = H(t)$ otherwise. This is perhaps easier in words: If the same arc is chosen for the transit just before t_i and for just after t_i, then do not go all the way to $Z_i = H(t_i)$ but instead wait at distance δ_i from Z_i during the interval $t_i^- \leqq t \leqq t_i^+$.

Let $\hat{\mathcal{S}}$ denote the subset of $\mathcal{S}$ consisting of all S satisfying condition (4.1) with respect to the ε chosen above, and such that

(4.3) For all i, either $S(t_i) = Z_i$ or $d(S(t_i), Z_i) > 2\delta_i$.

Our choices of ε and the δ_i ensure that

(4.4) $$s(\hat{\mathcal{S}}) > 1 - \varepsilon/2.$$

We will show that

(4.5) $$c(S, \bar{h}) \geqq 3 - 2\varepsilon \quad \text{for all } S \text{ in } \hat{\mathcal{S}}.$$

It will then follow from (4.4) and (4.5) that $V \geqq c(S, \bar{h}) \geqq (\varepsilon/2)(0) + (1-\varepsilon/2) \times (3-2\varepsilon)$ for all S in $\mathcal{S}$. By choosing ε sufficiently small we thus have $V \geqq 3 - \varepsilon^*$. It remains only to demonstrate (4.5).

Fix any S in $\hat{\mathcal{S}}$ and define a vector $x = (x_1, x_2, \cdots)$, which depends on S, by $x_i = 1$ if $S(t_i) = Z_i$ and $x_i = 0$ otherwise. Observe that $x_1 = 0$ and that

(4.6) $$x_i x_{i+1} = 0 \quad \text{for all } i$$

because of condition (4.1) with $\delta = 1$. We shall use the notation Pr ( ) to denote the probability of various events assuming that the searcher is using the pure strategy S and the hider is employing the mixed strategy $\bar{h}$. In particular we will need the following two important equations, which we justify below.

(4.7) $$\Pr(T > t_{i+1}^- / T > t_i) = (\tfrac{2}{3})^{1-x_i-x_{i+1}},$$

(4.8) $$\Pr(T > t_i^+ / T > t_i^-) = (\tfrac{1}{3})^{x_i}.$$

Since the hider always chooses his next arc equiprobably out of the three choices, and independently of his current arc, the respective probabilities of switching arcs (and thus going through a vertex) and staying on the same arc (avoiding the vertex) are $\frac{2}{3}$ and $\frac{1}{3}$. To justify equation (4.7), suppose the hider is "alive" at time t_i. Since S satisfies (4.1) and (4.3) the hider will certainly still be alive at time t_i^+, i.e. $T > t_i^+$. Of the three possible paths that the hider may take from t_i^+ to t_{i+1} at most one will intersect S. If neither $S(t_i) = Z_i$ nor $S(t_{i+1}) = Z_{i+1}$, then exactly one will intersect S (otherwise none will intersect S). In the former case $\Pr(T > t_{i+1}^- / T > t_i^-)$ is equal to $\frac{2}{3}$; in the latter it is equal to 1. To establish (4.8), simply observe that condition (4.3) guarantees that the only way the hider can be caught between times t_i^- and t_i^+ is if the hider switches arcs at Z_i at time t_i (which has probability $\frac{2}{3}$) and the searcher is also at Z_i at time t_i (in which case $x_i = 1$). Equations (4.7) and (4.8) may be combined to give

(4.9) $$\Pr(T > t_{i+1}^- / T > t_i^-) = (\tfrac{1}{3})^{x_i} (\tfrac{2}{3})^{1-x_i-x_{i+1}}.$$

We also make the obvious remark,

(4.10) $$\Pr(T > t_1^-) = 1.$$

Let $p_i = \Pr(T > t_i^-)$ so that $p_1 = 1$ and $p_{i+1} = p_i r_i$, where $r_i = r_i(x) = (\frac{1}{3})^{x_i}(\frac{2}{3})^{1-x_i-x_{i+1}}$. It follows that $p_i = r_1 r_2 \cdots r_{i-1}$. We may estimate

$$
\begin{aligned}
c(S, \bar{h}) &\geqq \sum_{n=1}^{\infty} t_n^- \Pr(t_n^- < T \leqq t_{n+1}) \\
&= \sum_{n=1}^{\infty} (n - \varepsilon - \delta_n)(p_n - p_{n+1}) \qquad (4.11) \\
&\geqq (1 - \varepsilon - \delta_2) + \sum_{n=2}^{\infty} p_n \geqq (1 - 2\varepsilon) + \sum_{n=2}^{\infty} p_n.
\end{aligned}
$$

But

$$
\sum_{n=2}^{\infty} p_n(x) = \sum_{n=2}^{\infty} \prod_{i=1}^{n-1} r_i(x) = \sum_{n=1}^{\infty} \prod_{i=1}^{n} r_i(x) = \phi(x),
$$

where ϕ is defined as in Lemma 2, and hence $\phi(x) = 2$. Thus (4.11) implies (4.5) completing the proof.

COROLLARY 1. *The search game on three arcs (without the nonloitering assumption) does not have an equilibrium in distribution.*

Proof. Since $V = 3$ (Theorem 1) it is sufficient to find a $\delta > 0$ such that for any h in $\mathcal{H}^*$, $U(F^+(h)) \leqq 3 - \delta$. It follows from the definitions that for any h in $\mathcal{H}^*$, $(F^+(h))(t) \geqq \sup_s (F^-(s))(t)$ so that in particular $1 - (F^+(h))(t) \leqq \min[1 - (F^-(s^*))(t), 1 - (F^-(\hat{s}))(t)]$ where s^* is Gal's oscillation strategy (§ 3) and $\hat{s}$ is the strategy of Lemma 1. Recall that $1 - (F^-(s^*))(t) = (\frac{2}{3})^{[t]}$ so that $U(F^-(s^*)) = 3$. The estimate for $T > 2$ in Lemma 1 shows that for t in the interval $[2, 3)$, $1 - (F^-(\hat{s}))(t) = \frac{3}{7}$, while $1 - (F^-(s^*))(t) = \frac{4}{9}$. Therefore

$$
\begin{aligned}
U(F^+(h)) &= \int_0^{\infty} 1 - (F^+(h))(t)\, dt \leqq \int_0^{\infty} \min[1 - (F^-(s^*))(t), 1 - (F^-(\hat{s}))(t)]\, dt \\
&\leqq \int_0^{\infty} [1 - (F^-(s^*))(t)]\, dt - \int_2^3 [(F^-(\hat{s}))(t) - (F^-(s^*))(t)]\, dt \\
&= 3 - (\tfrac{4}{7} - \tfrac{5}{9}) = 3 - \tfrac{1}{63}.
\end{aligned}
$$

REFERENCES

[1] S. ALPERN, *The search game with mobile hider on the circle*, in Differential Games and Control Theory, E. Roxin, P. T. Liu and R. L. Sternberg, eds., Marcel Dekker, New York, 1974, pp. 181-200.

[2] S. ALPERN AND M. ASIC, *Review* of [7], SIAM Rev., 24 (1982), pp. 235-236.

[3] ———, *Two search games on graphs*, Theoretical Economics Discussion Paper Series, International Center for Economics and Related Disciplines, London School of Economics, No. 60, 1982, pp. 1-40.

[4] ———, *The search value of a network*, Networks, 15 (1985), to appear.

[5] C. FITZGERALD, *The princess and monster differential game*, this Journal, 17 (1979), pp. 700-712.

[6] S. GAL, *Search games with mobile and immobile hider*, this Journal, 17 (1979), pp. 99-122.

[7] ———, *Search Games*, in Mathematics in Science and Engineering, Vol. 149, Academic Press, New York, 1980.

[8] R. ISAACS, *Differential Games*, John Wiley, New York, 1965.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 1, January 1986

003

LIMITING DISTRIBUTION FOR RANDOM OPTIMIZATION METHODS*

CHANG C. Y. DOREA†

Abstract. Let f be a function defined on some domain $\Omega \subset R^d$. We consider the problem of finding the global minimum of f subjected to some constraints, say $g_i(x) \leqq 0$, $i = 1, \cdots, m$. When differentiability is not assumed random optimization methods provide an alternative way to estimate the minimum. For two such methods we study the existence of the limiting distribution and the estimation of the parameter of the limiting distribution.

Key words. random optimization, limiting distribution, order statistics

1. Introduction. Let f be a measurable function defined on some domain $\Omega \subset R^d$, we like to find the global minimum of f subjected to some constraints represented by a measurable set A. That is, find l such that

$$l = \lim_{x \in A} f(x), \qquad A = \{g_i(x) \leqq 0, i = 1, \cdots, k\}. \tag{1}$$

The standard methods such as the gradient method or the steepest descent method, all require that f satisfies some differentiability property. In the cases where differentiability is not assumed random optimization methods could well be applied. Consider the random procedures described below.

Method (A). Let $\eta^1, \eta^2, \cdots$ be iid random vectors with a common distribution G. Let $X_0, X_1, \cdots$ be defined by

step 0. $X_0 = f(U)$, where U is uniformly distributed over A.

step $k+1$. Having defined X_k, let X_{k+1} be defined by

(a1) $X_{k+1} = f(\eta^{k+1})$, if $\eta^{k+1} \in A$ and $f(\eta^{k+1}) < f(\eta^k)$.

(a2) $X_{k+1} = X_k$, otherwise.

Method (B). Let $\xi^1, \xi^2, \cdots$ be iid random vectors with a common distribution H. Let $W^0, W^1, W^2, \cdots$ be defined by

step 0. $W^0 = U$, where U is uniformly distributed over A.

step $k+1$. Having defined W^k, let W^{k+1} be defined by

(b1) $W^{k+1} = W^k + \xi^{k+1}$, if $(W^k + \xi^{k+1}) \in A$.

(b2) $W^{k+1} = W^k$, otherwise

For $n \geqq 0$ define $Y_n = \min\{f(W^0), \cdots, f(W^n)\}$.

Note that the upper indices have been used to denote sequences of random vectors and the lower indices to denote sequences of random variables. For related results on Method (B) see Baba [1], Dorea [2] and Solis-Wets [6]. In fact, Method (B) can be viewed as a particular setting of the conceptual random algorithm considered by Solis-Wets [6].

Now a confidence statement for l can be made if the limiting distribution of X_n (or Y_n) properly normalized exists and the parameters of this limiting distribution can be estimated.

Notice that $X_n = \min\{X_0, \cdots, X_n\} \geqq \min\{f(U), f(\eta^1), \cdots, f(\eta^n)\}$ with equality holding if $P(A) = 1$. Since the X_j's are not independent the usual extreme value theory cannot be directly applied. Also W^n is not exactly a random vector whose components are partial sums of independent random variables or randomly stopped subsequences

* Received by the editors June 14, 1983, and in revised from May 15, 1984. This research was supported in part by CNPq 200607/82 and CNPq 301508/84.

† Departameto de Matematica, Universidade de Brasilia, 70910 Brasilia-DF, Brazil. This research was carried out while the author was at Iowa State University, Ames, Iowa 50011.

of partial sums. And again the results on the minimum of functions of partial sums cannot be used. In fact, W^n can be regarded as a partial sum of randomly selected r.v.'s

$$W^n = U + \sum_{i=1}^{n} \xi^{\sigma(i)}$$

where $\xi^0 = 0$, $\sigma(i) = i$ if $W^{i-1} + \xi^i \in A$, otherwise $\sigma(i) = 0$.

Thus our problem reduces to finding conditions on f and G (or H) that assure weak convergence under this dependence setting. In the next section it will be shown that if Condition (A) (Condition (B)) is satisfied then $X_n(Y_n)$ properly normalized converges weakly to a distribution of type $1 - b \exp(-x^\alpha)$, where $0 < b \leqq 1$ and $\alpha > 0$. Finally in § 3, following the approach of de Haan [4], we provide an asymptotic estimate of the parameter α based on the limiting behavior of the ratio of the differences of small order statistics.

2. Limiting distribution. Our problem is to find conditions under which there exist norming constants $a_n > 0$ that will assure the existence of a nondegenerate limiting distribution for $(X_n - l)/a_n$ (or $(Y_n - l)/a_n$). The introduction of norming constants is made necessary since in general we have convergence to the degenerate distribution, that is, $\lim_{n\to\infty} P(X_n = l) = 1$ (see Dorea [2]).

Let $A \subset R^d$ be a bounded and measurable set, say $A \subset [-L, L]^d$, and f a measurable function defined on $\Omega \supset A$. Let

Condition (A). (a1) $G = \pi G_i$ is a distribution such that $P(A) > 0$.

(a2) There exists a positive function $v(t)$, $t > 0$, with $\lim_{t\downarrow 0} v(t) = 0$ and $\alpha > 0$ such that for all $y > 0$ we have

$$\begin{aligned} &m(A_t(y)) \to y^\alpha \quad \text{as } t \downarrow 0, \\ &A_t(y) = \{u\colon tu \in (0,1)^d,\ G^*(tu) \in \bar{A},\ f(G^*(tu)) \leqq l + yv(t)\}. \end{aligned} \tag{2}$$

where $G = \pi G_i$ indicates that G is the product of its marginal distributions; $tu = (tu_1, \cdots, tu_d)$; m stands for the Lebesgue measure; and for $s \in (0,1)^d$, $G^*(s) = \{G_i^*(s_i), 1 \leqq i \leqq d\}$ with $G_i^*(s_i) = \inf\{r\colon G_i(r) \geqq s_i\}$.

Condition (B). (b1) H has a density function $h(x) \geqq \delta > 0$ for $x \in [-2L, 2L]^d$.

(b2) There exists a positive function $v(t)$, $t > 0$, with

$$\lim_{t\downarrow 0} v(t) = 0$$

and $\alpha > 0$ such that for all $y > 0$ we have as $t \downarrow 0$,

$$m\{u\colon tu \in A, f(tu) \leqq l + yv(t)\} \to y^\alpha. \tag{3}$$

Although conditions (2) and (3) are not easy to verify, examples (a) and (b) below show that they are not too restrictive. Note that if $G'(x) > 0$ on A then (2) reduces to (3). And, for (3) f need not satisfy sided Lipschitz type conditions or even be continuous. In de Haan [4, p. 469] we can find sufficient conditions for f to satisfy (2) when G is the uniform distribution on A and there exists $x_0 \in A$, $f(x_0) = l$ and $f(x) > l$, $x \neq x_0$. Namely, there exists $v(t) > 0$, $v(t) \to 0$ such that

$$\lim_{t\downarrow 0} \frac{f(x_0 + t_x) - l}{v(t)} \tag{4}$$

exists and is positive for all $x \neq 0$.

Examples. (a) Let $A = [-1, 1]^2$ and $f(x, y) = \max(|x|, |y|)$. Then (4) is verified with $v(t) = t$.

(b) Let $A=[-1, 1]$ and Q be any distribution with density $q(x)>0$ on A. Let

$$f(x)=\begin{cases} x^2, & -1\leqq x\leqq 0, \\ x\left(\sin\dfrac{1}{x}+2\right), & 0<x\leqq 1 \end{cases}$$

and $v(t)=(t/g(0))^2$ then (2) is satisfied with $\alpha=\frac{1}{2}$. Let $\beta>0$ and $f(x)=1$, $-1\leqq x\leqq 0$, $f(x)=x^{\beta}$. $0<x\leqq 1$, then (3) is satisfied with $v(t)=t^{\beta}$ and $\alpha=1/\beta$. Note that in both cases de Haan's condition (4) is not satisfied.

Let $E=[-2L, 2L]^d$, $a_n=v(n^{-1/d})$,

$$\tilde{X}_n=a_n^{-1}(X_n-l) \quad \text{and} \quad \tilde{Y}_n=a_n^{-1}(Y_n-l). \tag{5}$$

THEOREM (A). *Let X_n be defined by Method* (A) *and assume that Condition* (A) *holds. Then*

$$\lim_{n\to\infty} P(\tilde{X}_n\leqq y)=1-b\exp(-y^{\alpha}), \qquad y>0. \tag{6}$$

THEOREM (B). *Let Y_n be defined by Method* (B) *and assume that Condition* (B) *holds. Then for $y>0$*

$$\begin{aligned} 1-\exp(-cy^{\alpha}) &\leqq \liminf_{n\to\infty} P(\tilde{Y}_n\leqq y) \\ &\leqq \limsup_{n\to\infty} P(\tilde{Y}_n\leqq y)\leqq 1-\exp(-Cy^{\alpha}), \end{aligned} \tag{7}$$

where $c=\inf_{x\in E} h(x)$ and $C=\sup_{x\in E} h(x)$.

COROLLARY (B). *If H is such that it is uniform on E then for $a_n=v(n^{-1/d})c^{-1/\alpha}$ we have for $y>0$*

$$\lim_{n\to\infty} P(\tilde{Y}_n\leqq y)=1-\exp(-y^{\alpha}). \tag{8}$$

Remarks. Although we were not able to find the limiting distribution for Method (B) except for those satisfying Corollary (B), it is worth pointing out that with respect to the expected number of steps required to reach an ε-neighborhood of l, the uniform distribution on E is the optimizing one and it satisfies Corollary (B) (see Dorea [2]). If general distributions are considered in Condition (A) by dropping the requirement $G=\pi G_i$, the results of Theorem (A) can still be obtained. In this case one needs to replace (2) by a condition on the distribution of $f(\eta)$:

$$P(\eta\in A, f(\eta)\leqq l+yv(t))\to y^{\alpha}.$$

An alternative approach is to express the required dependence of G by means of constraint functions g_i that determine A.

Proof of Theorem (A). Let η be a r.v. with distribution G and let

$$\begin{aligned} &P_A=P(\eta\in A), \qquad q_A=1-p_A, \\ &\beta_n=P(f(\eta)>l+a_n y,\ \eta\in A), \\ &\gamma_n=P(f(U)>l+a_n y). \end{aligned} \tag{9}$$

Then we can write

$$\begin{aligned} P(\tilde{X}_n>y) &= P(f(U)>l+a_n y,\ \eta_i\notin A,\ i=1,\cdots,n)+\cdots \\ &\quad +P(f(U)>l+a_n y, f(\eta_i)>l+a_n y,\ \eta_i\in A,\ i=1,\cdots,n) \\ &= \gamma_n(q_A+\beta_n)^n. \end{aligned} \tag{10}$$

(a1) Since U is uniformly distributed and $a_n \downarrow 0$ we have

$$\lim_{n\to\infty} \gamma_n = b, \qquad 0 < b \leqq 1. \tag{11}$$

Note that since (2) is satisfied we have $b = 1$ whenever $G'(x) > 0$ on A.

(a2) Let $F_t(y) = \{u\colon u \in A, f(u) \leqq l + yv(t)\}$ so that

$$\beta_n = p_A - P(\eta \in F_{n^{-1/d}}(y)).$$

Let $V = (V_1, \cdots, V_d)$ be uniformly distributed on $[0, 1]^d$. Then, since $G = \pi G_i$ we have η and $G^*(V) = \{G_1^*(V_1), \cdots, G_d^*(V_d)\}$ with the same probability distribution. Hence we can write

$$\begin{aligned} P(\eta \in F_t(y)) &= P(G^*(V) \in F_t(y)) \\ &= P(G^*(V) \in A, f(G^*(V)) \leqq l + yv(t)) = t^d m(A_t(y)). \end{aligned}$$

From (2) it follows that

$$\frac{1}{t^d} P(\eta \in F_t(y)) \to y^\alpha \quad \text{as } t \downarrow 0;$$

hence

$$n \log (1 - P(\eta \in F_{n^{-1/d}}(y)) \to -y^\alpha \quad \text{as } n \to \infty. \tag{12}$$

(a3) Finally from (10), (11) and (12) we have

$$P(\tilde{X}_n \leqq y) = 1 - \gamma_n (1 - P(\eta \in F_{n^{-1/d}}(y)))^n$$

and

$$\lim_{n\to\infty} P(\tilde{X}_n \leqq y) = 1 - b \exp(-y^\alpha).$$

Proof of Theorem (B). First notice that $Y_0 \geqq Y_1 \geqq \cdots \geqq Y_n$ and

$$P(\tilde{Y}_n > y) = P(\tilde{Y}_{n-1} > y) - P(\tilde{Y}_n \leqq y,\ \tilde{Y}_{n-1} > y). \tag{13}$$

(b1) By (3)

$$P(\tilde{Y}_0 > y) = P(f(U) > l + a_n y) = \gamma_n \xrightarrow[n\to\infty]{} 1. \tag{14}$$

(b2) Clearly for all $x \in A$ the conditional density of $W^k + \xi^{k+1}$ given W^k is

$$h_{W^k + \xi^{k+1} | W^k}(z|x) = h(z - x) \tag{15}$$

and

$$\begin{aligned} &\inf_{z\in A, x\in A} h(z - x) \geqq c \geqq \delta > 0, \\ &\sup_{z\in A, x\in A} h(z - x) \leqq C. \end{aligned} \tag{16}$$

Now let F_{j-1} denote the distribution of W^{j-1}. Then by (15)

$$P(\tilde{Y}_j \leqq y,\ \tilde{Y}_{j-1} > y) = \int_{A \setminus Q_n} \int_{Q_n} h(z - x)\, dz\, dF_{j-1}(x)$$

where $Q_n = \{u\colon u \in A, f(u) \leqq l + ya_n\}$.

By (16) we have

$$cm(Q_n) P(\tilde{Y}_{j-1} > y) \leqq P(\tilde{Y}_j \leqq y,\ \tilde{Y}_{j-1} > y) \leqq Cm(Q_n) P(\tilde{Y}_{j-1} > y). \tag{17}$$

From (13), (14) and (17) we have

$$\gamma_n(1-Cm(Q_n))^n \leqq P(\tilde{Y}_n > y) \leqq \gamma_n(1-cm(Q_n))^n.$$

Now from (3) it follows that $nm(Q_n) \to y^\alpha$, so that (7) follows.

3. Estimation of the parameter. Let $Z_1, Z_2, \cdots$ be iid r.v.'s with common distribution F and let $Z_{(1)} \leqq Z_{(2)} \leqq \cdots \leqq Z_{(n)}$ denote the order statistics from $(Z_1, \cdots, Z_n)$. From de Haan [4] it follows that if for some $\alpha > 0$ we have

$$\lim_{t \downarrow 0} \frac{F(l+tx)}{F(l+t)} = x^\alpha \quad \text{for all } x > 0, \tag{18}$$

then for

$$\varphi_Z(n) = \log \frac{Z_{(k(n))} - Z_{(3)}}{Z_{(2)} - Z_{(1)}} \Big/ \log k(n) \tag{19}$$

we have for all $\varepsilon > 0$

$$\lim_{n \to \infty} P(|\varphi_Z(n) - 1/\alpha| > \varepsilon) = 0 \tag{20}$$

provided $k(n) \to \infty$ and $k(n)/n \to 0$.

It is easy to verify that (20) remains true if (18) holds with a positive function $v(t)$ with $\lim_{t \downarrow 0} v(t) = 0$ in place of t. We shall adapt the above results to our dependence setting of Methods (A) and (B). For $j \geqq 1$ define the following stopping times

$$\sigma_j = \inf\{i: i > \sigma_{j-1}, \eta^i \in A\}, \qquad \sigma_0 = 0,$$

$$\tau_j = \inf\{i: i > \tau_{j-1}, W^{\tau_{j-1}} + \xi^i \in A\}, \qquad \tau_0 = 0.$$

For $k \geqq 1$ let $R_k = f(\eta^{\sigma_k})$ and $S_k = f(W^{\tau_k})$ and define $\varphi_R(n)$ and $\varphi_S(n)$ analogously as in (19).

PROPOSITION (A). *Suppose Condition* (A) *is satisfied. Then*

$$\lim_{n \to \infty} P(|\varphi_R(n) - 1/\alpha| \geqq \varepsilon) = 0 \quad \forall \varepsilon > 0. \tag{21}$$

PROPOSITION (B). *Suppose Condition* (B) *is satisfied and H has a density* $h(x)$ *with* $h(x) = c > 0$, $-2L \leqq x \leqq 2L$. *Then*

$$\lim_{n \to \infty} P(|\varphi_S(n) - 1/\alpha| \geqq \varepsilon) = 0 \quad \forall \varepsilon > 0. \tag{22}$$

Proof of Proposition (A). First we will show that $R_1, R_2, \cdots$ are iid with the distribution

$$F_R(x) = \frac{1}{P(A)} \int_{\{u: u \in A, f(u) \leqq x\}} dG(u).$$

This follows from the relation

$$P(\eta^{\sigma_1} \leqq x^1, \cdots, \eta^{\sigma_k} \leqq x^k) = \sum_{1 \leqq i_1 < i_2 < \cdots < i_k} P(\sigma_j = i_j, \eta^{i_j} \leqq x^j, j = 1, \cdots, k)$$

$$= \sum_{1 \leqq i_1 < i_2 < \cdots < i_k} q_A^{i_1 - 1} p_A(x^1) \cdots q_A^{i_k - i_{k-1} - 1} p_A(x^k)$$

where

$$q_A = 1 - P(A), \; p_A(x) = \int_{\{u \in A, u \leqq x\}} dG(u) \quad \text{and} \quad (\eta \leqq x) = \{\eta_i \leqq x_i, i = 1, \cdots, d\}.$$

It remains to show that F_R satisfies (18); now

$$F_R(l+yv(t)) = \frac{1}{P(A)} P(\eta^1 \in A, f(\eta^1) \leqq l+yv(t)).$$

Proceeding as in step (a2) of Theorem (A) we can write

$$F_R(l+yv(t)) = \frac{1}{P(A)} t^d m(A_t(y)).$$

Now from (2)

$$\lim_{t \downarrow 0} \frac{m(A_t(y))}{m(A_t(1))} = y^\alpha,$$

so that (18) follows.

Remark. If we assume F_R absolutely continuous we have

$$P(R_{(1)} < R_{(2)} < \cdots < R_{(n)}) = 1$$

and (21) can be translated in terms of X_n's of Method (A). Let $\rho_1 = n$, $\rho_j = \sup\{k: k < \rho_{j-1}, X_k \neq X_{\rho_{j-1}}\}$ and from (21) $\log((X_{\rho_{k(n)}} - X_{\rho_3})/(X_{\rho_3} - X_{\rho_2}))/\log k(n)$ converges in probability to $1/\alpha$. Analogous results hold for the Y_n's of Method (B).

Proof of Proposition (B). Notice that

$$P(W^{\tau_1} \leqq x) = \sum_{i=1}^{\infty} P(U + \xi^k \leqq x, \tau_1 = k)$$

and

$$P(U + \xi^k \leqq x, \tau_1 = k) = (1 - cm(A))^{k-1} P(U + \xi^k \leqq x, U + \xi^k \in A)$$

since

$$P(U + \xi^k \in A) = \int_A \int_{x \in A \setminus u} \frac{c}{m(A)} du\, dx = cm(A).$$

So we can write

$$P(W^{\tau_1} \leqq x) = \frac{1}{P(A)} P(\xi^1 : \xi^1 \in A, \xi^1 \leqq x).$$

Now proceeding as in Proposition (A) we can show that $W^{\tau_1}, W^{\tau_2}, \cdots$ are iid r.v.'s. Moreover

$$F_S(x) = \frac{c}{P(A)} m\{u: u \in A, f(u) \leqq x\}$$

and (18) follows from (3).

Acknowledgment. The author wishes to thank the referee for helpful comments and suggestions.

REFERENCES

[1] N. BABA, *Convergence of a random optimization method for constrained optimization problems*, J. Optim. Theory Appl., 33 (1981), pp. 451-461.

[2] C. C. Y. DOREA, *Expected number of steps of a random optimization method*, J. Optim. Theory Appl., 39 (1983), pp. 165-171.

[3] P. B. GNEDENKO, *Sur la distribution limite du terme maximum d'une série aléatoire*, Ann. Math., 44 (1943), pp. 423-453.

[4] L. DE HAAN, *Estimation of the minimum of a function using order statistics*, J. Amer. Stat. Assoc., Theory and Methods Section, 76 (1981), pp. 467-469.

[5] P. HALL, *Representations and limit theorems for extreme value distributions*, J. Appl. Prob., 15 (1978), pp. 639-644.

[6] F. J. SOLIS AND R. J. B. WETS, *Minimization by random search techniques*, Math. Operations Res., 6 (1981), pp. 19-30.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 1, January, 1986

004

ON THE INTERPLAY OF SINGULAR PERTURBATIONS AND WIDE-BAND STOCHASTIC FLUCTUATIONS*

MOHAMED EL-ANSARY† AND HASSAN KHALIL‡

Abstract. A class of nonlinear singularly perturbed systems driven by wide-band noise is considered. The asymptotic behavior of the slow variables is studied when the fast variables are sufficiently fast (represented by $\mu \to 0$) and the wide-band noise is sufficiently wide (represented by $\varepsilon \to 0$). A reduced-order Markov model which represents the behavior of the slow variables is derived. It is shown that the slow variables converge weakly to the solution of this reduced-order model as ε and μ tend to zero. The coefficients of this reduced-order model depend, in general, on the speeds of ε and μ as they approach zero. The implication of such dependence on the engineering practice of neglecting parasitic elements is discussed. Special cases where the model is independent of the speeds of ε and μ are explored.

Key words. asymptotic methods, model order reduction, singular perturbation, wide-band stochastic fluctuations

1. Introduction. The purpose of this paper is to highlight the interplay of two asymptotic phenomena which arise in the analysis and design of control systems. The first phenomenon arises in singular perturbation analysis of deterministic systems containing parasitic elements while the second phenomenon arises in asymptotic stochastic analysis of systems driven by wide-band noise.

Consider the singularly perturbed system

$$\dot{x}(t) = f(x(t), y(t), u(t)), \tag{1.1}$$

$$\mu \dot{y}(t) = g(x(t), y(t), u(t)), \tag{1.2}$$

where μ is a small positive parameter representing parasitic elements. If the input $u(t)$ is smooth and deterministic, a reduced-order model of this system can be obtained by neglecting the parasitic elements, i.e. by setting $\mu = 0$. Assuming that the algebraic equation

$$0 = g(\bar{x}(t), \bar{y}(t), u(t)) \tag{1.3}$$

has a unique root

$$\bar{y}(t) = h(\bar{x}(t), u(t)), \tag{1.4}$$

the reduced-order model is given by

$$\dot{\bar{x}}(t) = f(\bar{x}(t), u(t)). \tag{1.5}$$

The literature on singular perturbation theory (see [1] for a survey) is full of analyses validating this order reduction procedure. In particular if both $x(t)$ and $\bar{x}(t)$ start from the same initial conditions, then under certain stability conditions imposed on the boundary layer system, $x(t) \to \bar{x}(t)$ as $\mu \to 0$, on compact time intervals.

Consider next a system described by the differential equation

$$\dot{x}(t) = f(x(t)) + G(x(t))v^{\varepsilon}(t) \tag{1.6}$$

* Received by the editors July 5, 1983, and in revised form July 10, 1984. A preliminary version of this paper was presented at the 21st IEEE Conference on Decision and Control, Orlando, Florida, 1982. The work of this paper was supported by the U.S. Department of Energy, Electric Energy Systems Division, under contract DE-AC01-80RA50425 with Michigan State University.

† Department of Mathematics, California State College-Bakersfield, Bakersfield, California 93311.

‡ Department of Electrical Engineering and Systems Science, Michigan State University, East Lansing, Michigan 48824.

where $v^\varepsilon(t)$ is a wide-band stationary process in the sense that its spectrum is flat up to a frequency of order $1/\varepsilon$ where ε is a small positive parameter. The process $v^\varepsilon(t)$ is not white noise but it tends to white noise as $\varepsilon \to 0$. Asymptotic stochastic analysis (cf. [2]-[4]) shows that $x(t)$ converges weakly to a diffusion process $\bar{x}(t)$ as $\varepsilon \to 0$, where $\bar{x}(t)$ satisfies the Ito equation

$$d\bar{x}(t) = (f(\bar{x}(t)) + g(\bar{x}(t)))\, dt + G(\bar{x}(t))\, dw. \tag{1.7}$$

The vector $g(\bar{x})$, which is usually referred to as the correction term, is formed of the elements of G and their partial derivatives with respect to x. Since most of the stochastic stability and control results have been developed for systems driven by white noise [5], the importance of the above limit is in extending the range of those results to cases when the noise is not white but only approximately so (see [3] for details).

The study of systems where both parasitic parameters and wide-band noise are present has been initiated in [6], [7]. The first paper treated linear systems and the second one nonlinear systems. The results of [6] and [7] did not show the interaction between the two asymptotic phenomena in which we are interested in this paper. Such interaction could not appear in [6] anyway since, as it will be shown, it is tied in with nonlinearities. The approach adopted in [7] did not reveal such interaction because implicit in that approach there was a sequential ordering of the two asymptotic phenomena in the sense that order reduction as in singular perturbation methods was performed first, followed by asymptotic stochastic analysis to compute the diffusion limit. The interaction between the two asymptotic phenomena has been brought to attention after a paper by Razvig [8]. In that paper Razvig considered the second order equation

$$\mu\ddot{x}(t) + \dot{x}(t) = a(x(t)) + b(x(t))v^\varepsilon(t) \tag{1.8}$$

where $v^\varepsilon(t)$ is exponentially correlated noise with correlation time ε. He studied the asymptotic behavior of $x(t)$ as ε and μ tend to zero and suggested that for sufficiently small ε and μ, $x(t)$ can be approximated by a diffusion process $\bar{x}(t)$ defined by the Ito equation

$$d\bar{x}(t) = \left[a(\bar{x}(t)) + \frac{0.5}{1+\varepsilon/\mu} \frac{\partial b}{\partial x}(\bar{x}(t))b(\bar{x}(t))S(0) \right] dt + b(x(t))\sqrt{S(0)}\, dw(t) \tag{1.9}$$

where $S(\omega/\varepsilon)$ is the spectrum of v^ε. In deriving this reduced-order model Razvig employed a formal intuitive reasoning. He assumed that over a time interval Δt which is very small with respect to the relaxation time of $x(t)$ while very large with respect to μ and ε, the process $x(t)$ will behave like a Markov process. With that assumption he went on to compute the first and second moments of $x(t+\Delta t) - x(t)$ given $x(t) = x$ which resulted in the drift and diffusion coefficients of (1.9). Razvig did not prove that $x(t)$ converges to $\bar{x}(t)$ as $\varepsilon, \mu \to 0$ in any stochastic sense. The remarkable feature of the reduced order model (1.9) is its dependence on the ratio ε/μ hinting to the interaction between the two asymptotic phenomena. Our work has been motivated by Razvig's example. Our objective has been to generalize the reduced-order model of Razvig to a wider class of systems and to provide a rigorous proof of convergence of $x(t)$ to the diffusion process defined by the reduced-order model. Section 2 gives a reduced-order model and convergence proof for a class of nonlinear singularly perturbed systems driven by wide-band noise. The class of singularly perturbed systems considered is a special case of (1.1), (1.2) in which y appears linearly. The linearity in y is assumed to avoid technical complications; yet, it is a realistic assumption and parasitics in many physical systems appear in this form. The convergence proof adapts

a martingale method developed by Kushner [9] for proving weak convergence of a sequence of non-Markovian processes to a diffusion process. In § 3 we discuss the implication of the results of § 2 on the robustness of the well established engineering practice of reducing the order of physical systems by neglecting parasitic elements.

2. Reduced-order model and convergence result. Consider the singularly perturbed system

$$\dot{x}^{\varepsilon,\mu}(t) = a_1(x^{\varepsilon,\mu}(t)) + A_{12}(x^{\varepsilon,\mu}(t))y^{\varepsilon,\mu}(t) + B_1(x^{\varepsilon,\mu}(t))v^{\varepsilon}(t), \qquad x^{\varepsilon,\mu}(0) = x_0, \tag{2.1}$$

$$\mu\dot{y}^{\varepsilon,\mu}(t) = a_{21}(x^{\varepsilon,\mu}(t)) + A_2 y^{\varepsilon,\mu}(t) + B_2(x^{\varepsilon,\mu}(t))v^{\varepsilon}(t), \qquad y^{\varepsilon,\mu}(0) = y_0 \tag{2.2}$$

where $x \in R^n$, $y \in R^m$ and x_0, y_0 are bounded random vectors. The stochastic process $v^{\varepsilon} \in R^r$ is defined as

$$v^{\varepsilon}(t) = \frac{1}{\sqrt{\varepsilon}} v(t/\varepsilon) \tag{2.3}$$

where $v(t)$ satisfies:

(A1) $v(t)$ is a stationary, zero mean, right continuous, uniformly bounded process on $[0, \infty)$. The σ-algebras induced by $v(t)$ are assumed to have a mixing property with an exponential mixing rate [10],

$$\sup_{A_i, t} |P(A_2/A_1) - P(A_2)| \leqq K e^{-\alpha\tau} \tag{2.4}$$

for some $\alpha > 0$, where $A_1 \in \sigma\{v(s), s \leqq t\}$ and $A_2 \in \sigma\{v(s), s \geqq t+\tau\}$. The exponential mixing rate assumption is taken for convenience but can be replaced by a more general mixing rate as in [2]-[4]. The process $v^{\varepsilon}(t)$ is said to be wide-band noise since its power spectral density matrix $S^{\varepsilon}(\omega) = S(\omega/\varepsilon)$ will have a frequency band of ω_0/ε when $S(\omega)$, the spectral matrix of v, has a frequency band ω_0. Indeed, the process $v^{\varepsilon}(t)$ converges to Gaussian white noise by the central limit theorem [2].

We make the following assumptions:

(A2) The coefficients a_1, a_{21}, A_{12}, B_1 and B_2 are continuous in x and have continuous partial derivatives up to the second order which are bounded uniformly in x;

(A3) The constant matrix A_2 is Hurwitz, i.e. $\operatorname{Re} \lambda(A_2) < 0$;

(A4) The positive parameters ε and μ satisfy $\varepsilon > \mu\gamma_0$ where $\gamma_0 > 0$ is arbitrary but fixed.

Under the smoothness conditions spelled out in (A2), the usual existence and uniqueness theory for ordinary differential equations gives us a solution for (2.1) and (2.2) on $[0, T]$ for each sample path of $v(\cdot)$. Condition (A3) is needed to guarantee asymptotic stability of the boundary-layer phenomena associated with y. Condition (A4) excludes the case $\varepsilon/\mu \to 0$ as $\mu \to 0$. This technicality is needed in the convergence proof as it will be pointed out later.

Our objective is to study the asymptotic behavior of $x^{\varepsilon,\mu}(\cdot)$ as $\varepsilon \to 0$ and $\mu \to 0$. The main result of this paper shows that $x^{\varepsilon,\mu}(\cdot)$ converges weakly to a diffusion process $\bar{x}(\cdot)$ with initial condition $\bar{x}(0) = x_0$. The infinitesimal generator associated with $\bar{x}(\cdot)$, whose form will follow from the proof of the result, is given by

$$L^{\gamma} f(x) = \sum_{i=1}^{n} b_i(x) \frac{\partial f}{\partial x_i}(x) + \frac{1}{2} \sum_{i,j=1}^{n} a_{ij}(x) \frac{\partial^2 f}{\partial x_i \, \partial x_j}(x), \tag{2.5}$$

where

$$b(x) = a_0(x) + h_1(x) - A_{12}(x)A_2^{-1}h_2(x) + h_3(x), \tag{2.6}$$

$$A(x) = B_0(x)S(0)B_0^T(x) \triangleq [a_{ij}(x)], \tag{2.7}$$

$$a_0(x) = a_1(x) - A_{12}(x)A_2^{-1}a_{21}(x), \tag{2.8}$$

$$B_0(x) = B_1(x) - A_{12}(x)A_2^{-1}B_2(x), \tag{2.9}$$

$S(\omega)$ is the spectral matrix of v,

$$h_{1i} = \operatorname{tr}[D_i'B_0W' + D_i'A_{12}A_2^{-1}\Sigma]^1, \tag{2.10}$$

$$h_{2i} = \operatorname{tr}[E_i'B_0W' + E_i'A_{12}A_2^{-1}\Sigma], \tag{2.11}$$

$$h_{3i} = \operatorname{tr}[-F_i'B_0W'B_2'(A_2')^{-1} - F_i'B_0\Sigma'(A_2')^{-1} + F_i'A_{12}A_2^{-1}P], \tag{2.12}$$

$$D_i = \left[\nabla_x\psi_{i1} \mid \nabla_x\psi_{i2} \mid \text{-----} \mid \nabla_x\psi_{ir}\right]_{n\times r}, \qquad B_1 = [\psi_{ij}]_{n\times r}, \tag{2.13}$$

$$E_i = \left[\nabla_x\eta_{i1} \mid \nabla_x\eta_{i2} \mid \text{-----} \mid \nabla_x\eta_{ir}\right]_{n\times r}, \qquad B_2 = [\eta_{ij}]_{m\times r}, \tag{2.14}$$

$$F_i = \left[\nabla_x\xi_{i1} \mid \nabla_x\xi_{i2} \mid \text{-----} \mid \nabla_x\xi_{im}\right]_{n\times m}, \qquad A_{12} = [\xi ij]_{n\times m}, \tag{2.15}$$

$$W = \int_0^\infty R(\tau)\,d\tau, \quad R \text{ is the correlation matrix of } v,$$

$$\Sigma = \int_0^\infty e^{A_2\gamma\tau}B_2R'(\tau)\,d\tau \quad \text{for some } \gamma \in [\gamma_0, \infty),\ \gamma_0 > 0, \tag{2.16}$$

$$P = \int_0^\infty e^{A_2\lambda}(B_2\Sigma' + \Sigma B_2')\,e^{A_2'\lambda}\,d\lambda. \tag{2.17}$$

We require that

(A5) $b(x)$ and $B_0(x)$ satisfy the growth and Lipschitz conditions

$$|b(x)| + |B_0(x)| \leqq K(1+|x|) \qquad \forall x \in R^n,$$

$$|b(x) - b(z)| + |B_0(x) - B_0(z)| \leqq K|x - z| \qquad \forall x, z \in R^n.$$

Condition (A5) implies that the martingale problem corresponding to L^γ is well-posed [11]. Condition (A5) follows from the smoothness condition (A2) if A_{12} is independent of x or if a_{21} is bounded and B_2 is independent of x. It is, however, a restrictive requirement that eliminates some interesting problems. For example, it does not allow both A_{12} and B_2 to be linear in x simultaneously.

Our main result is the following theorem.

THEOREM 1. *Under the assumptions* (A1) *to* (A5), $x^{\varepsilon,\mu}(\cdot)$ *converges weakly to* $\bar{x}(\cdot)$ *as* $\varepsilon \to 0$, $\mu \to 0$ *and* $\varepsilon/\mu \to \gamma$.

Proof. We utilize a technique for proving weak convergence of a sequence of non-Markovian processes to a diffusion process which was introduced by Kurtz [12] and further developed by Kushner [4], [9], [13]. The version used here is due to Kushner [9]. Our proof, in fact, follows Kushner's method step-by-step as it was applied in [13]. There are, however, two additional levels of complexity in our proof. The first one is due to the singularly perturbed equation (2.2). The second one is in computing the explicit form of the operator L^γ, which required lengthy and detailed manipulations. Before we state Kushner's method we need to introduce some definitions and terminology which are recalled from [9].

[1] (') denotes transposition.

Truncated processes. For every positive integer N, let $S_N = \{x \in R^n, \|x\| \leqq N\}$ and define the truncated process $x_N^{\varepsilon,\mu}(t)$ to be the solution of (2.1), (2.2) with the right-hand side of (2.1) multiplied by $q_N(x)$, i.e.,

$$\dot{x}_N^{\varepsilon,\mu} = q_N(x_N^{\varepsilon,\mu})[a_1(x_N^{\varepsilon,\mu}) + A_{12}(x_N^{\varepsilon,\mu})y_N^{\varepsilon,\mu} + B_1(x_N^{\varepsilon,\mu})v^\varepsilon], \quad x_N^{\varepsilon,\mu}(0) = x_0, \tag{2.18}$$

$$\mu\dot{y}_N^{\varepsilon,\mu} = a_{21}(x_N^{\varepsilon,\mu}) + A_2 y_N^{\varepsilon,\mu} + B_2(x_N^{\varepsilon,\mu})v^\varepsilon, \quad y_N^{\varepsilon,\mu}(0) = y_0, \tag{2.19}$$

where $q_N(x) = 1$ for $x \in S_N$, $q_N(x) = 0$ for $x \in R^n - S_{N+1}$ and $q_N(x) \in [0, 1]$ and has third derivatives that are bounded uniformly in x and N. For each N, $\{x_N^{\varepsilon,\mu}(\cdot)\}$ is bounded uniformly in μ and ε. As it will be seen, the actual technical proof involves only the truncated processes $\{x_N^{\varepsilon,\mu}(\cdot)\}$. See [9], [13] for similar treatment.

Terminology. Let $(\Omega, P, \mathscr{F})$ be the probability space in which $v(\cdot)$ is defined and let $\mathscr{F}_{t,N}^{\varepsilon,\mu}$ be the σ-algebra induced by $\{x_N^{\varepsilon,\mu}(s), y_N^{\varepsilon,\mu}(s), v^\varepsilon(s), 0 \leqq s \leqq t\}$ and $E_{t,N}^{\varepsilon,\mu}$ the corresponding conditional expectation. Let $\mathscr{L}^0$ be the class of measurable (ω, t) real valued functions such that if $f(\cdot) \in \mathscr{L}^0$ then $E|f(t+s) - f(t)| \to 0$ as $s \to 0^+$, $\sup_t E|f(t)| < \infty$ and $f(t)$ is adapted to $\mathscr{F}_{t,N}^{\varepsilon,\mu}$. We say $p - \lim_{s\to 0} f^s = 0 \Leftrightarrow \sup_{s,t} E|f^s(t)| < \infty$ and $E|f^s(t)| \to 0$ as $s \to 0^+$. Define an operator $A_N^{\varepsilon,\mu}$ and its domain $D(A_N^{\varepsilon,\mu})$ as follows: $f \in D(A_N^{\varepsilon,\mu})$ and $A_N^{\varepsilon,\mu} f = g \Leftrightarrow f, g \in \mathscr{L}^0$ and

$$p - \lim_{r\to 0} \left| \frac{E_{t,N}^{\varepsilon,\mu} f(t+r) - f(t)}{r} - g(t) \right| = 0.$$

Let L_N^γ be a diffusion operator of the form (2.7) such that the coefficients of L_N^γ and L^γ are equal for $x \in S_N$. Let $\hat{\mathscr{C}}_0$ be the space of continuous functions $f: R^n \to R$ which have compact support and $\hat{\mathscr{C}}_0^3$ be the space of functions which belongs to $\hat{\mathscr{C}}_0$ together with its partial derivatives up to the third order.

The following lemma is Theorems (1) and (2) of [9] adapted to our case.

LEMMA 1. *Assume that the martingale problem associated with L^γ is well-posed. For each fixed N, let $\{x_N^{\varepsilon,\mu}(\cdot)\}$ be the solution of* (2.18) *and* (2.19). *Suppose that for each $f \in \hat{\mathscr{C}}_0^3$, there is a sequence $f_N^{\varepsilon,\mu}(\cdot) \in D(A_N^{\varepsilon,\mu})$ and a random variable $M_{N,T}^{\varepsilon,\mu}(f)$, for each $T > 0$, such that*

$$p - \lim_{\substack{\varepsilon,\mu\to 0 \\ \varepsilon/\mu\to\gamma}} [f_N^{\varepsilon,\mu}(t) - f(x_N^{\varepsilon,\mu}(t))] = 0, \tag{2.20}$$

$$p - \lim_{\substack{\varepsilon,\mu\to 0 \\ \varepsilon/\mu\to\gamma}} [A_N^{\varepsilon,\mu} f_N^{\varepsilon,\mu}(t) - L_N^\gamma f(x_N^{\varepsilon,\mu}(t))] = 0, \tag{2.21}$$

$$p\{\sup_{t\leqq T} |f_N^{\varepsilon,\mu}(t) - f(x_N^{\varepsilon,\mu}(t))| \geqq \eta\} \to 0 \quad \text{as } \varepsilon, \mu \to 0, \varepsilon/\mu \to \gamma, \tag{2.22}$$

$$\sup_{t\leqq T} |A_N^{\varepsilon,\mu} f_N^{\varepsilon,\mu}(t)| \leqq M_{N,T}^{\varepsilon,\mu}(f), \tag{2.23}$$

$$\sup_{\varepsilon,\mu} p\{M_{N,T}^{\varepsilon,\mu}(f) \geqq K\} \to 0 \quad \text{as } K \to \infty. \tag{2.24}$$

Then $\{x^{\varepsilon,\mu}(\cdot)\}$ converges weakly to $\bar{x}(\cdot)$ as $\varepsilon \to 0$, $\mu \to 0$ and $\varepsilon/\mu \to \gamma$.

For notational convenience we write $x(t)$, $y(t)$, $A^{\varepsilon,\mu}$, L^γ, $f_i(t)$ and E_t instead of $x_N^{\varepsilon,\mu}(t)$, $y_N^{\varepsilon,\mu}(t)$, $A_N^{\varepsilon,\mu}$, L_N^γ, $f_{i,N}^{\varepsilon,\mu}(t)$ and $E_{t,N}^{\varepsilon,\mu}$ respectively but we are always working with the truncated process $\{x_N^{\varepsilon,\mu}(\cdot)\}$. Moreover, we omit the q_N terms for further simplification. Now we proceed with the proof of the theorem. Let $f \in \mathscr{C}_0^3$ be given, then the test function $f^{\varepsilon,\mu}(t)$ is constructed in three steps as in [13]. First, we have

$$A^{\varepsilon,\mu} f(x(t)) = \frac{\partial f}{\partial x}(x(t))[a_1(x(t)) + A_{12}(x(t))y(t) + B_1(x(t))v^\varepsilon(t)]. \tag{2.25}$$

The last two terms on the right-hand side of (2.25) are not uniformly bounded in ε and μ and cannot be part of the operator L^γ, so they are averaged out by defining $f_1(x, t)$ as:

$$f_1(x,t)=\int_0^\infty E_t\frac{\partial f}{\partial x}(x)[A_{12}(x)[\hat{y}(t+s,x)+A_2^{-1}a_{21}(x)]+B_1(x)v^\varepsilon(t+s)]\,ds \tag{2.26}$$

where

$$\begin{aligned}\hat{y}(t+s,x)&=e^{A_2 s/\mu}y(t)+(e^{A_2 s/\mu}-I)A_2^{-1}a_{21}(x)\\&\quad+\frac{1}{\mu}\int_t^{t+s}e^{A_2(t+s-\tau)/\mu}B_2(x)v^\varepsilon(\tau)\,d\tau.\end{aligned} \tag{2.27}$$

Here, $\hat{y}(t+s, x)$ is the solution of the singularly perturbed equation (2.19) starting at $s=0$ from initial condition $y(t)$ with x, on the right-hand side of (2.19), being frozen at $x=x(t)$. Subtracting the term $-A_2^{-1}a_{21}(x)$ in (2.26), in a sense, centers $\hat{y}$ at its steady-state mean. Using (2.27), it is straightforward to show that $f_1(x, t)$ is given by

$$\begin{aligned}f_1(x,t)&=-\mu\frac{\partial f}{\partial x}(x)A_{12}(x)A_2^{-1}[y(t)+A_2^{-1}a_{21}(x)]\\&\quad+\frac{\partial f}{\partial x}B_0(x)\int_0^\infty E_t v^\varepsilon(t+s)\,ds.\end{aligned} \tag{2.28}$$

But direct solution of (2.19) starting at $t=0$ shows that $y(t)$ is given by

$$y(t)=e^{A_2 t/\mu}y_0+\frac{1}{\mu}\int_0^t e^{A_2(t-\tau)/\mu}\left[a_{21}(x(\tau))+B_2(x(\tau))\frac{1}{\sqrt{\varepsilon}}v(\tau/\varepsilon)\right]d\tau. \tag{2.29}$$

Using the boundedness of v and y_0, the boundedness of a_{21} and B_2, which follows from the boundedness of the truncated process, and the exponentially decaying nature of the transition matrix $\exp[A_2 t/\mu]$, it can be shown that $\sqrt{\varepsilon}\,y(t)$ is uniformly bounded in ε and μ. Hence, using the mixing property (2.4), the compact support of $(\partial f/\partial x)(x)$ and the boundedness of the truncated process $x(t)$, we get

$$|f_1(x,t)|\leqq C_1\sqrt{\varepsilon}+C_2\mu/\sqrt{\varepsilon}. \tag{2.30}$$

The restriction $\varepsilon/\mu\geqq\gamma_0>0$ guarantees that the right-hand side of (2.30) is bounded by $C_1\sqrt{\varepsilon}+C_3\sqrt{\mu}$; that is why this restriction has been imposed. Thus we have shown that

$$|f_1(t)|=|f_1(x(t),t)|\leqq K_1(\sqrt{\varepsilon}+\sqrt{\mu}). \tag{2.31}$$

$K_1>0$ is independent of ε, μ and ω.

Operating on $f_1(t)$ by $A^{\varepsilon,\mu}$, we get

$$\begin{aligned}A^{\varepsilon,\mu}f_1(t)&=-\frac{\partial f}{\partial x}(x)[A_{12}(x)y(t)+A_{12}(x)A_2^{-1}a_{21}(x)+B_1(x)v^\varepsilon(t)]\\&\quad+\frac{\partial f_1}{\partial x}(x,t)[a_1(x)+A_{12}(x)y(t)+B_1(x)v^\varepsilon(t)].\end{aligned} \tag{2.32}$$

Adding (2.25) to (2.32) yields

$$A^{\varepsilon,\mu}(f(x)+f_1(t))=\frac{\partial f}{\partial x}(x)a_0(x)+\frac{\partial f_1}{\partial x}(x,t)[a_1(x)+A_{12}(x)y(t)+B_1(x)v^\varepsilon(t)]. \tag{2.33}$$

The last two terms of (2.33) are, again, not uniformly bounded in ε and μ and cannot be part of L^γ, so we average them out by defining f_2 as

$$f_2(x,t)=\int_0^\infty \left[E_t \frac{\partial f_1}{\partial x}(x,t+s)(A_{12}(x)\hat{y}(t+s,x)+A_{12}(x)A_2^{-1}a_{21}(x)+B_1(x)v^\varepsilon(t+s)) \right.$$

(2.34)

$$\left. +\frac{\partial f}{\partial x}(x)a_0(x)-L^{(\varepsilon/\mu)}f(x) \right] ds.$$

The form of $L^{(\varepsilon/\mu)}$, as defined by (2.5)-(2.17) with ε/μ replacing γ, results as a by product of showing that $|f_2(x(t),t)|$ is $O(\mu+\varepsilon)$, i.e., by identifying the parts of the first three terms on the right-hand side of (2.34) which are not $O(\varepsilon)$ or $O(\mu)$. This involves lengthy calculations which are given in detail in [17]. Using the compact support of f_x and f_{xx}, the mixing property (2.4) and the boundedness of the truncated process, it can be shown that

$$|f_2(t)|=|f_2(x(t),t)|\leqq K_2\varepsilon+K_3\mu \tag{2.35}$$

where K_2 and K_3 are positive constants independent of ε, μ and ω. Operating on $f_2(t)$ by $A^{\varepsilon,\mu}$ yields

$$A^{\varepsilon,\mu}f_2(t)=L^{(\varepsilon/\mu)}f(x)-\frac{\partial f_1}{\partial x}(x,t)(A_{12}(x)y(t)+A_{12}(x)A_2^{-1}a_{21}(x)+B_1(x)v^\varepsilon(t))$$

(2.36)

$$-\frac{\partial f}{\partial x}(x)a_0(x)+\frac{\partial f_2}{\partial x}(x,t)(a_1(x)+A_{12}(x)y(t)+B_1(x)v^\varepsilon(t)).$$

Adding (2.33) to (2.36) we get

$$A^{\varepsilon,\mu}(f(x)+f_1(t)+f_2(t))$$

(2.37)

$$=L^{\varepsilon/\mu}f(x)+\frac{\partial f_1}{\partial x}(x,t)a_0(x)+\frac{\partial f_2}{\partial x}(x,t)[a_1(x)+A_{12}(x)y(t)+B_1(x)v^\varepsilon(t)].$$

We define

$$f^{\varepsilon,\mu}(t)=f(x(t))+f_1(x(t),t)+f_2(x(t),t) \quad \text{for } 0\leqq t\leqq T. \tag{2.38}$$

Then condition (2.20) of Lemma 1 follows directly from (2.31), (2.35) and (2.38).

From the mixing property, the compact supports of the partial derivatives of f up to the third order and the boundedness of the truncated process we can show that

$$\left|\frac{\partial f_1}{\partial x}(x,t)a_0(x)\right|+\left|\frac{\partial f_2}{\partial x}(x,t)(A_{12}(x)y(t)+B_1v^\varepsilon(t))\right|\leqq K_4\sqrt{\varepsilon}+K_5\sqrt{\mu} \tag{2.39}$$

and

$$\left|\frac{\partial f_2}{\partial x}(x,t)a_1(x)\right|\leqq K_6\varepsilon+K_7\mu, \tag{2.40}$$

where the positive constants K_i are independent of ε, μ and ω.

By the smooth dependence of L^γ on γ (see (2.16)), it follows that there exists a constant $c>0$ such that

$$|L^{(\varepsilon/\mu)}f(x)-L^\gamma f(x)|\leqq c\left|\frac{\varepsilon}{\mu}-\gamma\right|. \tag{2.41}$$

From (2.37), (2.38) we have:

$$
\begin{aligned}
&|A^{\varepsilon,\mu}f^{\varepsilon,\mu}(t) - L^{\gamma}f(x(t))| \\
&\leqq \left| \frac{\partial f_1}{\partial x}(x,t)a_0(x) + \frac{\partial f_2}{\partial x}(x,t)(a_1(x) + A_{12}(x)y(t) + B_1(x)v^{\varepsilon}(t)) \right. \\
&\qquad \left. + L^{(\varepsilon/\mu)}f(x(t)) - L^{\gamma}f(x(t)) \right|.
\end{aligned} \tag{2.42}
$$

Then (2.21) of Lemma (1) follows immediately from (2.39)-(2.42).

Now we need to verify (2.22)-(2.24) of Lemma 1. The limit (2.22) follows directly from (2.31) and (2.35). Also (2.23) and (2.24) follow easily from (2.39)-(2.42) and the compact support of f. Thus the proof of the theorem is completed by applying Lemma 1. Q.E.D.

From the proof of Theorem 1 it is apparent that if L^{γ} is independent of the parameter γ then the requirement $\varepsilon/\mu \to \gamma$ in Theorem 1 can be dropped.

COROLLARY 1. *Suppose that assumptions* (A1) *to* (A5) *hold and that* $L^{\gamma} = L$ (*independent of* γ). *Then* $x^{\varepsilon,\mu}(\cdot)$ *converges weakly to* $\bar{x}(\cdot)$ *as* $\varepsilon \to 0$ *and* $\mu \to 0$.

In Theorem 1 we employed the noise condition (A1) which requires the noise process $v(\cdot)$ to be a stationary, uniformly bounded, ϕ-mixing process. The use of ϕ-mixing processes is typical in asymptotic stochastic analysis (e.g. [2]-[4]), although in those references ϕ does not have to be exponentially decaying. It is obvious, however, from Kushner's work [4] that the noise process $v(\cdot)$ can be a stationary Gaussian process with rational spectra. Since this type of unbounded noise process very often is more important in applications, we extend our result to that case.

Suppose that the initial states x_0 and y_0 are Gaussian random vectors with bounded second moments. Assumption (A1) is replaced by

(A1)′ The process $v(t)$ is defined by

$$
\begin{aligned}
d\xi &= Q\xi\, dt + U\, dw, \\
v(t) &= \Gamma\xi(t),
\end{aligned} \tag{2.43}
$$

where $w(\cdot)$ is a vector Brownian motion, Q, U and Γ are constant matrices and Q is Hurwitz.

THEOREM 2. *Under the assumptions* (A1)′ *and* (A2)-(A5), $x^{\varepsilon,\mu}(\cdot)$ *converges weakly to a diffusion process* $\bar{x}(\cdot)$ *with initial condition* $\bar{x}(0) = x_0$ *and infinitesimal generator defined by* (2.5).

Remark. A corollary similar to Corollary 1 holds in this case as well.

Proof. The proof is almost the same as that of Theorem 1. The only difference is that instead of using the mixing property we use inequalities that follow directly from (2.43), e.g., $|E_t v(t+s)| \leqq K e^{-\alpha s}|\xi(t)$. Proceeding as in the proof of Theorem 1 we arrive at four inequalities similar to (2.31), (2.35), (2.39) and (2.40) with the left-hand side replaced by its expectation. For example, (2.31) is replaced by

$$E|f_1(t)| \leqq K_1(\sqrt{\varepsilon} + \sqrt{\mu}).$$

Conditions (2.20) and (2.21) of Lemma 1 follow immediately from these four inequalities. To verify conditions (2.22)-(2.24) of Lemma 1, we use the fact that there is a finite w.p.1 w-function C such that $|\xi(t)| \leqq C$ for all $t \in [0, T]$ w.p.1. This fact leads to four inequalities similar, again, to (2.31), (2.35), (2.39) and (2.40) which hold w.p.1 and the right-hand side constants (e.g., K_1 in (2.31)) are replaced by finite w.p.1 w-functions. These inequalities lead easily to (2.22)-(2.24). Q.E.D.

3. Discussion and conclusions. Our main motivation for studying the asymptotic behavior of singularly perturbed systems driven by wide-band noise has been to highlight the interplay of the two asymptotic phenomena involved in the problem. The explicit form of the limiting diffusion model derived in § 2 clearly shows such interaction through the dependence of the matrix Σ, defined by (2.14), on $\gamma = \lim_{\varepsilon,\mu\to 0} \varepsilon/\mu$. Such interplay becomes significant when we study its impact on the well-established engineering practice of neglecting parasitic elements when writing down differential equations representing electrical networks, mechanical systems, etc. According to that practice, in modeling the singularly perturbed system (2.1), (2.2), the parasitic elements represented by μ are neglected. This is equivalent to setting $\mu = 0$ in (2.2) replacing (2.2) by the algebraic equation

$$0 = a_{21}(x) + A_2 y + B_2(x)v^{\varepsilon}. \tag{3.1}$$

Since A_2 is nonsingular, (3.1) can be solved to get

$$y = -A_2^{-1}[a_{21}(x) + B_2(x)v^{\varepsilon}]. \tag{3.2}$$

Substituting (3.2) in (2.1) results in the reduced-order model

$$\dot{x} = a_0(x) + B_0(x)v^{\varepsilon} \tag{3.3}$$

where a_0 and B_0 are defined in (2.8) and (2.9). Although in deriving the reduced-order model (3.3) we started with the higher-dimensional model (2.1) and (2.2) and arrived at (3.3) through the formal procedure of setting $\mu = 0$, it is typical in practical modeling situations that such order reduction is done even before writing down any equations representing the system. Parasitic elements are usually neglected by omitting them from the physical description of the system (e.g., omitting parasitic inductances or capacitances from electrical networks descriptions). As a result of that it is not uncommon that the only available mathematical model to describe the system is the reduced-order model (3.3). Using (3.3) to characterize the behavior of x may lead to wrong conclusions. In particular, since our interest here is in the case when ε is sufficiently small, we can study the asymptotic behavior of x as $\varepsilon \to 0$ using well-established techniques (e.g. [2]-[4]). As $\varepsilon \to 0$ the solution of (3.3) converges weakly to a diffusion process with infinitesimal generator $\tilde{L}$ given by

$$\tilde{L}f(x) = \sum_{i=1}^{n} \tilde{b}_i(x)\frac{\partial f}{\partial x}(x) + \frac{1}{2}\sum_{i,j=1}^{n} \tilde{a}_{ij}(x)\frac{\partial^2 f(x)}{\partial x_i\, \partial x_j} \tag{3.4}$$

where

$$\tilde{b} = a_0 + \tilde{h}_1 - A_{12}A_2^{-1}\tilde{h}_2 + \tilde{h}_3, \tag{3.5}$$

$$\tilde{A} = B_0 S(0) B_0', \tag{3.6}$$

$$\tilde{h}_{1i} = \operatorname{tr}[D_i' B_0 W], \tag{3.7}$$

$$\tilde{h}_{2i} = \operatorname{tr}[E_i' B_0 W], \tag{3.8}$$

and

$$\tilde{h}_{3i} = \operatorname{tr}[-F_i' B_0 W B_2' A_2^{-1}]. \tag{3.9}$$

Obviously this diffusion limit is different from the right diffusion limit given in § 2. In fact, comparing (3.5)-(3.9) with (2.6)-(2.12) shows that the two operators coincide only as $\gamma \to \infty$ (i.e., $\mu/\varepsilon \to 0$), indicating that the use of the reduced-order model (3.3) is acceptable only if $\mu \ll \varepsilon$.

As a generic example that illustrates the above situation consider the feedback control system of Fig. 1. Our interest here is not in designing such a control system, although this is an interesting problem for which the order reduction results of this paper could be useful, but in analyzing the behavior of x once the system has been built. The singularly perturbed equation in the forward loop represents actuator dynamics, amplified dynamics or *RC* or *RL* sections, which typically have small time constants represented here by the small parameter μ. In modeling such feedback systems, it is the practice of engineers to simplify the model by neglecting the fast dynamics of actuators, amplifiers, etc., when the time constant μ is sufficiently small with respect to the relaxation time of the plant. Such simplification leads to the feedback system of Fig. 2 in which the dynamic equation $\mu\dot{y} = Hy + Ke$ is replaced by the algebraic equation $y = -H^{-1}Ke$. Readers who are not familiar with this engineering practice can realize how common that practice is by noticing that whenever an amplifier in a control system is modeled by a constant gain k, such model simplification has been already employed since a more realistic model of the amplifier would be the first-order transfer function $k/(1+s\tau)$ or even higher-order transfer functions. If the driving inputs r and n are smooth deterministic functions of time, neglecting parasitic elements is justified by invoking singular perturbation results [1]. To analyze the situation when n is wide-band noise that can be modeled as $n(t) = (1/\sqrt{\varepsilon})v(t/\varepsilon)$, we invoke the result of § 2. The feedback system of Fig. 1 is represented by

$$\dot{x} = f(x) + G(x)y, \tag{3.10}$$

$$\mu\dot{y} = K(r-x) + Hy + Kn, \tag{3.11}$$

which fits the structure of equations (2.1), (2.2). Suppose that $f(x)$ and $G(x)$ are smooth enough that our technical conditions are satisfied, H is a stability matrix, and r is a constant set point. Then, for sufficiently small ε and μ, the behavior of x can be approximated by a diffusion process with infinitesimal generator (2.5) which is dependent on γ. If the simplified feedback system of Fig. 2 is used to analyze the behavior of x, then for sufficiently small ε, x can be approximated by a diffusion process whose infinitesimal generator is given by (3.4). This shows the invalidity of

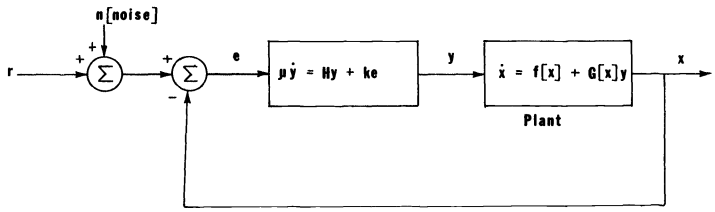


FIG. 1

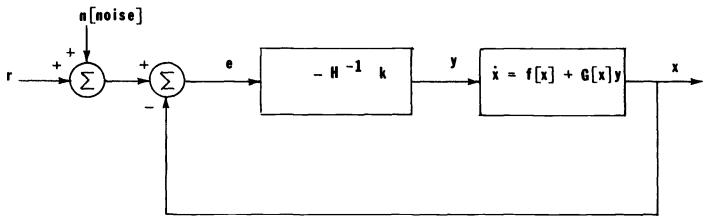


FIG. 2

the simplified feedback system of Fig. 2 as a model for analyzing x. To make the example more transparent, let us consider a specific situation where $x \in R$, $y \in R^2$,

$$G(x) = [c^{-x}, 1], \qquad H = \begin{bmatrix} -1 & 0 \\ 0 & -2 \end{bmatrix}, \qquad K = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

and suppose that the correlation function $R(\tau)$ is given by $R(\tau) = \sigma^2 e^{-|\tau|}$. Then, as $\varepsilon, \mu \to 0$, $x(\cdot)$ converges weakly to $\bar{x}(\cdot)$ which satisfies the Ito equation

$$d\bar{x} = \left[f(\bar{x}) + \left(\frac{1}{2} + e^{-\bar{x}}\right)(r - \bar{x}) - \frac{3\gamma^2 + 6\gamma + 2}{3(\gamma+1)(2\gamma+1)} \sigma^2 e^{-\bar{x}} - \sigma^2 e^{-2\bar{x}} \right] dt$$

$$+ \sqrt{2}\sigma^2 \left[\frac{1}{2} + e^{-\bar{x}} \right] dw, \qquad \bar{x}(0) = x_0$$

where $\gamma = \lim_{\varepsilon,\mu \to 0} (\varepsilon/\mu)$.

On the other hand, the diffusion process corresponding to (3.4) satisfies the Ito equation

$$d\tilde{x} = [f(\tilde{x}) + (\tfrac{1}{2} + e^{-\tilde{x}})(r - \tilde{x}) + \tfrac{1}{2}\sigma^2 e^{-\tilde{x}} - \sigma^2 e^{-2\tilde{x}}]\, dt + \sqrt{2}\sigma^2 [\tfrac{1}{2} + e^{-\tilde{x}}]\, dw, \qquad \tilde{x}(0) = x_0.$$

It is apparent that for all finite $\gamma > 0$, the two Ito equations are different. They coincide only as $\gamma \to \infty$.

The above discussions have shown that reducing the order of dynamic systems driven by wide-band stochastic inputs via neglecting parasitic elements is, in general, valid only when the frequency band of the neglected fast dynamics is much wider than the frequency band of the input. There are, however, several interesting special cases where neglecting parasitic elements is valid even when the frequency band of the input is of the same order of the frequency band of the fast dynamics. These are the special cases for which the operator L^γ is independent of γ. Using the explicit form of the operator L^γ given by (2.7)-(2.17), we can easily identify such special cases. For instance, it is apparent that for linear systems, as long as the behavior of the slow variable x is concerned, μ can be set to zero. This conclusion for linear systems can be interpreted in the following way. Although replacing the differential equation (2.2) by the algebraic equation (3.1) cannot be used to approximate y, it can be used as an input to the slow equation (2.1) since the error resulting from this approximation will be filtered out by the slow equation. The same conclusion holds even when the input is modeled as white noise [14]. From (2.7)-(2.17), we can see also that the nonlinearity of a_1 and a_{21} is not the source of the inconsistency encountered here. It is the nonlinearities in B_1, B_2 and A_{12} that give rise to the correction terms h_1, h_2 and h_3, respectively, bringing in the dependence on γ through the matrix Σ. Recall that

$$\Sigma = \int_0^\infty e^{A_2 \gamma \tau} B_2 R'(\tau)\, d\tau$$

and suppose, without loss of generality, that R is normalized in such a way that B_2 is proportional to the root-mean square of the noise input. It is apparent that if $B_2 = 0$ (no noise input to the singularly perturbed equation) or B_2 is sufficiently small (small noise input to the singularly perturbed equation), then the dependence of L^γ on γ will either vanish or be insignificant, respectively. The small noise case is particularly important and arises frequently in applications cf. [15]. For example, in the feedback system of Fig. 1, let $n(t)$ have a correlation function $\lambda^2 e^{-|\tau|/\theta}$ where λ is of order one and θ is sufficiently small. The power spectrum of $n(t)$ is $2\theta\lambda^2/(1 + \omega^2\theta^2)$ which is flat

over a frequency band of order $1/\theta$. As $\theta \to 0$, $n(t)$ does not tend to white noise but $n/\sqrt{\theta}$ tends to white noise. Taking $\varepsilon = \theta$, $n(t)$ can be modeled as $n(t) = v(t/\varepsilon)$ where $v(t)$ has $\lambda^2 e^{-\tau}$ as its correlation. For this $n(t)$, equation (3.10), (3.11) take the form

$$\dot{x} = f(x) + G(x)y,$$

$$\mu\dot{y} = K(r-x) + Hy + \sqrt{\varepsilon}\, Kv^{\varepsilon},$$

which is of the form (2.1), (2.2) except that B_2 is multiplied by $\sqrt{\varepsilon}$.

Checking the proof of Theorem 1, it can be easily seen that the following corollary holds for the small noise case.

COROLLARY 2. *Consider the system* (2.1), (2.2) *except that* B_2 *is replaced by* βB_2. *Assume that* (A1)-(A5) *hold. Let* $\hat{x}(\cdot)$ *be a diffusion process whose infinitesimal generator is given by* (2.7)-(2.17) *with* $B_2 = 0$ *and let* $x(0) = x_0$. *Then* $x^{\varepsilon,\mu,\beta}(\cdot)$ *converges weakly to* $\hat{x}(\cdot)$ *as* $\varepsilon \to 0$, $\mu \to 0$ *and* $\beta \to 0$.

A similar corollary holds for the Gaussian case of Theorem 2.

Finally, we conclude our discussions by pointing out that the restriction $\varepsilon/\mu \geqq \gamma > 0$ which has been imposed in this paper is purely technical and that the form of the operator L^{γ} derived in § 2 is actually valid as $\gamma \to 0$. This can be shown by studying the case $\varepsilon/\mu \to 0$ as two sequential limiting processes $\varepsilon \to 0$ followed by $\mu \to 0$, and employing asymptotic results from [3]-[5], [16]. The details of that are given in [17].

REFERENCES

[1] P. V. KOKOTOVIC, R. E. O'MALLEY, JR. AND P. SANNUTI, *Singular perturbation and order reduction in control theory—an overview*, Automatica, 12 (1976), pp. 123-132.

[2] G. C. PAPANICOLAOU AND W. KOHLER, *Asymptotic theory of mixing stochastic ordinary differential equations*, Comm. Pure Appl. Math., 27 (1974), pp. 641-668.

[3] G. BLANKENSHIP AND G. C. PAPANICOLAOU, *Stability and control of stochastic systems with wide-band noise disturbance*, SIAM J. Appl. Math., 34 (1978), pp. 437-476.

[4] H. J. KUSHNER, *Jump-diffusion approximations for ordinary differential equations with wide-band random right hand side*, this Journal, 17 (1979), pp. 729-744.

[5] ———, *Stochastic Stability and Control*, Academic Press, New York, 1967.

[6] G. BLANKENSHIP AND S. SACHS, *Singularly perturbed linear stochastic ordinary differential equations*, SIAM J. Math. Analysis, 10 (1979), pp. 306-320.

[7] G. BLANKENSHIP, *On the Separation of Time Scales in Stochastic Differential Equations*, Proc. 7th IFAC Congress, Helsinki, 1978, pp. 937-944.

[8] V. D. RAZVIG, *Reduction of stochastic differential equations with small parameters and stochastic integrals*, Int. J. Control, 28 (1978), pp. 707-720.

[9] H. J. KUSHNER, *A martingale method for the convergence of a sequence of processes to a jump-diffusion process*, Z. Wahrsch. Verw. Gebiete, 53 (1980), pp. 207-219.

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

[11] D. W. STROOCK AND S. R. S. VARADHAN, *Multidimensional Diffusion Processes*, Springer-Verlag, New York, 1979.

[12] T. G. KURTZ, *Semigroups of conditioned shifts and approximation of Markov processes*, Ann. Probab., 3 (1975), pp. 618-642.

[13] H. J. KUSHNER AND Y. BAR-NESS, *Analysis of nonlinear stochastic systems with wide-band inputs*, IEEE Trans. Automat. Control, AC-25 (1980), pp. 1072-1078.

[14] A. HADDAD, *Linear filtering of singularly perturbed systems*, IEEE Trans. Automat. Control, AC-21 (1976), pp. 515-519.

[15] Z. SCHUSS, *Singular perturbation methods in stochastic differential equations of mathematical physics*, SIAM Review, 22 (1980), pp. 119-155.

[16] G. C. PAPANICOLAOU, D. STROOCK AND S. R. S. VARADHAN, *Martingale approach to some limit theorems*, Statistical Mechanics and Dynamical Systems, Duke Turbulence Conference, M. Reed, ed., Duke Univ. Mathematics Series, 3 (1977), Durham, NC.

[17] M. G. EL-ANSARY, *Stability and control of nonlinear singularly perturbed stochastic systems*, Ph.D. Dissertation, Michigan State Univ., East Lansing, 1983.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 1, January 1986

ON SEMINORMALITY OF INTEGRAL FUNCTIONALS AND THEIR INTEGRANDS*

E. J. BALDER†

Dedicated to E. J. McShane on the occasion of his 80th birthday.

Abstract. We present a definition of seminormality which extends classical notions due to Tonelli [Fondamenti di Calcolo delle Variazioni, 1921], McShane [Ann Scuola Norm. Sup. Pisa (2), 3 (1934), pp. 181-211, 287-315] and Cesari [Trans. Amer. Math. Soc., 124 (1966), pp. 369-412; J. Optim. Theory Appl., 6 (1970), pp. 114-137; SIAM J. Control Optim., 9 (1971), pp. 287-315], and applies to both integrands ("seminormality in the small") and their integral functionals ("seminormality in the large"). Our main result states that under very general hypotheses seminormality in the small and large are equivalent. By introducing a notion called Nagumo tightness, we show that the usual sufficient conditions for the lower semicontinuity of an integral functional also imply a form of seminormality of the integral functional. Necessary conditions for the lower semicontinuity of an integral functional can also be obtained from our results.

Key words. seminormality, integral functionals, lower semicontinuity, lower closure, tightness, property (Q).

1. Introduction. Let (X, d) be a metric space and $(V, P, \langle \cdot, \cdot \rangle)$ a pair of locally convex spaces paired by the strict duality (or nondegenerate bilinear form) $\langle \cdot, \cdot \rangle$ on $V \times P$. Let $\bar{\mathbb{R}} \equiv [-\infty, +\infty]$ denote the set of extended real numbers; we shall use the convention $(+\infty) - (+\infty) \equiv +\infty$ throughout.

A function $a: X \times V \to \bar{\mathbb{R}}$ is defined to be *simple seminormal* (on $X \times V$) if there exist a lower semicontinuous function $f: X \to \bar{\mathbb{R}}$ and $p \in P$ such that

$$a(x, v) = f(x) + \langle v, p \rangle.$$

A function $a: X \times V \to \bar{\mathbb{R}}$ is defined to be *seminormal* (on $X \times V$) if it is the (pointwise) supremum of a collection of simple seminormal functions on $X \times V$. An equivalent, more traditional definition of seminormality will be given in § 2, where this property will be studied in detail.

Now suppose in addition that X, V and P are Suslin spaces (Appendix B). Let $(T, \mathcal{T}, \mu)$ be a σ-finite measure space and let $\mathcal{X}$ be a decomposable set of equivalence classes of $(\mathcal{T}, \mathcal{B}(X))$-measurable functions, equipped with the essential supremum metric

$$d(x, y) \equiv \operatorname*{ess\,sup}_{t \in T} d(x(t), y(t)).$$

Also, let $(\mathcal{V}, \mathcal{P}, \langle \cdot, \cdot \rangle)$ be a pair of decomposable vector spaces of equivalence classes of scalarly μ-integrable functions (cf. Appendix B) going from T into V and P respectively, paired by the strict duality

$$\langle v, p \rangle \equiv \int_T \langle v(t), p(t) \rangle \mu(dt),$$

where it is assumed that for every $v \in \mathcal{V}$, $p \in \mathcal{P}$ the function $t \mapsto \langle v(t), p(t) \rangle$ is μ-integrable.

Let $l: T \times X \times V \to \bar{\mathbb{R}}$ be a function which is such that for some $p_0 \in \mathcal{P}$ and $\phi_0 \in \mathcal{L}^1_{\mathbb{R}}$

$$l(t, x, v) \geqq \langle v, p_0(t) \rangle + \phi_0(t) \text{ for all } x \in X \text{ and } v \in V \ \mu\text{-a.e.} \tag{1.1}$$

* Received by the editors August 16, 1983, and in final revised form February 4, 1985.
† Mathematical Institute, University of Utrecht, Utrecht, the Netherlands.

By means of outer integration (Appendix A) we define the integral functional $I_l: \mathscr{X} \times \mathscr{V} \to \bar{\mathbb{R}}$ as follows:

$$I_l(x, v) \equiv \tilde{\int}_T l(t, x(t), v(t)) \mu\,(dt).$$

We can now view seminormality at the following two levels:

(i) *Seminormality in the small*: seminormality of the function $l(t, \cdot, \cdot): X \times V \to \bar{\mathbb{R}}$ μ-a.e. (i.e., for μ-almost every $t \in T$). Here seminormality is defined with respect to the framework consisting of (X, d) and $(V, P, \langle \cdot, \cdot \rangle)$.

(ii) *Seminormality in the large*: seminormality of the integral functional $I_l: \mathscr{X} \times \mathscr{V} \to \bar{\mathbb{R}}$ of l. Here seminormality is defined with respect to the framework which is composed of $(\mathscr{X}, d)$ and $(\mathscr{V}, \mathscr{P}, \langle \cdot, \cdot \rangle)$, introduced above.

While seminormality in the small has a long history, starting with the seminal work by Tonelli [27] and McShane [21], seminormality in the large seems to be an entirely new concept. The main explanation for this rather surprising observation can undoubtedly be found in the traditional occupation with integral functionals of a single variable, such as exemplified by

$$I(x) \equiv \int_0^1 l(t, x(t), \dot{x}(t))\, dt$$

where x ranges over a set of smooth curves; e.g. cf. [10].

The principal result of this paper, which will be presented in § 3, states that under quite general hypotheses seminormality in the small and in the large are *equivalent.* There are essentially two ways in which this result can be used to shed new light on lower semicontinuity and lower closure questions for integral functionals (§ 4). The most important one is as follows. A function $h: V \to (-\infty, +\infty]$ is said to be of *Nagumo type* on V if h is convex and sequentially inf-compact on V for every slope. A subset $\mathscr{V}_0$ of $\mathscr{V}$ is defined to be *Nagumo tight* if there exists a $\mathscr{T} \times \mathscr{B}(X)$-measurable function $h: T \times V \to [0, +\infty]$ such that

$$h(t, \cdot) \text{ is of Nagumo type on } V \ \mu\text{-a.e.},$$

$$\sup_{v \in \mathscr{V}_0} I_h(v) < +\infty.$$

Also, $\mathscr{V}_0 \subset \mathscr{V}$ is defined to be *almost Nagumo tight* if there exists a nonincreasing sequence $\{B_i\}_1^\infty$ in $\mathscr{T}$, whose intersection is a μ-null set, such that for every $i \in \mathbb{N}$ the restrictions to $T \setminus B_i$ of all elements of $\mathscr{V}_0$ form a Nagumo tight set. Important examples of almost Nagumo tightness are the following. (a) A subset $\mathscr{V}_0$ of $\mathscr{V}$ is Nagumo tight if there exists a scalarly measurable multifunction $\Gamma: T \rightrightarrows V$ having (sequentially) compact convex values such that $v(t) \in \Gamma(t)$ μ-a.e. for every $v \in \mathscr{V}_0$. (b) In case V is a separable reflexive Banach space, P its topological dual V' and $\mathscr{V} \equiv \mathscr{L}^1_V$, $\mathscr{P} \equiv \mathscr{L}^\infty_{V'}$, a subset $\mathscr{V}_0$ of $\mathscr{L}^1_V$ is Nagumo tight if it is relatively (sequentially) compact for the weak topology $\sigma(\mathscr{L}^1_V, \mathscr{L}^\infty_{V'})$. (c) In the same case as in (b) a sequence in $\mathscr{V}_0$ has an almost Nagumo tight subsequence if it is uniformly bounded in L^1-norm.

Now suppose that $l: T \times X \times V \to \bar{\mathbb{R}}$ satisfies the following well-known sufficient conditions for sequential lower semicontinuity of I_l on $\mathscr{X} \times \mathscr{V}$:

$l(t, \cdot, \cdot)$ is sequentially lower semicontinuous on $X \times V$ μ-a.e.,

$l(t, x, \cdot)$ is convex on V for every $x \in X$ μ-a.e.

Define $l_1: T \times X \times V \times \mathbb{R} \to \bar{\mathbb{R}}$ by

$$l_1(t, x, v, \lambda) \equiv \max(l(t, v, x), \lambda).$$

Then it follows from the main result that for every $\mathcal{V}_0 \subset \mathcal{V}$, $\mathcal{L}_0 \subset \mathcal{L}^1_{\mathbb{R}}$ such that $\mathcal{V}_0$ is almost Nagumo tight and $\mathcal{L}_0^- \equiv \{\lambda^-: \lambda \in \mathcal{L}_0\}$ is uniformly integrable, I_{l_1} has the following *coincident seminormality* property:

$$I_{l_1} \text{ coincides on } \mathcal{X} \times \mathcal{V}_0 \times \mathcal{L}_0 \text{ with a seminormal function,}$$

where $I_{l_1}: \mathcal{X} \times \mathcal{V} \times \mathcal{L}^1_{\mathbb{R}} \to \bar{\mathbb{R}}$ is defined by

$$I_{l_1}(x, v, \lambda) \equiv \int_T^{\tilde{}} l_1(t, x(t), v(t), \lambda(t))\, \mu\,(dt),$$

and seminormality is considered with respect to the framework consisting of $(\mathcal{X}, d)$ and the pair $(\mathcal{V} \times \mathcal{L}^1_{\mathbb{R}}, \mathcal{P} \times \mathcal{L}^\infty_{\mathbb{R}})$, equipped with the duality

$$\langle\langle v, \lambda; p, q \rangle\rangle \equiv \int_T (\langle v(t), p(t) \rangle + \lambda(t) q(t)) \mu\,(dt).$$

Coincident seminormality being a much stronger property than relative lower semicontinuity, one thus generalizes a whole class of results. These ideas can be carried over to the subject of lower closure without much difficulty.

The second way in which the main result can be used concerns the usual necessary conditions for lower semicontinuity of integral functionals. This line of approach could be of some interest as a contribution to unifying necessary and sufficient conditions for lower semicontinuity of integral functionals. Although our result is quite general in most respects, the boundedness condition (1.1) makes it somewhat incomparable to similar results obtained elsewhere.

2. Seminormality. In this section we study several aspects of seminormality. We introduce the seminormal hull of a function and show it to be representable by semicontinuous hulls and Fenchel conjugation. Also, we give some sufficient conditions for seminormality to hold. Finally, we show the close relationship which exists between the notion of seminormality developed here and Cesari's property (Q) for multifunctions [8]. Let us note in advance that although this section is strictly treated in terms of the framework for seminormality in the small, most results obtained here can be used in connection with seminormality in the large by a simple substitution of framework.

Let (X, d) be a metric space. For every $x \in X$ and $\delta > 0$ the set of all $y \in X$ such that $d(x, y) < \delta$ ($d(x, y) \leqq \delta$) will be denoted by $B(x; \delta)$ ($\bar{B}(x; \delta)$). Also, let $(V, P, \langle \cdot, \cdot \rangle)$ be a pair of locally convex spaces, paired by the strict duality $\langle \cdot, \cdot \rangle: V \times P \to \mathbb{R}$. The topologies on V and P are understood to be compatible with the duality; note that strictness of the duality causes these topologies to be Hausdorff [9, 5.22].

Let $a: X \times V \to \bar{\mathbb{R}}$ be a given function. Following Tonelli [27] and McShane [21] (see also Cesari [8]), we introduce a notion of seminormality for extended real-valued functions by defining a to be *seminormal* at the point $(x, v) \in X \times V$ if for every $\alpha \in \mathbb{R}$, $\alpha < a(x, v)$, there exist $p \in P$, $\beta \in \mathbb{R}$, $\delta > 0$ such that

$$a \geqq \langle \cdot, p \rangle + \beta - \chi(\cdot; B(x, \delta)), \tag{2.1}$$

$$\alpha < \langle v, p \rangle + \beta. \tag{2.2}$$

As usual, the indicator function $\chi(\cdot; B)$ of a subset B of X is given by $\chi(y; B) \equiv 0$ if

$y \in B$, $\chi(y; B) \equiv +\infty$ if $y \notin B$. Also, we say that $a: X \times V \to \bar{\mathbb{R}}$ is *seminormal* (on $X \times V$) if a is seminormal at every point of $X \times V$. The following observation is obvious, but important.

Remark 2.1. The (pointwise) supremum of an arbitrary collection of seminormal functions on $X \times V$ is seminormal. Also, the sum of a finite collection of seminormal functions on $X \times V$ is seminormal, provided that no addition of values $+\infty$ and $-\infty$ takes place. □

As a consequence of this remark, it makes sense to define the *seminormal hull* $\tilde{\tilde{a}}: X \times V \to \bar{\mathbb{R}}$ of the function $a: X \times V \to \bar{\mathbb{R}}$ to be the supremum of the collection of all seminormal functions on $X \times V$ which are (pointwise) no larger than a; note that $\tilde{\tilde{a}}$ is seminormal on $X \times V$ by Remark 2.1 and that

$$\tilde{\tilde{a}} \leqq a. \tag{2.3}$$

By means of the hull concept we can take a different look at local seminormality.

PROPOSITION 2.2. *For every $x \in X$, $v \in V$ the following are equivalent*:

$$a \text{ is seminormal at } (x, v), \tag{2.4}$$

$$\tilde{\tilde{a}}(x, v) = a(x, v). \tag{2.5}$$

Proof. If (2.5) holds, then (2.3) and seminormality of $\tilde{\tilde{a}}$ imply (2.4) directly. Conversely, if (2.4) is true, then for every $\alpha \in \mathbb{R}$, $\alpha < a(x, v)$, there exist $p \in P$, $\beta \in \mathbb{R}$, $\delta > 0$ such that $e \leqq a$ and $e(x, v) > \alpha$, where e is the function on the right in (2.1). Since e is obviously seminormal on $X \times V$, we find $\tilde{\tilde{a}}(x, v) > \alpha$, which shows that $\tilde{\tilde{a}}(x, v) \geqq a(x, v)$. By (2.3) we thus have (2.5). □

Remark 2.3. The above proof shows that the seminormal hull of a is precisely the supremum of all functions $e \equiv \langle \cdot, p \rangle + \beta - \chi(\cdot; B(y; n^{-1}))$, $p \in P$, $\beta \in \mathbb{R}$, $n \in \mathbb{N}$, satisfying $e \leqq a$; note that the latter inequality implies

$$-\beta \geqq c(n, y, p) \equiv \sup_{z \in X, w \in V} [\langle w, p \rangle - a(z, w) - \chi(z; B(y; n^{-1}))].$$

Hence, we obtain

$$\tilde{\tilde{a}}(x, v) = \sup_{n \in \mathbb{N}, y \in X, p \in P} [\langle v, p \rangle - c(n, y, p) - \chi(x; B(y; n^{-1}))].$$

Further, note that e as above is evidently simple seminormal, as defined in § 1. Hence, it has been shown in the above proof that the seminormal hull of a is the supremum of the collection of all simple seminormal functions on $X \times V$ which are (pointwise) no larger than a. By Proposition 2.2 this shows the equivalence of the seminormality definitions given in § 1 and § 2.

Let us now express the seminormal hull a of $\tilde{\tilde{a}}$ in terms of the functions $b: X \times P \to \bar{\mathbb{R}}$ and $\bar{b}: X \times P \to \bar{\mathbb{R}}$, which we define as follows by Fenchel conjugation of a with respect to the variable v and semicontinuous hulls with respect to the variable x [7, I.4]:

$$b(y, p) \equiv a^*(y, p) \equiv \sup_{v \in V} [\langle v, p \rangle - a(y, v)],$$

$$\bar{b}(x, p) \equiv \limsup_{y \to x} b(y, p).$$

THEOREM 2.4. *For every $x \in X$, $v \in V$*

$$\tilde{\tilde{a}}(x, v) = \bar{b}^*(x, v) \equiv \sup_{p \in P} [\langle v, p \rangle - \bar{b}(x, p)]. \tag{2.6}$$

Proof. By definition of $\bar{b}$ we have $\bar{b} \geqq b$; this implies $\bar{b}^* \leqq b^* = a^{**} \leqq a$. From (2.6) it is clear that $\bar{b}^*$ is the supremum of a collection of simple seminormal functions.

Hence, $\bar{b}^*$ is seminormal (cf. Remark 2.3). By the above this implies $\tilde{\tilde{a}} \geqq \bar{b}^*$. On the other hand, let e be a simple seminormal function on $X \times V$ such that $e \leqq a$. For such a function one has trivially that $e^*: X \times P \to \bar{\mathbb{R}}$ is upper semicontinuous in the variable x. Hence the obvious inequality $e^* \geqq b$ implies $e^* \geqq \bar{b}$. By [7, I.4] (or ad hoc inspection) we have $e = e^{**}$. Combined, this gives $e \leqq \bar{b}^*$. By Remark 2.3 we conclude that $\tilde{\tilde{a}} \leqq \bar{b}^*$. This finishes the proof. □

Remark 2.5. It will be convenient to rewrite (2.6) in the following two ways:

$$\tilde{\tilde{a}}(x, v) = \sup_{p \in P} \lim_{n \to \infty} \uparrow \inf_{y \in X, w \in V} [\langle v - w, p \rangle + a(y, w) + \chi(y; B(x; n^{-1}))]$$

$$= \lim_{n \to \infty} \uparrow \sup_{p \in P} \inf_{y \in X, w \in V} [\langle v - w, p \rangle + a(y, w) + \chi(y; \bar{B}(x; n^{-1}))].$$

Remark 2.6. For every function $f: X \to \bar{\mathbb{R}}$, considered as a function on $X \times V$, we have that the seminormal hull $\tilde{\tilde{f}}$ of f is given by

$$\tilde{\tilde{f}}(x) = \liminf_{y \to x} f(y) \equiv \bar{f}(x).$$

Also, for every function $g: V \to \bar{\mathbb{R}}$, considered as a function on $X \times V$, we have that the seminormal hull $\tilde{\tilde{g}}$ of g is given by

$$\tilde{\tilde{g}}(v) = g^{**}(v).$$

Hence, the seminormal hull of $f: X \to \bar{\mathbb{R}}$ coincides with its lower semicontinuous hull $\bar{f}$, and the seminormal hull of $g: V \to \bar{\mathbb{R}}$ coincides with the Fenchel biconjugate g^{**}; the latter is known to be the lower semicontinuous convex hull of g, provided that g is affinely bounded from below [7, I.5]. Thus, the seminormal hull concept straddles two important hulls.

Let us observe that for every $x \in X$ seminormality of a at every point of $\{x\} \times V$ implies trivially that a is lower semicontinuous at every point of $\{x\} \times V$ and that $a(x, \cdot)$ is convex on V. We shall now establish a kind of converse to this implication, which in a somewhat less general form constitutes a classical sufficient condition for seminormality [8], [27]. Let us recall that a function $h: V \to (-\infty, +\infty]$ is said to be *sequentially inf-compact on V for every slope* if for every $p \in P$, $\gamma \in \mathbb{R}$ the set of all $v \in V$ such that $h(v) - \langle v, p \rangle \leqq \gamma$, is sequentially compact [7], [20]; of course then for every $p \in P$ there exists $\beta \in \mathbb{R}$ such that

$$h \geqq \langle \cdot, p \rangle + \beta \tag{2.7}$$

since the infimum of $h - \langle \cdot, p \rangle$ over V is attained somewhere (Weierstrass' theorem). A function from V into $(-\infty, +\infty]$ which is both convex and sequentially inf-compact on V for every slope will be called of *Nagumo type* on V in this paper; cf. [22].

Example 2.7. Suppose that $(V, P, \langle \cdot, \cdot \rangle)$ is specified as follows: V is a reflexive Banach space, whose norm we denote by $\| \cdot \|$, P is the topological dual V' of $(V, \| \cdot \|)$, $\langle \cdot, \cdot \rangle$ is the usual duality, and the topology on V is $\sigma(V, V')$. Then a function $h: V \to (-\infty, +\infty]$ is of Nagumo type if there exists a nondecreasing lower semicontinuous convex function $h' = [0, +\infty) \to (-\infty, +\infty]$ such that $h(v) \equiv h'(\|v\|)$ and

$$\lim_{\gamma \to \infty} \gamma^{-1} h'(\gamma) = +\infty. \tag{2.8}$$

Of course, this follows from the fact that every norm-bounded subset of V is relatively sequentially compact for the topology $\sigma(V, V')$ on V [16].

From now on until Remark 2.14 we make the additional assumption (cf. [14, p. 30 ff.]):

(C) V has countably determined compactness.

THEOREM 2.8. *For every $x \in X$ we have that if*

(2.9) *a is sequentially lower semicontinuous at every point of $\{x\} \times V$,*

(2.10) *$a(x, \cdot)$ is convex on V,*

and if there exist a function $h: V \to (-\infty, +\infty]$ and $\delta > 0$ such that

(2.11) *h is sequentially inf-compact on V for every slope,*

(2.12) *$a(y, \cdot) \geqq h$ for all $y \in B(x; \delta)$,*

then

a is seminormal at every point of $\{x\} \times V$.

COROLLARY 2.9. *For every $x \in X$ and $\varepsilon > 0$ we have that if* (2.9)-(2.10) *hold, if further there exist $p_0 \in P$, $\beta_0 \in \mathbb{R}$ and $\delta > 0$ such that*

(2.13) *$a(y, v) \geqq \langle v, p_0 \rangle + \beta_0$ for all $y \in B(x; \delta)$ and $v \in V$,*

and if $h: V \to (-\infty, +\infty]$ is of Nagumo type on V, then

$a + \varepsilon h$ is seminormal at every point of $\{x\} \times V$.

Proof. The conditions (2.9)-(2.12) of Theorem 2.8 are obviously satisfied if we substitute $a + \varepsilon h$ for a and $\langle \cdot, p_0 \rangle + \beta_0 + \varepsilon h$ for h. □

Remark 2.10. By assumption (C) any subset of V is (relatively) sequentially compact if and only if it is (relatively) compact. Hence, (2.9)-(2.12) in Theorem 2.8 imply

(2.10′) $a(x, \cdot)$ is lower semicontinuous and convex on V.

A similar observation holds for $a + \varepsilon h$ in Corollary 2.9.

We shall prepare the proof of Theorem 2.8 by giving a simple but very useful generalization of Dini's theorem. This result can also be regarded as a generalization of the well-known "theorem of the maximum" [2], [20, p. 358]; it is a special case of [29, Thm. 1].

LEMMA 2.11. *Let E be a topological Hausdorff space and $\{f_n\}_0^\infty$ a nondecreasing sequence of functions from E into $(-\infty, +\infty]$. If*

(2.14) $$\lim_{n \to \infty} \uparrow f_n(x) = f_0(x) \text{ for every } x \in E,$$

(2.15) $$\gamma_n \equiv \inf_{x \in E} f_n(x) > -\infty \text{ for every } n \in \mathbb{N},$$

and if there exists $\Omega \subset E$ such that for every $n \in \mathbb{N}$

(2.16) *f_n is sequentially lower semicontinuous at every point of Ω,*

and such that for every $\varepsilon > 0$ and every sequence $\{x_n\}_1^\infty \subset E$ which satisfies

(2.17) $$f_n(x_n) \leqq \gamma_n + \varepsilon \text{ for every } n \in \mathbb{N},$$

there exists a subsequence of $\{x_n\}_1^\infty$ which converges to some point in Ω. Then

$$\lim_{n \to \infty} \uparrow \gamma_n = \inf_{x \in \Omega} f_0(x).$$

Proof. Let us write $\gamma_0 \equiv \inf_\Omega f_0$; it is obvious that $\gamma_0 \geqq \lim_n \gamma_n$. Conversely, for arbitrary $\varepsilon > 0$ there exists by (2.15) a sequence $\{x_n\}$ satisfying (2.17). By hypothesis there exists a subsequence $\{x_{n'}\}$ of $\{x_n\}$ which converges to some point x_* in Ω. By using (2.14), (2.16) we find for every $k \in \mathbb{N}$

$$\liminf_{n'} f_{n'}(x_{n'}) \geqq \liminf_{n'} f_k(x_{n'}) \geqq f_k(x_*).$$

By (2.14), (2.17) this gives $\lim_n \gamma_n \geqq f_0(x_*) - \varepsilon$. It is now easy to finish the proof. □

Proof of Theorem 2.8. Let $v \in V$ be arbitrary. If $\tilde{\tilde{a}}(x, v) = +\infty$, then seminormality of a at (x, v) is a consequence of (2.3) and Proposition 2.2. So suppose now that $\tilde{\tilde{a}}(x, v) < +\infty$. Let $p \in P$ be arbitrary. We define

$$f_n(y, w) \equiv \langle v - w, p\rangle + a(y, w) + \chi(y; \bar{B}(x; n^{-1})),$$
$$f_0(y, w) \equiv \langle v - w, p\rangle + a(y, w) + \chi(y; \{x\}).$$

We shall apply Lemma 2.11 with $E \equiv X \times V$ and $\Omega \equiv \{x\} \times V$ (note that we could also have worked with open balls $B(x; n^{-1})$ in the definition of f_n). The conditions (2.14)-(2.16) obviously hold by virtue of (2.8)-(2.9) and (2.11)-(2.12). Let $\{(y_n, w_n)\}$ be as in (2.17) (viz., a sequence of ε-almost minimizers of $\{f_n\}$). Then by Remark 2.5, (2.12) and evident properties of indicator functions we have for every $n \in \mathbb{N}$

$$h(w_n) - \langle w_n, p\rangle \leqq \tilde{\tilde{a}}(x, v) - \langle v, p\rangle + \varepsilon,$$
$$y_n \in \bar{B}(x; n^{-1}).$$

From (2.11) and the above it follows that $\{(y_n, w_n)\}$ has a subsequence which converges to some point in $\{x\} \times V$. Hence we may apply Lemma 2.11 and we obtain

$$\lim_{n\to\infty} \uparrow \inf_{y \in X, w \in V} f_n(y, w) = \inf_{y \in X, w \in V} f_0(y, w).$$

By Remark 2.5 this gives

$$\tilde{\tilde{a}}(x, v) = \sup_{p \in P} [\langle v, p\rangle - b(x, p)] = a^{**}(x, v).$$

By [7, I.4] it follows from (2.10′) that $a^{**}(x, v) = a(x, v)$. In view of Proposition 2.2 this finishes the proof. □

Thus far, we studied seminormality of a in the framework composed of (X, d) and $(V, P, \langle \cdot, \cdot \rangle)$. Next to this, we shall now also consider the seminormality of functions from $X \times V \times \mathbb{R}$ into $\bar{\mathbb{R}}$ with respect to the framework which consists of (X, d) and $(V \times \mathbb{R}, P \times \mathbb{R}, \langle\langle \cdot; \cdot \rangle\rangle)$, where the duality $\langle\langle \cdot; \cdot \rangle\rangle$ is defined by

$$\langle\langle v, \lambda; p, q\rangle\rangle \equiv \langle v, p\rangle + \lambda q.$$

Let $a_1 : X \times V \times \mathbb{R} \to \bar{\mathbb{R}}$ correspond to a by the relation

$$a_1(x, v, \lambda) \equiv \max(a(x, v), \lambda).$$

We shall first present two analogues of Corollary 2.9. The first of these is straightforward and will not be used later on. The second analogue shows that it is possible to relinquish the boundedness condition (2.13) by adding to a_1 an additional growth term for the negative part of the variable λ.

PROPOSITION 2.12. *For every $x \in X$ and $\varepsilon > 0$ we have that if* (2.9)-(2.10) *and* (2.13) *hold and if $h : V \to (-\infty, +\infty]$ is of Nagumo type on V, then*

$$a_1 + \varepsilon h \text{ is seminormal at every point of } \{x\} \times V.$$

Proof. We have by definition of a_1

$$(a_1+\varepsilon h)(x, v, \lambda)=\max((a+\varepsilon h)(x, v), \varepsilon h(v)+\lambda).$$

By Corollary 2.9 the function $(x, v, \lambda)\mapsto(a+\varepsilon h)(x, v)$ is seminormal at every point of $\{x\}\times V\times\mathbb{R}$. The function $(x, v, \lambda)\mapsto\varepsilon h(v)+\lambda$ is simply seminormal. Hence, the result follows by Remark 2.1. □

THEOREM 2.13. *For every $x\in X$ and $\varepsilon>0$ we have that if* (2.9)-(2.10) *hold, if $h: V\to(-\infty, +\infty]$ is of Nagumo type on V and if $h': [0, +\infty)\to(-\infty, +\infty]$ is nondecreasing lower semicontinuous and convex, satisfying* (2.8), *then for the function $a_{1,\varepsilon}: X\times V\times\mathbb{R}\to\bar{\mathbb{R}}$, defined by*

$$a_{1,\varepsilon}(x, v, \lambda)\equiv a_1(x, v, \lambda)+\varepsilon h(v)+\varepsilon h'(\lambda^-),$$

where $\lambda^-\equiv\max(-\lambda, 0)$, we have

$$a_{1,\varepsilon} \text{ is seminormal at every point of } \{x\}\times V\times\mathbb{R}.$$

Proof. Let $v\in V$, $\lambda\in\mathbb{R}$ be arbitrary. Let $\tilde{\tilde{a}}_{1,\varepsilon}$ be the seminormal hull of $a_{1,\varepsilon}$. If $\tilde{\tilde{a}}_{1,\varepsilon}(x, v, \lambda)=+\infty$, then seminormality of $a_{1,\varepsilon}$ is immediate by (2.3) and Proposition 2.2. So suppose now that $\tilde{\tilde{a}}_{1,\varepsilon}(x, v, \lambda)<+\infty$. Let $p\in P$, $q\in\mathbb{R}$ be arbitrary and define

$$g_n(y, w, \kappa)\equiv\langle\langle v-w, \lambda-\kappa; p, q\rangle\rangle+a_{1,\varepsilon}(y, w, \kappa)+\chi(y; \bar{B}(x; n^{-1})),$$

$$g_0(y, w, \kappa)\equiv\langle\langle v-w, \lambda-\kappa; p, q\rangle\rangle+a_{1,\varepsilon}(y, w, \kappa)+\chi(y; \{x\}).$$

The first case to be considered is when $q<1$. We can then apply Lemma 2.11 with $E\equiv X\times V\times\mathbb{R}$, $\Omega\equiv\{x\}\times V\times\mathbb{R}$. Namely, conditions (2.14)-(2.16) hold obviously. Also, if $\{(y_n, w_n, \kappa_n)\}$ is a sequence of δ-almost minimizers of $\{g_n\}$ (cf. (2.17)), then by Remark 2.5 and obvious properties of indicator functions we have for every $n\in\mathbb{N}$

$$\varepsilon h(w_n)+\varepsilon h'(\kappa_n^-)+(1-q)\kappa_n-\langle w_n, p\rangle\leqq\tilde{\tilde{a}}_{1,\varepsilon}(x, v, \lambda)-\langle\langle v, \lambda; p, q\rangle\rangle+\delta,$$

$$y_n\in\bar{B}(x; n^{-1}).$$

From the Nagumo type property for h and h', and the inequality $q<1$ it follows then easily that $\{(y_n, w_n, \kappa_n)\}$ contains a subsequence converging to some point $\{x\}\times V\times\mathbb{R}$. By applying Lemma 2.11 we find

$$\lim_{n\to\infty}\uparrow\inf_{X\times V\times\mathbb{R}} g_n=\inf_{X\times V\times\mathbb{R}} g_0. \tag{2.18}$$

Next, we consider the case $q=1$. Then (2.18) continues to hold since for every $n\in\mathbb{N}\cup\{0\}$, $y\in X$ and $w\in V$ with $a(y, w)<+\infty$:

$$\inf_{\kappa\in\mathbb{R}} g_n(y, w, \kappa)=\langle v-w, p\rangle+\lambda+\varepsilon h(w)+\varepsilon h'(0)+\chi(y; \bar{B}(x; n^{-1})).$$

So this time we can apply Lemma 2.11 for $E\equiv X\times V$ and $\Omega\equiv\{x\}\times V$, in complete analogy to what was done in the proof of Theorem 2.8. It remains to consider the case when $q>1$. Then it is easy to see that $\inf_{X\times V\times\mathbb{R}} g_n=-\infty$ for every $n\in\mathbb{N}\cup\{0\}$; hence (2.18) continues to hold. We conclude that (2.18) holds in all three cases. By Remark 2.5 this leads to

$$\tilde{\tilde{a}}_{1,\varepsilon}(x, v, \lambda)=\sup_{p\in P, q\in\mathbb{R}}[\langle\langle v, \lambda; p, q\rangle\rangle-a_{1,\varepsilon}^*(x, p, q)]=a_{1,\varepsilon}^{**}(x, v, \lambda).$$

By [7, I.4] it follows from (2.10) and the properties of h and h' that $a_{1,\varepsilon}^{**}(x, v, \lambda)=a_{1,\varepsilon}(x, v, \lambda)$. The proof is finished by applying Proposition 2.2. □

Remark 2.14. The additional assumption (C) will now be dropped. For later reference we point out that (C) holds in either of the following two cases:

(i) P is separable (for example, Suslin).

(ii) V is normable for some topology for which P is its dual.

This follows from the Eberlein-Smulian theorem [14, p. 39].

The remainder of this section will be used to demonstrate the strong connections which exist between the notion of seminormality developed here and property (Q) for multifunctions, due to Cesari [8]. These connections can be phrased as follows: seminormality of a *finite-valued* function is equivalent to property (Q) of its epigraphic multifunction (cf. [8b, p. 134], [8d, p. 486] and Proposition 2.20 below) and property (Q) of a multifunction is equivalent to seminormality of the indicator function of the multifunction (cf. Proposition 2.15 below). Although these connections will not play a role further on, they do clarify the position of the present section in relation to a substantial part of the literature on lower semicontinuity and lower closure.

Let $Q: X \rightrightarrows V$ be a given multifunction; we shall not exclude the possibility that some or all values of Q are empty. Following Cesari [8] we say that the multifunction $Q: X \rightrightarrows V$ has *property* (Q) at $x \in X$ if

$$Q(x) = \bigcap_{\delta>0} \overline{\text{co}}\, Q(x; \delta), \tag{2.19}$$

where we denote the union of all sets $Q(y)$ with $d(y, x) < \delta$ by $Q(x; \delta)$. Note that one inclusion holds trivially in (2.19); as will soon be apparent, this corresponds precisely to (2.3) (note that the implicit hull concept thus given has not been used at all in the literature on property (Q)). We define the *indicator function* $\chi_Q: X \times V \to \bar{\mathbb{R}}$ of the multifunction Q by

$$\chi_Q(x, v) \equiv \begin{cases} 0 & \text{if } v \in Q(x), \\ +\infty & \text{if } v \notin Q(x). \end{cases}$$

In terms of our previous notation this means that $\chi_Q(x, v) = \chi(v; Q(x))$. Our next result states that property (Q) of Q is in fact a seminormality property of the indicator function χ_Q.

PROPOSITION 2.15. *For every $x \in X$ the following are equivalent:*

Q has property (Q) at x,

χ_Q is seminormal at every point of $\{x\} \times V$.

Proof. It is easy to see that

$$b_Q(y, p) \equiv \chi_Q^*(y, p) = \chi^*(p; Q(y)),$$
$$\bar{b}_Q(x, p) \equiv \limsup_{y \to x} b_Q(y, p) = \inf_{\delta>0} \chi^*(p; \overline{\text{co}}\, Q(x; \delta)).$$

For every closed convex subset C of V we have $\chi^{**}(\cdot; C) = \chi(\cdot; C)$ by [7, I.4]. Thus, by Theorem 2.4 the seminormal hull $\bar{\bar{\chi}}_Q$ of χ_Q is given by

$$\bar{\bar{\chi}}_Q(x, v) = \sup_{\delta>0} \chi(v; \overline{\text{co}}\, Q(x; \delta)) = \chi\Big(v; \bigcap_{\delta>0} \overline{\text{co}}\, Q(x; \delta)\Big).$$

Hence, the desired equivalence follows directly from Proposition 2.2 by obvious properties of indicator functions. □

We shall now work towards a characterization of seminormality in terms of property (Q). For this purpose, let $Q_1: X \rightrightarrows V \times \mathbb{R}$ be a given multifunction. Of course, the definition of property (Q) extends to Q_1 by an obvious substitution of framework. Following [1b], we define the *modified Lagrangian* $\zeta_{Q_1}: X \times V \times \mathbb{R} \to \bar{\mathbb{R}}$ of Q_1 by

$$\zeta_{Q_1}(x, v, \lambda) \equiv \inf\{\eta: \eta \geqq \lambda, (v, \eta) \in Q_1(x)\}.$$

It is useful to express ζ_{Q_1} differently. For every subset C of $V \times \mathbb{R}$ we define the function $\zeta(\cdot, \cdot; C): V \times \mathbb{R} \to \bar{\mathbb{R}}$ by

$$\zeta(v, \lambda; C) \equiv \inf_{\eta \geqq \lambda} [\eta + \chi(v, \eta; C)]; \tag{2.20}$$

then clearly we have $\zeta_{Q_1}(x, v, \lambda) = \zeta(v, \lambda; Q_1(x))$. For every subset C of $V \times \mathbb{R}$ we shall use the following terminology: the *section* of C at $v \in V$ is the set, denoted by C_v, which consists of all $\lambda \in \mathbb{R}$ such that $(v, \lambda) \in C$; we shall say that C_v is *closed from the right* if for every nonincreasing sequence $\{\lambda_k\}_1^\infty \subset C_v$ with $\lambda_0 \equiv \lim_{k\to\infty} \downarrow \lambda_k \in \mathbb{R}$, we have $\lambda_0 \in C_v$.

LEMMA 2.16. *If $C, C' \subset V \times \mathbb{R}$ are such that*

$$\zeta(\cdot, \cdot; C) = \zeta(\cdot, \cdot; C'),$$

$$C_v, C'_v \text{ are closed from the right for every } v \in V,$$

then $C = C'$.

Proof. The result follows directly from the fact that if C_v is closed from the right, the infimum in (2.20) is attained, provided that it is finite. □

Remark 2.17. Closedness from the right of the v-sections is an essential condition for the above lemma. Consider for instance the case where $C \equiv V \times \mathbb{Q}$, $C' \equiv V \times (\mathbb{R} \setminus \mathbb{Q})$, with $\mathbb{Q}$ denoting the set of rational numbers.

PROPOSITION 2.18. *For every $x \in X$ we have*

$$\zeta_{Q_1} \text{ is seminormal at every point of } \{x\} \times V \times \mathbb{R}, \tag{2.21}$$

$$(Q_1(x))_v \text{ is closed from the right for every } v \in V \tag{2.22}$$

if and only if

$$Q_1 \text{ has property } (Q) \text{ at } x. \tag{2.23}$$

Proof. An easy computation shows

$$\zeta^*(p, q; C) = \begin{cases} \chi^*(p, q-1; C) & \text{if } q \geqq 0, \\ +\infty & \text{otherwise.} \end{cases}$$

Since we have

$$b_{Q_1}(y, p, q) \equiv \zeta^*_{Q_1}(y, p, q) = \zeta^*(p, q; Q_1(y)),$$

this gives

$$\bar{b}_{Q_1}(x, p, q) \equiv \limsup_{y \to x} b_{Q_1}(y, p, q) = \inf_{\delta > 0} \zeta^*(p, q; \overline{\text{co}}\, Q_1(x; \delta)),$$

where $Q_1(x; \delta)$ stands for the union of all sets $Q_1(y)$ with $d(y, x) < \delta$. By [7, I.4] we have for every closed convex subset C of $V \times \mathbb{R}$ that $\zeta(\cdot, \cdot; C) = \zeta^{**}(\cdot, \cdot; C)$. Therefore, the seminormal hull $\tilde{\zeta}_{Q_1}$ of ζ_{Q_1} is given by

$$\tilde{\zeta}_{Q_1}(x, v, \lambda) = \sup_{\delta > 0} \zeta(v, \lambda; \overline{\text{co}}\, Q_1(x; \delta)), \tag{2.24}$$

as follows from applying Theorem 2.4. We can rewrite this as

$$\tilde{\tilde{\zeta}}_{Q_1}(x, v, \lambda) = \lim_{n\to\infty} \uparrow \inf_{\eta \geqq \lambda} [\eta + \chi(v, \eta; \overline{\text{co}}\, Q_1(x; n^{-1}))]. \tag{2.25}$$

If $\tilde{\tilde{\zeta}}_{Q_1}(x, v, \lambda) = +\infty$, then it follows trivially from (2.24) that

$$\tilde{\tilde{\zeta}}_{Q_1}(x, v, \lambda) = \zeta\Big(v, \lambda; \bigcap_{\delta>0} \overline{\text{co}}\, Q_1(x; \delta)\Big).$$

This identity remains valid if $\tilde{\tilde{\zeta}}_{Q_1}(x, v, \lambda) < +\infty$, since it then follows from (2.25) by an obvious application of Lemma 2.11 (or Dini's theorem). Taking into account the fact that (2.23) trivially implies (2.22), we conclude from Lemma 2.16 and Proposition 2.2 that (2.21)-(2.22) are equivalent to (2.23). □

Our next result states that seminormality for a and seminormality for a_1 are very closely related. Here we use the *effective domain multifunction* $D_a: X \rightrightarrows V$, defined by

$$D_a(x) \equiv \text{dom}\, a(x, \cdot) \equiv \{v \in V: a(x, v) < +\infty\}.$$

PROPOSITION 2.19. *The seminormal hull* $\tilde{\tilde{a}}_1$ *of* a_1 *is given by*

$$\tilde{\tilde{a}}_1(x, v, \lambda) = \max(\tilde{\tilde{a}}(x, v), \lambda + \tilde{\tilde{\chi}}_{D_a}(x, v)).$$

Proof. We define $b_1, \bar{b}_1: X \times P \times \mathbb{R} \to \bar{\mathbb{R}}$ by

$$b_1(y, p, q) \equiv a_1^*(y, p, q) \equiv \sup_{v\in V, \lambda\in\mathbb{R}} [\langle\langle v, \lambda; p, q\rangle\rangle - a_1(y, v, \lambda)],$$

$$\bar{b}_1(x, p, q) \equiv \limsup_{y\to x} b_1(y, p, q).$$

An easy computation shows

$$\bar{b}_1(x, p, q) = \begin{cases} (1-q)\bar{b}(x, p(1-q)^{-1}) & \text{if } 0 \leqq q < 1, \\ \bar{\chi}^*_{D_a}(x, p) & \text{if } q = 1, \\ +\infty & \text{if } q < 0 \text{ or } q > 1. \end{cases}$$

This yields directly

$$\bar{b}_1^*(x, v, \lambda) = \max(\bar{b}^*(x, v), \lambda + \tilde{\tilde{\chi}}_{D_a}(x, v)).$$

Hence, the result follows by applying Theorem 2.4. □

As a certain converse to Proposition 2.15 we can now demonstrate that seminormality of the function $a: X \times V \to \bar{\mathbb{R}}$ can be characterized in terms of property (Q) for the *epigraphic multifunction* $Q_a: X \rightrightarrows V \times \mathbb{R}$ of a, defined by

$$Q_a(x) \equiv \text{epi}\, a(x, \cdot) \equiv \{(v, \eta) \in V \times \mathbb{R}: \eta \geqq a(x, v)\}.$$

PROPOSITION 2.20. *For every* $x \in X$ *we have that if*

$$a \text{ is seminormal at every point of } \{x\} \times V, \tag{2.26}$$

then

$$Q_a \text{ has property } (Q) \text{ at } x. \tag{2.27}$$

Moreover, if $a(x, v) < +\infty$ *for all* $v \in V$, *then* (2.26) *and* (2.27) *are equivalent.*

Proof. The modified Lagrangian ζ_{Q_a} of Q_a is easily seen to be precisely the function a_1. Since Q_a obviously satisfies (2.22), the result follows immediately from combining Propositions 2.2, 2.18 and 2.19. □

We shall now briefly discuss the connections between the results of this section and work by others. Surprisingly enough, the adaptation of the seminormality concept of Tonelli and McShane to extended real-valued functions seems not to have been considered before. Hence, our results can only be compared to results which have been formulated in terms of property (Q) for multifunctions. Because of Propositions 2.2, 2.15, Theorem 2.4 generalizes [15, Thm. 3.1] (see also [8d, 17.6]). Also by Proposition 2.15 we can observe that Theorem 2.8 is well-known, albeit in slightly weaker forms [8d, 10.5i], [13, VIII.2.1], [26, Thm. 2.1], [3, Lemma 0.7]. It should be observed that Theorem 2.8 can also be extended as in [8d, 10.5ii] by the introduction of monotonicity in certain components of the variable v (the proof then follows again by Lemma 2.11 and Remark 2.5—this goes similar to the more subtle proof of Theorem 2.13). We also wish to point out that the result of [26, Thm. 2.1], which at first sight appears to belong to a class of more general results, is indeed covered by Theorem 2.8. This can be seen by invoking Proposition 2.20 and constructing a suitable minorant function h under the conditions given in [26]; the details are left to the reader.

Propositions 2.12, 2.19 and especially Theorem 2.13 all seem to be new. Proposition 2.15 is also a novel result; observe that it crucially depends on defining seminormality for extended real-valued functions. Finally, Propositions 2.18 and 2.20 are related to similar, more involved characterizations of seminormality in the sense of Tonelli and McShane, which were given in [8b], [8d, 17.3].

3. Seminormality in the small and in the large. In this section we shall establish the connections between seminormality in the small and in the large, announced in the introduction. We shall repeatedly use results involving outer integration and measurable selections; we refer to Appendices A and B for some basic definitions in this regard.

Let $(T, \mathcal{T}, \mu)$ be a σ-finite measure space. For the moment we shall also assume that the σ-algebra $\mathcal{T}$ is μ-complete; later, we will show how this assumption can be lifted. Let (X, d) be a metric Suslin space (Appendix B) and let $(V, P, \langle \cdot, \cdot \rangle)$ be a pair of locally convex Suslin spaces, paired by the strict duality $\langle \cdot, \cdot \rangle$ (i.e., V and P are Suslin for topologies compatible with $\langle \cdot, \cdot \rangle$).

Let $\mathcal{X}$ be a decomposable set of equivalence classes of $(\mathcal{T}, \mathcal{B}(X))$-measurable functions from T into X (the equivalence relation being equality μ-a.e.), which we equip with the essential supremum metric

$$d(x, y) \equiv \operatorname{ess\,sup}_{t \in T} d(x(t), y(t)).$$

Also, let $(\mathcal{V}, \mathcal{P}, \langle \cdot, \cdot \rangle)$ be a pair of decomposable vector spaces of equivalence classes of scalarly μ-integrable functions (Appendix B) going from T into V and P respectively, paired by the duality

$$\langle v, p \rangle \equiv \int_T \langle v(t), p(t) \rangle \mu(dt),$$

where it is assumed that for every $v \in \mathcal{V}$, $p \in \mathcal{P}$ the function $t \mapsto \langle v(t), p(t) \rangle$ belongs to $\mathcal{L}^1_{\mathbb{R}}$ (this function is certainly $\mathcal{T}$-measurable by [7, III.36]). As a consequence of the decomposability hypothesis, the duality $\langle \cdot, \cdot \rangle$ is strict [7, VII.5].

Let $l: T \times X \times V \to \bar{\mathbb{R}}$ be a given function. By means of outer integration (Appendix A) we define the integral functional $I_l: \mathcal{X} \times \mathcal{V} \to \bar{\mathbb{R}}$ as follows:

$$I_l(x, v) \equiv \tilde{\int_T} l(t, x(t), v(t)) \mu(dt).$$

We can now consider seminormality of the integrand l with respect to the framework consisting of (X, d) and $(V, P, \langle \cdot, \cdot \rangle)$ (*seminormality in the small*) and seminormality of the integral functional I_l with respect to the framework which consists of $(\mathscr{X}, d)$ and $(\mathscr{V}, \mathscr{P}, \langle \cdot, \cdot \rangle)$ (*seminormality in the large*). Our main result on the relationship between seminormality in the small and seminormality in the large can now be stated.

THEOREM 3.1. *For every $x \in \mathscr{X}$ such that there exist $p_0 \in \mathscr{P}$, $\varphi_0 \in \mathscr{L}^1_{\mathbb{R}}$ and $\delta > 0$ with*

$$l(t, y, v) \geqq \langle v, p_0(t) \rangle + \varphi_0(t) \text{ for all } y \in B(x(t); \delta) \text{ and } v \in V \quad \mu\text{-a.e.}, \tag{3.1}$$

we have that if

$$l(t, \cdot, \cdot) \text{ is seminormal at every point of } \{x(t)\} \times V \quad \mu\text{-a.e.}, \tag{3.2}$$

then

$$I_l \text{ is seminormal at every point of } \{x\} \times \mathscr{V}. \tag{3.3}$$

Moreover, if

$$l \text{ is } \mathscr{T} \times \mathscr{B}(X \times V)\text{-measurable}, \tag{3.4}$$

$$I_l(x, \cdot) \text{ is not identically equal to } +\infty \text{ on } \mathscr{V}, \tag{3.5}$$

then (3.2) *and* (3.3) *are equivalent.*

We shall avail ourselves of some lemmas to prove this main result.

LEMMA 3.2. *If* (3.4) *holds, then for every $n \in \mathbb{N}$ and $x \in \mathscr{X}$*

(a) *The function $l_{n,x}: T \times X \times V \to \bar{\mathbb{R}}$, defined by*

$$l_{n,x}(t, y, w) \equiv \begin{cases} l(t, y, w) & \text{if } d(y, x(t)) \leqq n^{-1}, \\ +\infty, \end{cases}$$

is $\mathscr{T} \times \mathscr{B}(X \times V)$-measurable.

(b) *The function $m_{n,x}: T \times P \to \bar{\mathbb{R}}$, defined by*

$$m_{n,x}(t, p) \equiv \sup_{y \in X, w \in V} [\langle w, p \rangle - l_{n,x}(t, y, w)],$$

is $\mathscr{T} \times \mathscr{B}(\mathscr{P})$-measurable.

(c) *The function $m^*_{n,x}: T \times V \to \bar{\mathbb{R}}$, defined by*

$$m^*_{n,x}(t, v) \equiv \sup_{p \in P} [\langle v, p \rangle - m_{n,x}(t, p)],$$

is $\mathscr{T} \times \mathscr{B}(V)$-measurable.

Proof. (a) A simple consequence of [7, III.36]. (b) (Compare with the proof of [7, VII.1].) By the von Neumann-Aumann theorem [7, III.22] the epigraphic multifunction $t \mapsto \operatorname{epi} l_{n,x}(t, \cdot, \cdot)$, whose graph is $\mathscr{T} \times \mathscr{B}(X \times V \times \mathbb{R})$-measurable by (a), has a sequence $\{(y_j, w_j, r_j)\}_1^\infty$ of measurable selections such that for every $t \in T$ the sequence $\{(y_j(t), w_j(t), r_j(t))\}$ is dense in $\operatorname{epi} l_{n,x}(t, \cdot, \cdot)$. By continuity of $\langle \cdot, p \rangle$ this gives

$$m_{n,x}(t, p) = \sup_{j \in \mathbb{N}} [\langle w_j(t), p \rangle - r_j(t)],$$

so measurability of $m_{n,x}$ is evident from [7, III.36]. (c) The proof here is entirely similar to that of (b). □

Let us introduce the function $l: T \times X \times V \to \bar{\mathbb{R}}$ by setting

$$\tilde{\tilde{l}}(t, \cdot, \cdot) \equiv \text{seminormal hull of } l(t, \cdot, \cdot).$$

LEMMA 3.3. *If* (3.4) *holds, then the function $\tilde{\tilde{l}}$ is $\mathscr{T} \times \mathscr{B}(X \times V)$-measurable.*

Proof. By Remark 2.3 we know

$$\tilde{\tilde{l}}(t, x, v) = \sup_{n\in\mathbb{N}, y\in X, p\in P} [\langle v, p\rangle - c(t, n, y, p) - \chi(x; B(y; n^{-1}))]$$

with $c: T\times\mathbb{N}\times X\times P\to\bar{\mathbb{R}}$ defined by

$$c(t, n, y, p) \equiv \sup_{z\in X, w\in V} [\langle w, p\rangle - l(t, z, w) - \chi(z; B(y; n^{-1}))].$$

Let us first show that $c(\cdot, n, \cdot, \cdot)$ is $\mathcal{T}\times\mathcal{B}(X\times P)$-measurable for arbitrary $n\in\mathbb{N}$. By the von Neumann-Aumann theorem [7, III.22] the epigraphic multifunction $t\mapsto$ epi $l(t, \cdot, \cdot)$ has a sequence $\{(z_j, w_j, r_j)\}$ of $(\mathcal{T}, \mathcal{B}(X\times V\times\mathbb{R}))$-measurable selections such that for every $t\in T$ the sequence $\{(z_j(t), w_j(t), r_j(t))\}$ is dense in epi $l(t, \cdot, \cdot)$. By continuity of $\langle\cdot, p\rangle$ and upper semicontinuity of the indicator function of an open set this gives

$$c(t, n, y, p) = \sup_{j\in\mathbb{N}} [\langle w_j(t), p\rangle - r_j(t) - \chi(z_j(t); B(y; n^{-1}))].$$

The desired measurability of $c(\cdot, n, \cdot, \cdot)$ is now obvious by [7, III.36]. The measurability proof for $\tilde{\tilde{l}}$ now proceeds by an obvious repetition of moves, which is left to the reader. □

Note that the crucial difference between Lemmas 3.2 and 3.3 is that Lemma 3.2 is concerned with closed balls in the space X, and Lemma 3.3 with open balls. The reasons for this distinction will be apparent in the proof of the following key result.

LEMMA 3.4. *If* (3.1), (3.4) *hold, then*

$$I_{\tilde{\tilde{l}}} \text{ is the seminormal hull of } I_l: \mathcal{X}\times\mathcal{V}\to\bar{\mathbb{R}}.$$

Proof. Let $x\in\mathcal{X}$ and $v\in V$ be arbitrary. By Remark 2.5 we have

$$\tilde{\tilde{I}}_l(x, v) = \lim_{n\to\infty} \uparrow \sup_{p\in\mathcal{P}} \alpha_{n,p},$$

where

$$\alpha_{n,p} \equiv \inf_{y\in\mathcal{X}, d(x,y)\leqq 1/n, w\in\mathcal{V}} [\langle v - w, p\rangle + I_l(y, w)].$$

By definition of the essential supremum metric d we have

$$\alpha_{n,p} = \inf_{y\in\mathcal{X}, w\in\mathcal{V}} [\langle v - w, p\rangle + I_{l_{n,x}}(y, w)],$$

with $l_{n,x}$ as defined in Lemma 3.2(a). If $\alpha_{n,p} > -\infty$, it follows from Lemma 3.2(b) and the decomposability of $\mathcal{X}$ and $\mathcal{V}$ by the reduction theorem (Theorem B.1) that

$$\alpha_{n,p} = \langle v, p\rangle - I_{m_{n,x}}(p).$$

Note that by (3.1) for n large enough

$$m_{n,x}(t, p_0(t)) \leqq -\varphi_0(t);$$

hence, it follows from the above, Lemma 3.2(c) and decomposability of $\mathcal{P}$ that

$$\sup_{p\in\mathcal{P}} \alpha_{n,p} = \sup_{p\in\mathcal{P}} [\langle v, p\rangle - I_{m_{n,x}}(p)] = I_{m^*_{n,x}}(v)$$

by yet another application of the reduction theorem. Note that by the above inequality

$$m^*_{n,x}(t, v(t)) \geqq \langle v(t), p_0(t)\rangle + \varphi_0(t)$$

for n large enough. Hence, it follows by the monotone convergence theorem that

$$\lim_{n\to\infty} \uparrow \sup_{p\in\mathscr{P}} \alpha_{n,p} = \int_T \lim_{n\to\infty} \uparrow m^*_{n,x}(t, v(t))\mu(dt).$$

By combination of the definitions leading up to that of $m^*_{n,x}$ in Lemma 3.2 and Remark 2.5 it is easily seen that

$$\lim_{n\to\infty} \uparrow m^*_{n,x}(t, w) = \tilde{\tilde{I}}(t, x(t), w).$$

Hence, we may conclude that $\tilde{\tilde{I}}_I = I^*_I$. □

At the present stage we can prove Theorem 3.1 under the additional hypothesis that (3.4) be valid throughout:

Proof of Theorem 3.1 (simplified version). Here we suppose that (3.4) is valid throughout. By Proposition 2.2, (3.2) is equivalent to

$$\tilde{\tilde{I}}(t, x(t), \cdot) = I(t, x(t), \cdot) \ \mu\text{-a.e.} \tag{3.6}$$

Thus, if (3.2) holds, Lemma 3.4 gives

$$\tilde{\tilde{I}}_I(x, \cdot) = I^*_I(x, \cdot) = I_I(x, \cdot).$$

Hence (3.3) follows by invoking Proposition 2.2. Conversely, if (3.3) holds, then Proposition 2.2 and Lemma 3.4 give

$$I_I(x, \cdot) = \tilde{\tilde{I}}_I(x, \cdot) = I^*_I(x, \cdot).$$

Hence, (3.6) follows by Theorem B.2 from (3.5), Lemma 3.3 and the decomposability of $\mathscr{V}$. □

We shall now introduce an additional lemma that will enable us to remove condition (3.4) in proving one of the implications in Theorem 3.1.

LEMMA 3.5. *There exists a* $\mathscr{T}\times\mathscr{B}(X\times V)$*-measurable function* $\hat{I}: T\times X\times V\to\bar{\mathbb{R}}$ *such that*

$$\hat{I}(t, \cdot, \cdot) \text{ is seminormal on } X\times V \ \mu\text{-a.e.}, \tag{3.7}$$

$$\hat{I}(t, \cdot, \cdot) \geqq \tilde{\tilde{I}}(t, \cdot, \cdot) \ \mu\text{-a.e.}, \tag{3.8}$$

$$I_{\hat{I}} = I^*_I. \tag{3.9}$$

Proof. Consider the set $\mathscr{G}$ of all normal integrands g on $T\times(\mathbb{N}\times X\times P)$ of the type

$$g(t, n, y, p) \equiv \langle v(t), p\rangle - \varphi(t) - \chi(x(t); B(y; n^{-1}))$$

for all $x\in\mathscr{X}$, $v\in\mathscr{V}$ and all $\mathscr{T}$-measurable functions $\varphi: T\to\bar{\mathbb{R}}$ satisfying $\varphi(t)\geqq \tilde{\tilde{I}}(t, x(t), v(t))$ μ-a.e. By Theorem B.3 there exists a countable subset $\mathscr{G}_0$ of $\mathscr{G}$ such that $\hat{g}\equiv\sup_{g\in\mathscr{G}_0} g$ satisfies the inequality $\hat{g}(t, \cdot)\geqq g(t, \cdot)$ μ-a.e. for every $g\in\mathscr{G}$. We define

$$\hat{I}(t, z, w) \equiv \sup_{n\in\mathbb{N}, y\in X, p\in P} [\langle w, p\rangle - \hat{g}(t, n, y, p) - \chi(y; B(z; n^{-1}))].$$

Since $\hat{I}(t, \cdot, \cdot)$ is the supremum of a collection of simple seminormal functions on $X\times V$, (3.7) follows immediately. Also, $\hat{g}$ is by its definition a normal integrand on $T\times(\mathbb{N}\times X\times P)$; hence the desired measurability of $\hat{I}$ is proven in complete analogy to what was done in Lemma 3.3. Also, it follows from the definition of $\hat{g}$ that

$$\hat{g}(t, n, y, p) \leqq \sup_{z\in X, w\in V} [\langle w, p\rangle - \tilde{\tilde{I}}(t, z, w) - \chi(z; B(y; n^{-1}))].$$

Therefore it follows by Remark 2.3 from the definition of $\hat{I}$ that for μ-a.e. $t\in T$ the

seminormal hull of $\tilde{\tilde{l}}(t,\cdot,\cdot)$, which is $\tilde{\tilde{l}}(t,\cdot,\cdot)$ itself, is no larger than $\hat{l}(t,\cdot,\cdot)$. This proves (3.8). Finally, let $x\in\mathscr{X}$ and $v\in\mathscr{V}$ be arbitrary. For every $\mathscr{T}$-measurable function $\varphi: T\to\bar{\mathbb{R}}$ such that $\varphi(t)\geqq\tilde{\tilde{l}}(t,x(t),v(t))$ μ-a.e. we have for μ-a.e. $t\in T$

$$\varphi(t)\geqq\langle v(t),p\rangle-\hat{g}(t,n,y,p)-\chi(y;B(x(t);n^{-1}))\text{ for all }n\in\mathbb{N},y\in X,p\in P,$$

by the essential supremum property of $\hat{g}$; this implies $\varphi(t)\geqq\hat{l}(t,x(t),v(t))$ μ-a.e. By definition of outer integration this means that $I_{\tilde{l}}(x,v)\geqq I_{\hat{l}}(x,v)$; the converse inequality follows directly from (3.8) and measurability of $\hat{l}$. We conclude that (3.9) has been shown to hold. □

Proof of Theorem 3.1 (remainder). Let $\hat{l}$ correspond to $\tilde{\tilde{l}}$ as in Lemma 3.5. Since (3.2) holds, we have (3.6). In view of (3.9) this gives

$$I_l(x,\cdot)=I_{\hat{l}}(x,\cdot).\tag{3.10}$$

As $\hat{l}$ is $\mathscr{T}\times\mathscr{B}(X\times V)$-measurable, substitution of $\hat{l}$ in lieu of l in Lemma 3.4 gives $\tilde{\tilde{I}}_{\hat{l}}=I_{\hat{l}}$ by using (3.7). This shows $I_{\hat{l}}$ to be seminormal on $\mathscr{X}\times\mathscr{V}$ (Proposition 2.2). By (3.9) and (2.3) we have $\tilde{\tilde{I}}_l\geqq I_{\hat{l}}$. By Proposition 2.2, (3.3) thus follows from (3.10). □

From the proof of Theorem 3.1 it is evident that the following result also holds; its proof will be left to the reader.

THEOREM 3.6. *For l such that there exist $p_0\in\mathscr{P}$ and $\varphi_0\in\mathscr{L}^1_{\mathbb{R}}$ with*

$$l(t,y,v)\geqq\langle v,p_0(t)\rangle+\varphi_0(t)\tag{3.11}$$

we have that if

$$l(t,\cdot,\cdot)\text{ is seminormal on }X\times V\ \mu\text{-a.e.},\tag{3.12}$$

then

$$I_l\text{ is seminormal on }\mathscr{X}\times\mathscr{V}.\tag{3.13}$$

Moreover, if (3.4) *holds and*

$$I_l\text{ is not identically equal to }+\infty\text{ on }\mathscr{X}\times\mathscr{V},$$

then (3.12) *and* (3.13) *are equivalent.*

We can rid Theorems 3.1 and 3.6 of the completeness assumption for the measure space $(T,\mathscr{T},\mu)$, hitherto in force.

Remark 3.7. Theorems 3.1 and 3.6 continue to hold if $(T,\mathscr{T},\mu)$ is not assumed to be complete. This follows from the fact that every equivalence class of $\mathscr{T}_\mu$-measurable functions in $\mathscr{X}$, $\mathscr{V}$ or $\mathscr{P}$ has a $\mathscr{T}$-measurable representant by [7, III.36] and elementary facts concerning completion. Here $\mathscr{T}_\mu$ stands for the μ-completion of the σ-algebra $\mathscr{T}$. Further details are left to the reader. □

4. Seminormality and semicontinuity of integral functionals. In this section we shall combine the results of §§ 2, 3 so as to obtain very general sufficient conditions for coincident seminormality, lower closure and lower semicontinuity of integral functionals (of course the point of these conditions is that they do not explicitly refer to seminormality). At the same time, a new approach is initiated to necessary conditions for the lower semicontinuity of integral functionals.

Let $(T,\mathscr{T},\mu)$ be a finite measure space; in contrast to § 3, this measure space is supposed to be finite, since we shall work with weak convergence in $\mathscr{L}^1$-spaces. Otherwise, the framework of this section is that of § 3: (X,d) is a metric Suslin space, $(V,P,\langle\cdot,\cdot\rangle)$ is a pair of locally convex spaces, paired by a strict duality $\langle\cdot,\cdot\rangle$, such

that V and P are Suslin spaces for topologies that are compatible with the duality. Further, $\mathscr{X}$ is a decomposable set of equivalence classes of $(\mathscr{T}, \mathscr{B}(X))$-measurable functions from T into X, and $(\mathscr{V}, \mathscr{P}, \langle \cdot, \cdot \rangle)$ is a pair of decomposable vector spaces of equivalence classes of scalarly $\mathscr{T}$-measurable functions, going from T into V and P respectively, equipped with the strict duality

$$\langle v, p \rangle \equiv \int_T \langle v(t), p(t) \rangle \mu(dt).$$

We recall from § 2 that a function $h: V \to (-\infty, +\infty]$ is said to be of *Nagumo type* if h is convex and sequentially inf-compact on V for every slope. A subset $\mathscr{V}_0$ of $\mathscr{V}$ is defined to be *Nagumo tight* if there exists a $\mathscr{T} \times \mathscr{B}(V)$-measurable function $h: T \times V \to [0, +\infty]$ such that

(4.1) $h(t, \cdot)$ is of Nagumo type on V μ-a.e.,

(4.2) $$\sup_{v \in \mathscr{V}_0} I_h(v) < +\infty.$$

Further, $\mathscr{V}_0$ is defined to be *almost Nagumo tight* if there exists a nonincreasing sequence $\{B_i\}_1^\infty$ in $\mathscr{T}$, whose intersection is a μ-null set, such that for every $i \in \mathbb{N}$ there exists a $\mathscr{T} \times \mathscr{B}(V)$-measurable function $h_i: T \times V \to [0, +\infty]$ with

$(4.1)_i$ $h_i(t, \cdot)$ is of Nagumo type on V μ-a.e.,

$(4.2)_i$ $$\sup_{v \in \mathscr{V}_0} \int_{T \setminus B_i} h_i(t, v(t)) \mu(dt) < +\infty.$$

The following examples show that (almost) Nagumo tightness occurs in some interesting cases.

Example 4.1. Every subset $\mathscr{V}_0$ of $\mathscr{V}$ for which there exists a scalarly $\mathscr{T}$-measurable multifunction $\Gamma: T \rightrightarrows V$ having (sequentially) compact convex values, such that for every $\varepsilon > 0$ there exists a set A_ε in $\mathscr{T}$, $\mu(A_\varepsilon) \leqq \varepsilon$, with

$$v(t) \in \Gamma(t) \quad \text{for all } v \in \mathscr{V}_0 \quad \text{for every } t \in T \setminus A_\varepsilon,$$

is almost Nagumo tight, as is seen by taking B_i to be the intersection of the sets $A_{1/j}$, $j \leqq i$, and h_i to be the indicator function χ_Γ of Γ; note that χ_Γ is $\mathscr{T} \times \mathscr{B}(V)$-measurable by [7, III.37].

Example 4.2. Suppose that $(V, P, \langle \cdot, \cdot \rangle)$ is specified as follows: V is a separable reflexive Banach space, whose norm we denote by $\| \cdot \|$, P is the topological dual V' of $(V, \| \cdot \|)$, $\langle \cdot, \cdot \rangle$ is the usual duality, V is equipped with the topology $\sigma(V, V')$ and P with the topology $\sigma(V', V)$ (in this case V and P are both locally convex Suslin [7, p. 202]). Suppose further that $(\mathscr{V}, \mathscr{P}, \langle \cdot, \cdot \rangle)$ is as follows: $\mathscr{V}$ is the usual L^1-space of $\mathscr{L}^1_V$ of equivalence classes of μ-integrable functions from T into V (note that strong and scalar μ-integrability coincide here by separability of $(V, \| \cdot \|)$), and $\mathscr{P}$ is the usual L^∞-space $\mathscr{L}^\infty_{V'}$ of equivalence classes of essentially bounded scalarly $\mathscr{T}$-measurable functions from T into V'.

Then every subset $\mathscr{V}_0$ of $\mathscr{L}^1_V$ which is relatively compact or relatively sequentially compact for the weak topology $\sigma(\mathscr{L}^1_V, \mathscr{L}^\infty_{V'})$, is Nagumo tight. Namely, by the Dunford–Pettis theorem [12, IV.2.1] the set of all functions $t \mapsto \|v(t)\|$, $v \in \mathscr{V}_0$, is uniformly integrable. Hence, there exists by the de la Vallée Poussin theorem [11, II.22] a nondecreasing continuous convex function $h': [0, +\infty) \to [0, +\infty)$ satisfying (2.8), such

that

$$\sup_{v\in\mathcal{V}_0}\int_T h'(\|v(t)\|)\mu(dt)<+\infty.$$

Setting $h(v)\equiv h'(\|v\|)$, we see that Nagumo tightness holds by Example 2.7.

Example 4.3. Suppose that $(V, P, \langle\cdot,\cdot\rangle)$ and $(\mathcal{V}, \mathcal{P}, \langle\cdot,\cdot\rangle)$ are as in Example 4.2. Then for every sequence $\{v_k\}_1^\infty$ in $\mathcal{V}$ such that

$$\sup_{k\in\mathbb{N}}\int_T \|v_k(t)\|\mu(dt)<+\infty,$$

there exists a subsequence $\{k'\}$ of $\{k\}$ such that $\{v_{k'}\}$ is almost Nagumo tight. Namely, the set of all functions $t\mapsto\|v_k(t)\|$, $k\in\mathbb{N}$, is uniformly bounded in $\mathcal{L}^1_{\mathbb{R}}$ in L^1-norm. Hence by the biting lemma of Chacon [4] there exist a nonincreasing sequence $\{B_i\}$ in $\mathcal{T}$, having a μ-null set as its intersection, and a subsequence $\{k'\}$ of $\{k\}$ such that for every $i\in\mathbb{N}$ the restrictions of the functions $t\mapsto\|v_{k'}(t)\|$ to $T\backslash B_i$ form a uniformly μ-integrable sequence. At this point the previous example can be taken up in an obvious way.

Let $l: T\times X\times V\to\bar{\mathbb{R}}$ be a given function; correspondingly we define the function $l_1: T\times X\times V\times\mathbb{R}\to\bar{\mathbb{R}}$ by

$$l_1(t, x, v, \lambda)\equiv\max(l(t, x, v), \lambda).$$

The integral functional $I_{l_1}: \mathcal{X}\times\mathcal{V}\times\mathcal{L}^1_{\mathbb{R}}\to\bar{\mathbb{R}}$, of course defined by

$$I_{l_1}(x, v, \lambda)\equiv\int_T l_1(t, x(t), v(t), \lambda(t))\mu(dt),$$

will be considered in connection with the seminormality framework which consists of $(\mathcal{X}, d)$ and the pair $(\mathcal{V}\times\mathcal{L}^1_{\mathbb{R}}, \mathcal{P}\times\mathcal{L}^\infty_{\mathbb{R}}, \langle\!\langle\cdot;\cdot\rangle\!\rangle)$ of decomposable vector spaces, whose duality (strict) is given by

$$\langle\!\langle v, \lambda; p, q\rangle\!\rangle\equiv\int_T(\langle v(t), p(t)\rangle+\lambda(t)q(t))\mu(dt).$$

THEOREM 4.4. *For every $x\in\mathcal{X}$ if*

(4.3) *$l(t,\cdot,\cdot)$ is sequentially lower semicontinuous at every point of $\{x(t)\}\times V$ μ-a.e.,*

(4.4) *$l(t, x(t),\cdot)$ is convex on V μ-a.e.,*

then for every $\mathcal{V}_0\subset\mathcal{V}$ and $\mathcal{L}_0\subset\mathcal{L}^1_{\mathbb{R}}$ satisfying

(4.5) *$\mathcal{V}_0$ is almost Nagumo tight,*

(4.6) *$\mathcal{L}_0^-\equiv\{\lambda^-: \lambda\in\mathcal{L}_0\}$ is uniformly μ-integrable,*

we have

(4.7) *I_{l_1} coincides on $\mathcal{X}\times\mathcal{V}_0\times\mathcal{L}_0$ with a function which is seminormal at every point of $\{x\}\times\mathcal{V}\times\mathcal{L}^1_{\mathbb{R}}$.*

Proof. To simplify matters, we shall first assume that $\mathcal{V}_0$ is Nagumo tight. Let h be as in (4.1)-(4.2). By (4.6) there exists a nondecreasing continuous convex function

$h'\colon[0,+\infty)\to[0,+\infty)$, satisfying (2.8), such that

$$\sup_{\lambda\in\mathscr{L}_0}\int_T h'(\lambda^-(t))\mu(dt)<+\infty, \tag{4.8}$$

as follows by the theorem of de la Vallée Poussin. For every $\varepsilon>0$ we define the function $l_{1,\varepsilon}\colon T\times X\times V\times\mathbb{R}\to\bar{\mathbb{R}}$ by

$$l_{1,\varepsilon}(t,y,v,\lambda)\equiv l_1(t,y,v,\lambda)+\varepsilon h(v)+\varepsilon h'(\lambda^-).$$

In view of (4.3)–(4.4) and the properties of h and h' we have by Theorem 2.13 that

$$l_{1,\varepsilon}(t,\cdot,\cdot,\cdot)\text{ is seminormal at every point of }\{x(t)\}\times V\times\mathbb{R}\ \mu\text{-a.e.}$$

Hence condition (3.2) of Theorem 3.1 holds. Also the boundedness condition (3.1)—in fact (3.11)—is valid, since we can take $p_0(t)\equiv 0$, $q_0(t)\equiv 1$ and $\varphi_0(t)\equiv 0$ for all $t\in T$. By Theorem 3.1 we obtain

$$I_{l_{1,\varepsilon}}\text{ is seminormal at every point of }\{x\}\times\mathscr{V}\times\mathscr{L}^1_{\mathbb{R}}.$$

Let σ stand for the value of the supremum in (4.2) and σ' for the value of the supremum in (4.8). The function $J\colon\mathscr{X}\times\mathscr{V}\times\mathscr{L}^1_{\mathbb{R}}\to\bar{\mathbb{R}}$ defined by

$$J\equiv\sup_{\varepsilon>0}(I_{l_{1,\varepsilon}}-\varepsilon\sigma-\varepsilon\sigma'),$$

is seminormal at every point of $\{x\}\times\mathscr{V}\times\mathscr{L}^1_{\mathbb{R}}$, as follows from the above by Remark 2.1. It is also elementary to verify that

$$J(y,v,\lambda)=I_{l_1}(y,v,\lambda)\quad\text{for every } y\in\mathscr{X},\ v\in\mathscr{V}_0\text{ and }\lambda\in\mathscr{L}_0,$$

so the proof under the simplifying assumption has come to an end. Next, we suppose that merely (4.5) holds. Let $\{B_i\}_1^\infty$ be as asserted in the definition of almost Nagumo tightness. For every $i\in\mathbb{N}$ we define $I_i\colon\mathscr{X}\times\mathscr{V}\times\mathscr{L}^1_{\mathbb{R}}\to\bar{\mathbb{R}}$ by

$$I_i(y,v,\lambda)\equiv\int_T 1_{T\setminus B_i}(t)l_1(t,y(t),v(t),\lambda(t))\mu(dt)+\int_{B_i}\lambda(t)\mu(dt). \tag{4.9}$$

From the above it follows by making an obvious substitution for the integrand that for every $i\in\mathbb{N}$, I_i coincides on $\mathscr{X}\times\mathscr{V}_0\times\mathscr{L}_0$ with a function $J_i\colon\mathscr{X}\times\mathscr{V}\times\mathscr{L}^1_{\mathbb{R}}\to\bar{\mathbb{R}}$ which is seminormal at every point of $\{x\}\times\mathscr{V}\times\mathscr{L}^1_{\mathbb{R}}$. By Proposition A.1 and the monotone convergence theorem we have $I_{l_1}=\lim_i\uparrow I_i$ on $\mathscr{X}\times\mathscr{V}\times\mathscr{L}^1_{\mathbb{R}}$ (pointwise), so the desired result now follows by Remark 2.1. □

In complete analogy to the derivation of Theorem 4.4 from Theorem 3.1 we can now obtain the following result from Theorem 3.6.

THEOREM 4.5. *If*

$l(t,\cdot,\cdot)$ *is sequentially lower semicontinuous on* $X\times V$ μ*-a.e.,*

$l(t,x,\cdot)$ *is convex on* V *for every* $x\in X$ μ*-a.e.,*

then for every $\mathscr{V}_0\subset\mathscr{V}$ *and* $\mathscr{L}_0\subset\mathscr{L}^1_{\mathbb{R}}$ *satisfying* (4.5)–(4.6) *we have*

I_{l_1} *coincides on* $\mathscr{X}\times\mathscr{V}_0\times\mathscr{L}_0$ *with a seminormal function.*

Remark 4.6. The coincident seminormality results obtained in Theorems 4.4 and 4.5 immediately imply lower semicontinuity results for I_{l_1}. For instance, (4.7) implies

I_{l_1} is lower semicontinuous at every point of $\{x\}\times\mathscr{V}_0\times\mathscr{L}_0$, relative to the set $\mathscr{X}\times\mathscr{V}_0\times\mathscr{L}_0$,

where $\mathscr{X}$ is equipped with the d-topology and $\mathscr{V} \times \mathscr{L}^1_{\mathbb{R}}$ with the weak topology $\sigma(\mathscr{V} \times \mathscr{L}^1_{\mathbb{R}}, \mathscr{P} \times \mathscr{L}^\infty_{\mathbb{R}})$.

Rather than the modes of convergence employed in the above remark, it is desirable to work with weaker modes of convergence. For this purpose we shall work from now on with given sequences $\{x_k\}_0^\infty$ in $\mathscr{X}$ and $\{v_k\}_1^\infty$ in $\mathscr{V}$ (formally we could consider generalized sequences, but with little effect—see our comments below). We shall say that the sequence $\{x_k\}_1^\infty$ *almost* converges in d to the function x_0 if there exists a nonincreasing sequence $\{B_i\}_1^\infty$ in $\mathcal{T}$, whose intersection is a μ-null set, such that for every $i \in \mathbb{N}$

$$\operatorname*{ess\,sup}_{t \in T \setminus B_i} d(x_k(t), x_0(t)) \to 0.$$

We shall also say that the sequence $\{v_k\}_1^\infty$ *almost* converges in $\sigma(\mathscr{V}, \mathscr{P})$ to $v \in \mathscr{V}$ if there exists a nonincreasing sequence $\{B_i\}_1^\infty$ in $\mathcal{T}$, whose intersection is a μ-null set, such that for every $i \in \mathbb{N}$

$$\int_{T \setminus B_i} \langle v_k(t), p(t) \rangle \mu(dt) \to \int_{T \setminus B_i} \langle v(t), p(t) \rangle \mu(dt) \text{ for every } p \in \mathscr{P}.$$

Example 4.7. If the sequence $\{x_k\}_1^\infty$ converges in measure μ to x_0 then there exists a subsequence $\{\ell\}$ of $\{k\}$ such that $\{x_k\}$ almost converges in d to x_0. Namely, $\{k\}$ contains a subsequence $\{\ell\}$ such that $\{x_\ell\}$ converges to x_0 μ-a.e. Therefore there exists by Egorov's theorem [23, II.4] a sequence $\{A_j\}_1^\infty$ in $\mathcal{T}$, $\mu(A_j) \leqq j^{-1}$, such that $\operatorname{ess\,sup}_{T \setminus A_j} d(x_\ell(t), x_0(t)) \to 0$ for every $j \in \mathbb{N}$. Now take B_i to be the intersection of the sets A_j, $j \leqq i$.

Example 4.8. Suppose that $(V, P, \langle \cdot, \cdot \rangle)$, $\langle \mathscr{V}, \mathscr{P}, \langle \cdot, \cdot \rangle \rangle$ and $\{v_k\}_1^\infty$ are as in Example 4.3. Then $\{k\}$ contains a subsequence $\{\ell\}$ such that $\{v_\ell\}$ almost converges in $\sigma(\mathscr{L}^1_V, \mathscr{L}^\infty_{V'})$ to some $v_* \in \mathscr{L}^1_V$. Namely, let $\{k'\}$ and $\{B_i\}^\infty$ be as in Example 4.3. For every $i \in \mathbb{N}$ the restrictions of the functions $t \mapsto \|v_{k'}(t)\|$ to $T \setminus B_i$ form a uniformly μ-integrable subset of $\mathscr{L}^1_{\mathbb{R}}$, so by the Dunford-Pettis theorem [12, IV.2.1] a subsequence of the sequence of restrictions of v_k to $T \setminus B_i$, $k \in \mathbb{N}$, converges in $\sigma(\mathscr{L}^1_V(T \setminus B_i), \mathscr{L}^\infty_{V'}(T \setminus B_i))$ to some element of $\mathscr{L}^1_V(T \setminus B_i)$. By the obvious extraction of a diagonal sequence one thus finds the desired $\{\ell\}$ and $v_* \in \mathscr{L}^1_V$ (μ-integrability of v_* follows from the uniform L^1-boundedness of the initial sequence).

Note that neither example above—nor any other of which we know—applies to generalized sequences. Thus, apart from Theorems 4.4, 4.5 and Remark 4.6, nothing new ensues for generalized sequences.

THEOREM 4.9. *If*

(4.10) $l(t, \cdot, \cdot)$ *is sequentially lower semicontinuous at every point of* $\{x_0(t)\} \times V$ μ-*a.e.*,

(4.11) $l(t, x_0(t), \cdot)$ *is convex on* V μ-*a.e.*,

if also

$\{x_k\}_1^\infty$ *almost converges in* d *to* x_0,

$\{v_k\}_1^\infty$ *almost converges in* $\sigma(\mathscr{V}, \mathscr{P})$ *to* $v \in \mathscr{V}$,

$\{v_k\}_1^\infty \cup \{v\}$ *is almost Nagumo tight,*

and if there exists a uniformly μ-*integrable sequence* $\{\lambda_k\}_1^\infty$ *in* $\mathscr{L}^1_{\mathbb{R}}$ *such that*

(4.12) $l(t, x_k(t), v_k(t)) \geqq \lambda_k(t)$ μ-*a.e. for every* $k \in \mathbb{N}$,

then

$$\liminf_{k\to\infty} I_l(x_k, v_k) \geqq I_l(x_0, v). \tag{4.13}$$

Proof. Suppose first that $\{x_k\}$ converges in d to x_0 and $\{v_k\}_1^\infty$ converges in $\sigma(\mathcal{V}, \mathcal{P})$ to v. Let α stand for the left side in (4.13). There exists a subsequence $\{k'\}$ of $\{k\}$ such that $\lim_{k'} I_l(x_{k'}, v_{k'}) = \alpha$. By the Dunford-Pettis theorem there exist a subsequence $\{k''\}$ of $\{k'\}$ and $\lambda_* \in \mathcal{L}^1_{\mathbb{R}}$ such that $\{\lambda_{k''}\}$ converges to λ_* in $\sigma(\mathcal{L}^1_{\mathbb{R}}, \mathcal{L}^\infty_{\mathbb{R}})$. Hence the sequence $\{(v_{k''}, \lambda_{k''})\}$ converges to (v, λ_*) in the topology $\sigma(\mathcal{V}\times\mathcal{L}^1_{\mathbb{R}}, \mathcal{P}\times\mathcal{L}^\infty_{\mathbb{R}})$. By Remark 4.6 this gives (4.13), in view of the definition of l_1. Next, we prove the result in full generality. Let $\{B_i\}_1^\infty$ be as in the definitions of almost d-convergence and almost $\sigma(\mathcal{V}, \mathcal{P})$-convergence; rather than taking unions pairwise, we may work with one such sequence. Since (4.9) is also valid here (admittedly with a different interpretation of the B_i), the result follows now from the above by imitating the final step in the proof of Theorem 4.4. □

COROLLARY 4.10. *If* (4.10)–(4.12) *hold, if also*

$$\{x_k\}_1^\infty \text{ converges in measure } \mu \text{ to } x_0,$$

$$\{v_k\}_1^\infty \text{ converges in } \sigma(\mathcal{V}, \mathcal{P}) \text{ to } v_0 \in \mathcal{V},$$

and if there exists a scalarly measurable multifunction $\Gamma: T \rightrightarrows V$ *having compact convex values, such that for every* $\varepsilon > 0$ *there exists* A_ε *in* $\mathcal{T}$, $\mu(A_\varepsilon) \leqq \varepsilon$, *with*

$$v_k(t) \in \Gamma(t) \text{ for all } k \in \mathbb{N}\cup\{0\} \text{ for every } t \in T\setminus A_\varepsilon,$$

then

$$\liminf_{k\to\infty} I_l(x_k, v_k) \geqq I_l(x_0, v_0).$$

Proof. Combine Examples 4.1, 4.7 and Theorem 4.9. □

COROLLARY 4.11. *Suppose that* $(V, P, \langle\cdot,\cdot\rangle)$, $(\mathcal{V}, \mathcal{P}, \langle\cdot,\cdot\rangle)$ *are as in Example* 4.2, *that* V *is the Cartesian product of two separable reflexive Banach spaces* U *and* W *and that* P *is the direct sum of their topological duals* U' *and* W'. *Correspondingly, we write* $v_k \equiv (u_k, w_k)$. *If* (4.10), (4.12) *hold and*

$$l(t, x_0(t), \cdot) \text{ is convex on } V \ \mu\text{-a.e.}, \tag{4.11'}$$

if also

$$\{x_k\}_1^\infty \text{ converges in measure } \mu \text{ to } x_0,$$

$$\{u_k\}_1^\infty \text{ converges in } \sigma(\mathcal{L}^1_U, \mathcal{L}^\infty_{U'}) \text{ to } u_0 \in \mathcal{L}^1_U,$$

$$\sup_{k\in\mathbb{N}} \int_T \|w_k(t)\|_W \mu(dt) < +\infty,$$

then there exists a function $w_* \in \mathcal{L}^1_W$ *such that*

$$\liminf_{k\to\infty} I_l(x_k, (u_k, w_k)) \geqq I_l(x_0, (u_0, w_*)).$$

Proof. By Examples 4.2–4.3 it is easy to see that the set $\mathcal{V}_0$ consisting of all (u_k, w_k), $k \in \mathbb{N}$, and (u_0, w_*) is almost Nagumo tight. Combine now Examples 4.7–4.8 with Theorem 4.9. □

It is a simple exercise to merge Corollaries 4.10–4.11 into one lower closure result, using Theorem 4.9; we shall leave this to the interested reader.

Our next subject concerns necessary conditions for the lower semicontinuity of the integral functional I_l. These will be derived in a very simple way from Theorem 3.1, which would seem to open up a new approach to the subject.

THEOREM 4.12. *Suppose that* $(V, P, \langle\cdot,\cdot\rangle)$, $(\mathcal{V}, \mathcal{P}, \langle\cdot,\cdot\rangle)$ *are as in Example* 4.2. *For every* $x \in \mathcal{X}$ *we have that if* (3.1) *and* (3.4)-(3.5) *hold, and if*

(4.14) I_l *is lower semicontinuous at every point of* $\{x\} \times \mathcal{V}$,

(4.15) *the measure* μ *is nonatomic,*

then

(4.16) $l(t,\cdot,\cdot)$ *is sequentially lower semicontinuous at every point of* $\{x(t)\} \times V$ μ*-a.e.,*

(4.17) $l(t, x(t), \cdot)$ *is convex* μ*-a.e.*

Proof. By (3.5) there exists $v_0 \in \mathcal{V}$ such that $I_l(x, v_0) < +\infty$. The singleton $\{v_0\}$ being compact for $\sigma(\mathcal{L}^1_V, \mathcal{L}^\infty_{V'})$, there exists a nondecreasing continuous convex function $h' : [0, +\infty) \to [0, +\infty)$, satisfying (2.8), such that $I_h(v_0) < +\infty$ where we have set $h \equiv h'(\|\cdot\|)$ (Example 4.2). It follows directly from the norm continuity and convexity of h that the integral functional I_h is lower semicontinuous and convex on $\mathcal{L}^1_V$ (for instance, directly by Theorem 3.1). For arbitrary $p \in \mathcal{L}^\infty_{V'}$ the function $\gamma \mapsto h'(\gamma) - \|p\|_\infty \gamma$ satisfies the growth condition (2.8). Hence, it follows by the Dunford Pettis theorem [12, IV.2.1] and the above that

$$I_h \text{ is of Nagumo type on } \mathcal{V}.$$

Also, it follows from (4.14)-(4.15) that

$$I_l(x,\cdot) \text{ is convex on } \mathcal{V}.$$

To see this, note that the epigraph epi $I_l(x,\cdot)$ of $I_l(x,\cdot)$ is closed in $\mathcal{L}^1_V \times \mathbb{R}$ by (4.14). Hence, for epi $I_l(x,\cdot)$ to be convex, it is enough that for every $m \in \mathbb{N}$ and $\{p_1, \cdots, p_m\} \subset \mathcal{L}^\infty_{V'}$ and every $\rho \in \mathbb{R}$ the set consisting of $(\langle v, p_1\rangle, \cdots, \langle v, p_m\rangle, \rho r) \in \mathbb{R}^{m+1}$, for all $(v, r) \in$ epi $I_l(x,\cdot)$, is convex. The latter is evident from (4.15) by Lyapunov's theorem [7]. Further, by (3.1) there exist $p_0 \in \mathcal{P}$, $\beta_0 \in \mathbb{R}$ and $\delta > 0$ such that

$$I_l(y, v) \geqq \langle v, p_0\rangle + \beta_0 \text{ for all } y \in \mathcal{X},\ d(y, x) < \delta, \text{ and all } v \in \mathcal{V},$$

where we take $\beta_0 \equiv \int_T \varphi_0\, d\mu$. Thus all conditions of Corollary 2.9 (cf. Remark 2.14) have been shown to hold. It follows that for every $\varepsilon > 0$

$$I_l + \varepsilon I_h \text{ is seminormal at every point of } \{x\} \times \mathcal{V}.$$

By Theorem 3.1 this implies that for every $\varepsilon > 0$

$$l_\varepsilon(t,\cdot,\cdot) \text{ is seminormal at every point of } \{x(t)\} \times V \ \mu\text{-a.e.}$$

where $l_\varepsilon : T \times X \times V \to \bar{\mathbb{R}}$ is defined by

$$l_\varepsilon(t, y, v) \equiv l(t, y, v) + \varepsilon h(v).$$

From this it follows obviously that (4.16)-(4.17) are valid with l replaced by l_ε; letting ε go to zero then gives (4.17). Also, since $h : V \to (-\infty, +\infty]$ is nondecreasing (4.16) follows directly by using the uniform boundedness principle, letting ε go to zero [16]. □

We shall now briefly discuss the main results of this section. The tightness concept introduced in the beginning is in the spirit of [1e], [1f]; nevertheless the difference

with the tightness concept given there, which does not require superlinear growth of the function h, is considerable. Also the way in which tightness is made to work here is completely different from the approach followed in [1e], [1f]. The coincident seminormality result of Theorem 4.4 is quite new. The abstract lower closure and lower semicontinuity result in Theorem 4.9 is also new (but well-known in less general forms, of course). The lower semicontinuity result of Corollary 4.10 generalizes [1a, Case 2] and [6, Thms. 4, 5, 7, 8, 9]; cf. [1f]. The lower closure and lower semicontinuity result of Corollary 4.11 generalizes a multitude of results, e.g. [8d, 10.6i–ii, 10.7i–ii, 10.8i–ii], [17, Thm. 1], and the sufficiency part of [24b, Thm.] (where the space V is finite dimensional) and [5, Thm. 4.1(ii)] and the sufficiency part of [3, 1.0] (the latter two results on semicontinuity deal with a separable reflexive Banach space V). Note that none of the above results deals with outer integration. More importantly, none of the lower closure results mentioned above—nor any other of which we know—has been formulated for infinite-dimensional V, the one exception being [1f, Thm. 3.1], due to the present author. This result coincides precisely with Corollary 4.11, even though it was derived by a totally different approach. The necessary conditions for sequential lower semicontinuity obtained in Theorem 4.12 are quite general but for the boundedness condition (3.1); this makes the result somewhat incomparable to similar results obtained by others; e.g., cf. [17, Thm. 2], [24b, Thm.].

Appendix A. Let $(T, \mathcal{T}, \mu)$ be a σ-finite measure space. For every $\mathcal{T}$-measurable function $\varphi: T \to \bar{\mathbb{R}}$ we define

$$\int_T \varphi \, d\mu \equiv \int_T \varphi^+ \, d\mu - \int_T \varphi^- \, d\mu, \tag{A.1}$$

where $\varphi^+ \equiv \max(\varphi, 0)$, $\varphi^- \equiv \max(-\varphi, 0)$. Note that we use the convention $(+\infty) - (+\infty) \equiv +\infty$. Next, we define for every function $\psi: T \to \bar{\mathbb{R}}$

$$\tilde{\int}_T \psi \, d\mu \equiv \inf\left\{\int_T \varphi \, d\mu : \varphi: T \to \bar{\mathbb{R}} \text{ is } \mathcal{T}\text{-measurable}, \ \varphi \geqq \psi\right\}. \tag{A.2}$$

This notion of *outer integration* can be seen as an extension of the outer measure of μ. Note that outer integration in the sense of (A.2) reduces to integration in the sense of (A.1) for $\mathcal{T}$-measurable functions.

PROPOSITION A.1. *For every function $\psi: T \to \bar{\mathbb{R}}$ there exists a $\mathcal{T}$-measurable function $\hat{\varphi}: T \to \bar{\mathbb{R}}$, $\varphi \geqq \psi$, such that*

$$\tilde{\int}_T 1_B \psi \, d\mu = \int_T 1_B \hat{\varphi} \, d\mu \text{ for every } B \in T. \tag{A.3}$$

Proof. Let $\hat{\varphi}$ be the essential infimum of the collection Φ of all $\mathcal{T}$-measurable functions $\varphi: T \to \bar{\mathbb{R}}$ such that $\varphi \geqq \psi$ [23, I]. Then $\hat{\varphi} \geqq \psi$ and $\hat{\varphi} \leqq \varphi$ μ-a.e. for every $\varphi \in \Phi$. Let $B \in \mathcal{T}$ be arbitrary. For every $\mathcal{T}$-measurable function $\varphi: T \to \bar{\mathbb{R}}$ with $\varphi \geqq 1_B \psi$ we have evidently that the function $\varphi': T \to \bar{\mathbb{R}}$, defined by $\varphi'(t) \equiv \varphi(t)$ if $t \in B$ and $\varphi'(t) \equiv \hat{\varphi}(t)$ if $t \notin B$, belongs to the set Φ. Hence $\varphi' \geqq \hat{\varphi}$ μ-a.e., and in particular $\varphi(t) \geqq \hat{\varphi}(t)$ for μ-a.e. $t \in B$. Since $\varphi(t) \geqq 0$ for every $t \notin B$, this proves that $\int_T \varphi \, d\mu \geqq \int_T 1_B \hat{\varphi} \, d\mu$. Thus one inequality in (A.3) has been proven, and the converse inequality holds trivially. □

Appendix B. Let S be a *Suslin* space; by definition this means that there exist a Polish space R and a continuous surjection π from R onto S. Let $\mathcal{B}(S)$ denote the Borel σ-algebra on S. Let $(T, \mathcal{T}, \mu)$ be a σ-finite complete measure space and let $\mathcal{M}_S$

be the set of all $(\mathcal{T}, \mathcal{B}(S))$-measurable functions $u: T \to S$ such that $u(T)$ is a relatively compact subset of S. A set $\mathcal{U}$ of $(\mathcal{T}, \mathcal{B}(S))$-measurable functions from T into S is said to be *decomposable* if for every $u \in \mathcal{U}$, $v \in \mathcal{M}_S$ and $A \in \mathcal{T}$, $\mu(A) < +\infty$, the function $u': T \to S$, defined by $u'(t) \equiv v(t)$ if $t \in A$, $u'(t) \equiv u(t)$ if $t \notin A$, belongs to $\mathcal{U}$. For every function $g: T \times S \to \bar{\mathbb{R}}$ and every set $\mathcal{V}$ of $(\mathcal{T}, \mathcal{B}(S))$-measurable functions the integral functional $I_g: \mathcal{V} \to \bar{\mathbb{R}}$ is defined by

$$I_g(v) \equiv \tilde{\int}_T g(t, v(t)) \mu(dt),$$

where we use outer integration (cf. Appendix A). Our first result is due to Ioffe-Tikhomirov and Rockafellar [18], [25]; its present general form can be gleaned from the work by Castaing-Valadier [7, VII].

THEOREM B.1 (reduction theorem). *For every $\mathcal{T} \times \mathcal{B}(S)$-measurable function $g: T \times S \to \bar{\mathbb{R}}$ and every decomposable set $\mathcal{U}$ of $(\mathcal{T}, \mathcal{B}(S))$-measurable functions from T into S we have*

$$\inf_{u \in \mathcal{U}} I_g(u) = \int_T \inf_{s \in S} g(t, s) \mu(dt), \tag{B.1}$$

provided that the left-hand side does not equal $+\infty$.

Proof. Define $\gamma: T \to \mathbb{R}$ by $\gamma(t) \equiv \inf_S g(t, s)$; then it follows immediately from a well-known projection theorem [7, III.23] that γ is $\mathcal{T}$-measurable. Evidently, one inequality in (B.1) is trivial; to prove the other, let $\alpha \in \mathbb{R}$ be arbitrary, $\alpha > \int_T \gamma \, d\mu$. We will show that there exists $u \in \mathcal{U}$ such that $I_g(u) < \alpha$. By hypothesis, there exists $u_0 \in \mathcal{U}$ such that $I_g(u_0) < +\infty$. As μ is a σ-finite measure, there exists a strictly positive function φ_0 in $\mathcal{L}^1_{\mathbb{R}}$. Define for every $\varepsilon > 0$ the function $\varphi_\varepsilon: T \to \mathbb{R}$ by

$$\varphi_\varepsilon(t) \equiv \varepsilon \varphi_0(t) + \max(\gamma(t), -\varepsilon^{-1}).$$

For $\varepsilon > 0$ small enough φ_ε will satisfy $\int_T \varphi_\varepsilon \, d\mu < \alpha$ by the monotone convergence theorem. Since $\varphi_\varepsilon(t) > \gamma(t)$ for all $t \in T$, the multifunction $\Gamma: T \rightrightarrows S$, defined by

$$\Gamma(t) \equiv \{s \in S: g(t, s) < \varphi_\varepsilon(t)\}$$

is nonempty-valued. By measurability of g, φ_0 and γ, the graph of Γ is $\mathcal{T} \times \mathcal{B}(S)$-measurable. Hence, it follows from the Von Neumann-Aumann theorem [7, III.22] that there exists a $(\mathcal{T}, \mathcal{B}(S))$-measurable function $v: T \to S$ such that $g(t, v(t)) < \varphi_\varepsilon(t)$ for every $t \in T$. Consider now the image μ^v of the measure μ under the mapping v. Since S is Suslin, every finite measure on $(S, \mathcal{B}(S))$ is Radon [11, III.69]; μ being σ-finite, this quickly gives the existence of a nondecreasing sequence $\{K_i\}_1^\infty$ of compact subsets of S such that $\mu^v(T \setminus \bigcup_{i=1}^\infty K_i) = 0$ and $\mu^v(K_i) < +\infty$ for every $i \in \mathbb{N}$. Define $T_i \in \mathcal{T}$ to be the inverse image of K_i under v. Then clearly $v(T_i)$ is relatively compact in S and $\mu(T_i) < +\infty$ for every $i \in \mathbb{N}$; also, $\mu(T \setminus \bigcup_{i=1}^\infty T_i) = 0$. Defining $u_i: T \to S$ by $u_i(t) \equiv v(t)$ if $t \in T_i$, $u_i(t) \equiv u_0(t)$ if $t \notin T_i$, we know that $u_i \in \mathcal{U}$ for every $i \in \mathbb{N}$ by the decomposability of $\mathcal{U}$. Now

$$I_g(u_i) \leqq \int_{T_i} \varphi_\varepsilon \, d\mu + \int_{T \setminus T_i} g(t, u_0(t)) \mu(dt),$$

so for $i \in \mathbb{N}$ sufficiently large it follows from the above that $I_g(u_i) < \alpha$. □

It is easy to see that the above result applies to taking suprema as well, mutatis mutandis.

THEOREM B.2. *For every pair of $\mathcal{T} \times \mathcal{B}(S)$-measurable functions $g_1, g_2: T \times X \to (-\infty, +\infty]$ and every decomposable set $\mathcal{U}$ of $(\mathcal{T}, \mathcal{B}(S))$-measurable functions from T into*

S such that $g_1 \leqq g_2$ *and*

$$I_{g_1}(u) = I_{g_2}(u) \quad \textit{for all } u \in \mathcal{U},$$

we have

$$g_1(t, \cdot) = g_2(t, \cdot) \ \mu\text{-a.e.}, \tag{B.2}$$

provided that there exists $u_0 \in \mathcal{U}$ *for which both* $t \mapsto g_1(t, u_0(t))$ *and* $t \mapsto g_2(t, u_0(t))$ *are* μ*-integrable.*

Proof. If (B.2) were false, there would exist a set B in $\mathcal{T}$, $0 < \mu(B) < +\infty$, such that for every $t \in B$ the set

$$\Delta(t) \equiv \{s \in S: g_1(t, s) < g_2(t, s)\}$$

is nonempty. By the von Neumann–Aumann theorem [7, III.22] there would exist a $\mathcal{T}$-measurable function $v: B \to S$ such that $v(t) \in \Delta(t)$ for all $t \in B$. Just as in the proof of Theorem B.1, the fact that S is Suslin implies the existence of a nondecreasing sequence $\{B_i^\infty\}_1$ of $\mathcal{T}$-measurable subsets of B, $\lim_i \mu(B_i) = \mu(B)$ and with $v(B_i)$ relatively compact in S for every $i \in \mathbb{N}$. We now define a nondecreasing sequence $\{B_i'\}^\infty$ of $\mathcal{T}$-measurable sets by

$$B_i' \equiv \{t \in B_i: -i \leqq g_1(t, v(t)) \leqq i\}.$$

It is easy to see that $\lim_i \mu(B_i') = \mu(B)$ and that $B_i' \subset B_i$ for every $i \in \mathbb{N}$. Defining $u_i: T \to S$ by $u_i(t) \equiv v(t)$ if $t \in B_i'$ and $u_i(t) \equiv u_0(t)$ if $t \notin B_i'$, we have $u_i \in \mathcal{U}$ by decomposability and the above. Finally, this gives $I_{g_1}(u_i) < I_{g_2}(u_i)$ for every $i \in \mathbb{N}$ such that $\mu(B_i') > 0$, since we obviously have $g_1(t, u_0(t)) = g_2(t, u_0(t))$ μ-a.e. Thus we have arrived at the desired contradiction. □

A function $g: T \times S \to \bar{\mathbb{R}}$ is defined to be a *normal integrand* on $T \times S$ if g is $\mathcal{T} \times \mathcal{B}(S)$-measurable and $g(t, \cdot)$ is lower semicontinuous on S for every $t \in T$. The following result can be found in [1d]; it is also known for multifunctions in a rather similar form [28].

THEOREM B.3. *For every collection* $\mathcal{G}$ *of normal integrands on* $T \times S$ *there exists a countable subset* $\mathcal{G}_0$ *of* $\mathcal{G}$ *such that* $\hat{g} \equiv \sup_{g \in \mathcal{G}_0} g$ *satisfies*

$$\hat{g}(t, \cdot) \geqq g(t, \cdot) \ \mu\textit{-a.e. for every } g \in \mathcal{G}.$$

Proof. In case S is a metric Suslin space, the result follows by [1d, Thm. 2.2, Example A.2]. In general, let the Polish space R and the continuous surjection $\pi: R \to S$ be as at the beginning of this appendix. Define a collection of normal integrands on $T \times R$ by setting $g^\pi(t, r) \equiv g(t, \pi(r))$, $g \in \mathcal{G}$; the result then follows directly from the preceding one. □

Let $(V, P, \langle \cdot, \cdot \rangle)$ be a pair of Suslin locally convex spaces, paired by a strict duality $\langle \cdot, \cdot \rangle$, equipped with topologies which are compatible with the duality and which make V and P into Suslin spaces. A $(\mathcal{T}, \mathcal{B}(V))$-measurable function $v: T \to V$ which is such that for every $p \in P$ the function $t \mapsto \langle v(t), p \rangle$—$\mathcal{T}$-measurable by [7, III.36]—is μ-integrable is said to be *scalarly* μ*-integrable.* Incidentally, note that the obvious notion of scalar $\mathcal{T}$-measurability coincides with $(\mathcal{T}, \mathcal{B}(V))$-measurability [7, III.36].

Acknowledgments. The author is indebted to Dr. Paulette Clauzure for pointing out an error in earlier versions of Propositions 2.19 and 2.20, and for suggesting a number of other improvements. He also wishes to thank the referees for their constructive comments and for providing useful references.

REFERENCES

[1a] E. J. BALDER, *Lower semicontinuity of integral functionals with nonconvex integrands by relaxation-compactification*, this Journal, 19 (1981), pp. 533-542.

[1b] ———, *Lower closure problems with weak convergence conditions in a new perspective*, this Journal, 20 (1982), pp. 198-210.

[1c] ———, *On lower closure and lower semicontinuity in the existence theory for optimal control*, in System Modelling and Optimization, R. F. Drenick and F. Kozin, eds., Lecture Notes in Control and Information Sciences 38, Springer-Verlag, Berlin, 1982, pp. 158-164.

[1d] ———, *An extension of the essential supremum concept with applications to normal integrands and multifunctions*, Bull. Austral. Math. Soc., 27 (1983), pp. 407-418.

[1e] ———, *A general approach to lower semicontinuity and lower closure in optimal control theory*, this Journal, 22 (1984), pp. 570-598.

[1f] ———, *An extension of Prohorov's theorem for transition probabilities with applications to infinite-dimensional lower closure problems*, Rend. Circ. Mak. Palermo (2), 34 (1985), to appear.

[2] C. BERGE, *Espaces topologiques, fonctions multivoques*, Dunod, Paris, 1959.

[3] G. BOTTARO AND P. OPPEZZI, *Semicontinuità inferiore di un funzionale integrale dipendente da funzioni a valori in uno spazio di Banach*, Boll. Un. Mat. Ital., 17-B (1980), pp. 1290-1307.

[4] J. K. BROOKS AND R. V. CHACON, *Continuity and compactness of measures*, Adv. Math., 37 (1980), pp. 16-26.

[5] O. CALIGARIS AND P. OLIVA, *Nonconvex control problems in Banach spaces*, Appl. Math. Optim., 5 (1979), pp. 315-329.

[6] C. CASTAING AND P. CLAUZURE, *Semicontinuité des fonctionelles integrales*, Travaux du Séminaire d'Analyse Convexe, Université des Sciences et Techniques du Languedoc, Montpellier, 1981, pp. 15.1-15.45.

[7] C. CASTAING AND M. VALADIER, *Convex Analysis and Measurable Multifunctions*, Lecture Notes in Mathematics 580, Springer-Verlag, Berlin, 1977.

[8a] L. CESARI, *Existence theorems for weak and usual optimal solutions in Lagrange problems with unilateral constraints* I, Trans. Amer. Math. Soc., 124 (1966), pp. 369-412.

[8b] ———, *Seminormality and upper semicontinuity in optimal control*, J. Optim. Theory Appl., 6 (1970), pp. 114-137.

[8c] ———, *Closure, lower closure and semicontinuity theorems in optimal control*, this Journal, 9 (1971), pp. 287-315.

[8d] ———, *Optimization—Theory and Applications*, Springer-Verlag, Berlin, 1983.

[9] G. CHOQUET, *Lectures on Analysis*, Benjamin, Reading, MA, 1969.

[10] B. DACOROGNA, *Weak Continuity and Weak Lower Semicontinuity of Non-Linear Functionals*, Lecture Notes in Mathematics 580, Springer-Verlag, Berlin, 1982.

[11] C. DELLACHERIE AND P.-A. MEYER, *Probabilities and Potential*, Hermann, Paris, 1975; English transl., North-Holland, Amsterdam, 1976.

[12] J. DIESTEL AND J. J. UHL, *Vector-Measures*, Mathematical Surveys 15, American Mathematical Society, Providence, RI, 1977.

[13] I. EKELAND AND R. TEMAM, *Convex Analysis and Variational Problems*, Dunod, Paris, 1972; English transl., North-Holland, Amsterdam, 1976.

[14] K. FLORET, *Weakly Compact Sets*, Lecture Notes in Mathematics 801, Springer-Verlag, Berlin, 1980.

[15] G. S. GOODMAN, *The duality of convex functions and Cesari's property* (*Q*), J. Optim. Theory Appl., 19 (1976), pp. 17-23.

[16] R. B. HOLMES, *Geometric Functional Analysis and Its Applications*, Springer-Verlag, Berlin, 1975.

[17] A. D. IOFFE, *On lower semicontinuity of integral functionals* I, II, this Journal, 15 (1977), pp. 521-538, pp. 991-1000.

[18] A. D. IOFFE AND V. M. TIKHOMIROV, *Duality of convex functions and extremum problems*, Uspekhi Mat. Nauk, 23, 6 (1968), pp. 51-116; Russian Math. Surveys, 23, 6 (1968), pp. 53-124.

[19] P. J. KAISER, *Seminormality properties of convex sets*, Rend. Circ. Mat. Palermo (2), 28 (1979), pp. 161-182.

[20] P.-J. LAURENT, *Approximation et optimisation*, Hermann, Paris, 1972.

[21] E. J. MCSHANE, *Existence theorems for ordinary problems of the calculus of variations*, Ann. Scuola Norm. Sup. Pisa (2), 3 (1934), pp. 181-211, pp. 287-315.

[22] N. NAGUMO, *Über die gleichmässige Summierbarkeit und ihre Anwendung auf ein Variationsproblem*, Japan J. Math., 6 (1929), pp. 173-182.

[23] J. NEVEU, *Foundations of the Calculus of Probability*, Masson, Paris, 1964; English transl., Holden-Day, San Francisco, 1965.

[24a] C. OLECH, *Weak lower semicontinuity of integral functionals*, J. Optim. Theory Appl., 19 (1976), pp. 3–16.

[24b] ———, *A characterization of L_1-weak lower semicontinuity of integral functionals*, Bull. Acad. Pol. Sci., Ser. Sci. Math. Astron. Phys., 25 (1977), pp. 135–142.

[25a] R. T. ROCKAFELLAR, *Integrals which are convex functionals* I, II, Pacific J. Math., 24 (1968), pp. 525–539, 39 (1971), pp. 439–469.

[25b] ———, *Existence theorems for general control problems of Bolza and Lagrange*, Adv. Math., 15 (1975), pp. 321–333.

[25c] ———, *Integral functionals, normal integrands and measurable selections*, in Nonlinear Operators and the Calculus of Variations, Lecture Notes in Mathematics 543, Springer-Verlag, Berlin, 1976, pp. 157–207.

[26] R. D. RUPP, *Hypotheses implying Cesari's property (Q)*, J. Optim. Theory Appl., 19 (1976), pp. 119–123.

[27a] L. TONELLI, *Fondamenti di Calcolo delle Variazioni*, Zanichelli, Bologna, 1921.

[27b] ———, *Sugli integrali del calcolo delle variazioni in forma ordinaria*, Ann. Scuola Norm. Sup. Pisa (2), 3 (1934), pp. 401–450.

[28] M. VALADIER, *Multi-applications mesurables à valeurs convexes compactes*, J. Math. Pures Appl., 50 (1971), pp. 265–297.

[29] T. ZOLEZZI, *On stability analysis in mathematical programming*, in Mathematical Programming with Data Perturbations, H. V. Fiacco, ed., Mathematical Programming Study 21, North-Holland, Amsterdam, 1984.

SIAM J. CONTROL AND OPTIMIZATION.
Vol. 24, No. 1, January 1986

006

THE OPTIMAL PROJECTION EQUATIONS FOR FINITE-DIMENSIONAL FIXED-ORDER DYNAMIC COMPENSATION OF INFINITE-DIMENSIONAL SYSTEMS*

DENNIS S. BERNSTEIN† AND DAVID C. HYLAND†

Abstract. One of the major difficulties in designing implementable finite-dimensional controllers for distributed parameter systems is that such systems are inherently infinite dimensional while controller dimension is severely constrained by on-line computing capability. While some approaches to this problem initially seek a correspondingly infinite-dimensional control law whose finite-dimensional approximation may be of impractically high order, the usual engineering approach involves first approximating the distributed parameter system with a high-order discretized model followed by design of a relatively low-order dynamic controller. Among the numerous approaches suggested for the latter step are model/controller reduction techniques used in conjunction with the standard *LQG* result. An alternative approach, developed in [36], relies upon the discovery in [31] that the necessary conditions for optimal fixed-order dynamic compensation can be transformed into a set of equations possessing remarkable structural coherence. The present paper generalizes this result to apply directly to the distributed parameter system itself. In contrast to the pair of operator Riccati equations for the "full-order" *LQG* case, the optimal finite-dimensional fixed-order dynamic compensator is characterized by *four* operator equations (two modified Riccati equations and two modified Lyapunov equations) coupled by an oblique projection whose rank is precisely equal to the order of the compensator and which determines the optimal compensator gains. This "optimal projection" is obtained by a full-rank factorization of the product of the finite-rank nonnegative-definite Hilbert-space operators which satisfy the pair of modified Lyapunov equations. The coupling represents a graphic portrayal of the demise of the classical separation principle for the finite-dimensional reduced-order controller case. The results obtained apply to a semigroup formulation in Hilbert space and thus are applicable to control problems involving a broad range of specific partial and functional differential equations.

Key words. optimality conditions, finite-dimensional fixed-order dynamic compensator, infinite-dimensional system, distributed parameter system, semisimple operator, oblique projection, Drazin generalized inverse

1. Introduction.

One of the major difficulties in designing active controllers for distributed parameter systems is that such systems are inherently infinite dimensional while implementable controllers are necessarily finite dimensional with controller dimension severely constrained by on-line computing capability. As pointed out by Balas ([1], see also [2]), control design for distributed parameter systems entails the *practical constraints* of 1) finitely many sensors and actuators, 2) a finite-dimensional controller and 3) natural system dissipation. The validity of 2) is apparent from the fact that processing and transmitting electrical signals by conventional analog or digital components constitutes finite-dimensional action. Although distributed parameter devices can also be utilized, their fabrication and implementation can incorporate at most a finite number of design specifications.[1] Hence, although distributed parameter systems are most accurately represented by infinite-dimensional models, real-world

* Received by the editors December 6, 1983, and in revised form September 15, 1984. This work was performed at Lincoln Laboratory/MIT and was sponsored by the Department of the Air Force.

† Harris Corporation, Government Aerospace Systems Division, Controls Analysis and Synthesis Group, Melbourne, Florida 32901.

[1] Examples of such components include tapped delay lines and surface acoustic wave devices. Although acoustoelectric convolvers [3, p. 465] can perform continuous-time integration, synthesis of the desired impulse-response kernel can incorporate only finitely many specified parameters. The obvious fact should also be noted that physical limitations impose an upper bound on the number of design parameters that can be incorporated in the construction of *any* device. For an extensive treatment of this subject, see [72].

constraints require that implementable controllers be modelled as lumped parameter systems.

Clearly, the above observations effectively preclude the possibility of realizing infinite-dimensional controllers that involve full-state feedback or full-state estimation (see, e.g., [4]-[6] and the numerous references therein). Although finite-dimensional approximation schemes have been applied to optimal infinite-dimensional control laws ([7]-[9]), these results only guarantee optimality in the limit, i.e., as the order of the approximating controller increases without bound. Hence, there is no guarantee that a particular approximate (i.e., discretized) controller is actually optimal over the class of approximate controllers of a given order dictated by implementation constraints. Moreover, even if an optimal *approximate* finite-dimensional controller could be obtained, it would almost certainly be suboptimal in the class of *all* controllers of the given order.

Although the usual engineering approach to this problem is to replace the distributed parameter system with a high-order finite-dimensional model, analogous, fundamental difficulties remain since application of *LQG* leads to a controller whose order is identical to that of the high-order approximate model. Attempts to remedy this problem usually rely upon some method of open-loop model reduction or closed-loop controller reduction (see, e.g., [10]-[15]). Most of these techniques (with the exception of [11]) are ad hoc in nature, however, and hence guarantees of optimality and stability may be lacking.

A more direct approach that avoids both model and controller reduction is to fix the controller structure and optimize the performance criterion with respect to the controller parameters. Although much effort was devoted to this approach (see, e.g., [16]-[30]), progress in this direction was impeded by the extreme complexity of the nonlinear matrix equations arising from the first-order necessary conditions. What was lacking, to quote the insightful remarks of [24], was a "deeper understanding of the structural coherence of these equations." The key to unlocking these unwieldy equations was subsequently discovered by Hyland in [31] and developed in [32]-[36]. Specifically, it was found that these equations harbored the definition of an *oblique projection* (i.e., idempotent matrix) which is a consequence of optimality and *not* the result of an ad hoc assumption. By exploiting the presence of this "optimal projection," the originally *very* complex stationary conditions can be transformed without loss of generality into much simpler and more tractable forms. The resulting equations (see [36, (2.10)-(2.17)]) preserve the simple form of *LQG* relations for the gains in terms of covariance and cost matrices which, in turn, are determined by a coupled system of two modified Riccati equations and two modified Lyapunov equations. This coupling, by means of the optimal projection, represents a graphic portrayal of the demise of the classical separation principle for the reduced-order controller case. When, as a special case, the order of the compensator is required to be equal to the order of the plant, the modified Riccati equations immediately reduce to the standard *LQG* Riccati equations and the modified Lyapunov equations express the proviso that the compensator be minimal, i.e., controllable and observable. Since the *LQG* Riccati equations as such are nothing more than the necessary conditions for full-order compensation, the "optimal projection equations" appear to provide a clear and simple generalization of standard *LQG* theory.

The fact that the optimal projection equations consist of *four* coupled matrix equations, i.e., two modified Riccati equations and two modified Lyapunov equations, can readily be explained by the following simple reason. Reduced-order control-design methods often involve either *LQG* applied to a reduced-order model or model reduction

applied to a full-order *LQG* design, and hence both approaches require the solution of precisely four equations: two Riccati equations (for *LQG*) plus two Lyapunov equations (for system reduction via balancing, as in [12], [14]). The *coupled* form of the optimal projection equations is thus a strong reminder that the *LQG* and order-reduction operations *cannot* be iterated but must, in a precise sense, be performed *simultaneously*. This situation is partly due to the fact that the optimal projection matrix may not be of the form $\left[\begin{smallmatrix} I & 0 \\ 0 & 0 \end{smallmatrix}\right]$ even in the basis corresponding to the "balanced" realization [12], [14]. This point is explored in [37], [37a] where the solution to the optimal model-reduction problem is characterized by a pair of modified Lyapunov equations which are also coupled by an oblique projection.

Returning now to the distributed parameter problem, it should be mentioned that notable exceptions to the previously mentioned work on distributed parameter controllers are the contributions of Johnson [38] and Pearson [39], [40] who suggest fixing the order of the finite-dimensional compensator while retaining the distributed parameter model. Progress in this direction, however, was impeded not only by the intractability of the optimality conditions that were available for the finite-dimensional problem (as in [16]–[30]), but also by the lack of a suitable generalization of these conditions to the infinite-dimensional case. The purpose of the present paper is to make significant progress in filling these gaps, i.e., by deriving explicit optimality conditions which directly characterize the optimal finite-dimensional fixed-order dynamic compensator for an infinite-dimensional system and which are exactly analogous to the highly simplified optimal projection equations obtained in [31]–[34], [36] for the finite-dimensional case. Specifically, instead of a system for four matrix equations we obtain a system of four *operator* equations whose solutions characterize the optimal finite-dimensional fixed-order dynamic compensator. Moreover, the optimal projection now becomes a bounded idempotent Hilbert-space operator whose rank is precisely equal to the order of the compensator.

The mathematical setting we use is standard: a linear time-invariant differential system in Hilbert space with additive white noise, finitely many controls and finitely many noisy measurements (thus satisfying the first practical constraint mentioned above). The input and output maps are assumed to be bounded. Since the only explicit assumption on the unbounded dynamics operator is that it generate a strongly continuous semigroup, the results are potentially applicable to a broad range of specific partial and functional differential equations. The *actual* applicability of our results is essentially limited by practical constraint 3). Since we are concerned with the steady-state problem, we implicitly assume that the distributed parameter system is stabilizable, i.e., that there exists a dynamic compensator of a given order such that the closed-loop system is uniformly stable. We note that stabilizing compensators do exist for the wide class of problems considered in [41] and [42] which includes delay, parabolic and damped hyperbolic systems. The question of *how much* damping is required for stabilizability of hyperbolic systems is a crucial issue in designing controllers for large flexible space structures [7], [43]–[49a].

It is important to point out that the results of this paper can immediately be specialized to finite-dimensional systems by requiring that the Hilbert space characterizing the dynamical system be finite-dimensional. Then all unboundedness considerations can be ignored, adjoints can be interpreted as transposes and other obvious simplifications can be invoked. The only mathematical aspect requiring attention is the treatment of white noise which, for general handling of the infinite-dimensional case, is interpreted according to [6].[2] For the finite-dimensional case, however, the standard

classical notions suffice and the results go through with virtually no modifications.

The contents of the paper are as follows. Section 2 contains preliminary notation in addition to particular results for use later in the paper. Section 3 presents the optimal steady-state finite-dimensional fixed-order dynamic-compensation problem and the Main Theorem gives the necessary conditions in the form of the optimal projection equations (3.15)-(3.18). We then develop a series of results which serve to elucidate several aspects of the Main Theorem. Section 4 is devoted to the proof of the Main Theorem. The reader is alerted to the two crucial steps required. The first step involves generalizing to the infinite-dimensional case the derivation of the necessary conditions in their "primitive" form (see (4.27)-(4.29) and (4.48)-(4.53)). The derivation in [31]-[33], [36] involving Lagrange multipliers is invalid in the infinite-dimensional case due to the presence of the unbounded system-dynamics operator. Instead, we use the gramian form of the closed-loop covariance operator to obtain a dual problem formulation and then proceed to derive the primitive necessary conditions by means of a lengthy, but direct, computation (Lemma 4.7). The second crucial step involves transforming the primitive form of the necessary conditions to the final form given in the Main Theorem. This laborious computation was first carried out in [31], [32] and was subsequently facilitated in [33], [36] by means of a judicious change of variables (see (4.32), (4.33)). Finally, some concluding remarks are given in § 5.

2. Preliminaries. In this section we introduce general notation along with basic definitions and results for use in later sections. Our principal references are [6], [50] and [51].

Throughout this section let $\mathscr{H}$, $\mathscr{H}'$ and $\mathscr{H}''$ denote real separable Hilbert spaces with norm $\|\cdot\|$ and inner product $\langle\cdot,\cdot\rangle$ and let $\mathscr{B}(\mathscr{H},\mathscr{H}')$ denote the space of bounded linear operators from $\mathscr{H}$ into $\mathscr{H}'$. For $L\in\mathscr{B}(\mathscr{H},\mathscr{H}')$, $\|L\|$ is the norm of L, $\mathscr{R}(L)$ is the range of L, $\mathscr{N}(L)$ is the null space of L, $\rho(L)$ is the rank of L (set $\rho(L)=\infty$ if L does not have finite rank), L^{-1} is the inverse of L when L is invertible, i.e., when L has a bounded inverse, L^* is the adjoint of L and $L^{-*}\triangleq(L^*)^{-1}$. Recall that $\|L\|=\|L^*\|$ and that $\rho(L)=\rho(L^*)$ [50, p. 161]. Now suppose that $\mathscr{H}=\mathscr{H}'$ so that $L\in\mathscr{B}(\mathscr{H})\triangleq\mathscr{B}(\mathscr{H},\mathscr{H})$. If $LL^*=L^*L$ then L is *normal* and if $L=L^*$ then L is *selfadjoint.* If L is selfadjoint and $\langle Lx,x\rangle\geqq 0$, $x\in\mathscr{H}$, then L is *nonnegative definite.* Note that the selfadjointness assumption is included in the definition since the Hilbert spaces are assumed real. If L is nonnegative definite then $L^{1/2}$ denotes the (unique) nonnegative-definite square root of L. Call L semisimple (resp., real semisimple, nonnegative semisimple) if there exists invertible $S\in\mathscr{B}(\mathscr{H})$ such that SLS^{-1} is normal (resp., selfadjoint, nonnegative definite). This implies that SLS^{-1} has a complete set of orthonormal eigenvectors and, in the real-semisimple or nonnegative-semisimple cases, has real or nonnegative eigenvalues.

Recall that if $L\in\mathscr{B}(\mathscr{H})$ is compact then L has at most a countable number of eigenvalues and all nonzero eigenvalues have finite multiplicity. Hence, for $L\in\mathscr{B}(\mathscr{H},\mathscr{H}')$ compact, let $\{\alpha_i\}$ be the (at most countable) sequence of eigenvalues of $(LL^*)^{1/2}$ with appropriate multiplicity and $\alpha_1\geqq\alpha_2\geqq\cdots>0$ [50, p. 261]. Then $\mathscr{B}_1(\mathscr{H},\mathscr{H}')$ denotes the set of *trace class* (or *nuclear*) operators, i.e., the set of compact

[2] Alternatively, we could have adopted the white noise formulation of [4]. The main difference between the two white noise formalisms is that Balakrishnan works with finitely additive rather than countably additive measures. Strictly speaking, then, even in finite dimensions Balakrishnan's white noise is different from the standard notion (see [6, pp. 307, 315]).

$L \in \mathcal{B}(\mathcal{H}, \mathcal{H}')$ for which $\sum_i \alpha_i < \infty$ [50, p. 521]. $\mathcal{B}_1(\mathcal{H}, \mathcal{H}')$ is a Banach space with norm

$$\|L\|_1 \triangleq \sum_i \alpha_i.$$

If $\sum_i \alpha_i^2 < \infty$ then $L \in \mathcal{B}_2(\mathcal{H}, \mathcal{H}')$, the set of Hilbert-Schmidt operators, which is a Banach space with norm

$$\|L\|_2 \triangleq \left[\sum_i \alpha_i^2\right]^{1/2}.$$

Note that $\|L\| \leqq \|L\|_2 \leqq \|L\|_1$, $\|L\| = \|L^*\|$, $\|L\|_1 = \|L^*\|_1$ and $\|L\|_2 = \|L^*\|_2$. If $\mathcal{H} = \mathcal{H}'$, then we write $\mathcal{B}_1(\mathcal{H})$ and $\mathcal{B}_2(\mathcal{H})$ for $\mathcal{B}_1(\mathcal{H}, \mathcal{H})$ and $\mathcal{B}_2(\mathcal{H}, \mathcal{H})$, respectively. Note that if nonnegative-definite $L \in \mathcal{B}_1(\mathcal{H})$ then $L^{1/2} \in \mathcal{B}_2(\mathcal{H})$.

If $L \in \mathcal{B}_1(\mathcal{H}, \mathcal{H}')$ and $S \in \mathcal{B}(\mathcal{H}', \mathcal{H}'')$ then

$$\|SL\|_1 \leqq \|S\| \, \|L\|_1$$

and hence $SL \in \mathcal{B}_1(\mathcal{H}, \mathcal{H}')$. Similarly, under suitable hypotheses,

$$\|LS\|_1 \leqq \|S\| \, \|L\|_1,$$

and

$$\|SL\|_1 \leqq \|S\|_2 \|L\|_2.$$

LEMMA 2.1. *Suppose $L \in \mathcal{B}_1(\mathcal{H})$ and let $\{\lambda_i\}$ denote the nonzero eigenvalues of L with appropriate multiplicity. Then* [51, *p.* 89]

$$\sum_i |\lambda_i| \leqq \|L\|_1.$$

If L is selfadjoint then [50, *p.* 522]

$$\sum_i |\lambda_i| = \|L\|_1.$$

If L is nonnegative definite then

$$\sum_i \lambda_i = \|L\|_1.$$

Let $L \in \mathcal{B}_1(\mathcal{H})$. Then define [50, p. 523] the trace functional tr: $\mathcal{B}_1(\mathcal{H}) \to \mathbb{R}$ by

$$\operatorname{tr} L \triangleq \sum_i \langle L\phi_i, \phi_i \rangle,$$

where the summation is independent of the choice of orthonormal basis $\{\phi_i\}$. The trace satisfies $\operatorname{tr} L = \operatorname{tr} L^*$, $\operatorname{tr} SL = \operatorname{tr} LS$ for all $S \in \mathcal{B}(\mathcal{H})$, $\operatorname{tr} ST = \operatorname{tr} TS$ for all $S, T \in \mathcal{B}_2(\mathcal{H})$ and $\operatorname{tr}(\alpha T + \beta S) = \alpha(\operatorname{tr} T) + \beta(\operatorname{tr} S)$ for all $\alpha, \beta \in \mathbb{R}$ and $S, T \in \mathcal{B}_1(\mathcal{H})$.

LEMMA 2.2. *Suppose $L \in \mathcal{B}_1(\mathcal{H})$ and let $\{\lambda_i\}$ denote the nonzero eigenvalues of L with appropriate multiplicity. Then* [51, *p.* 139]

$$\operatorname{tr} L = \sum_i \lambda_i$$

and hence (by Lemma 2.1)

$$|\operatorname{tr} L| \leqq \|L\|_1.$$

If L is nonnegative definite then

$$\operatorname{tr} L = \|L\|_1.$$

COROLLARY 2.1. *For each* $S \in \mathcal{B}(\mathcal{H})$ *the linear functionals*

$$L \to \operatorname{tr} SL\colon \mathcal{B}_1(\mathcal{H}) \to \mathbb{R},$$

$$L \to \operatorname{tr} LS\colon \mathcal{B}_1(\mathcal{H}) \to \mathbb{R}$$

are continuous. For each $L \in \mathcal{B}_1(\mathcal{H})$ *the linear functionals*

$$S \to \operatorname{tr} LS\colon \mathcal{B}(\mathcal{H}) \to \mathbb{R},$$

$$S \to \operatorname{tr} SL\colon \mathcal{B}(\mathcal{H}) \to \mathbb{R}$$

are continuous.

Although showing that a bounded linear operator is trace class is slightly more involved than the above characterizations of $\mathcal{B}_1(\mathcal{H})$, the following result will suffice for our purposes (see [52, p. 96], or [52a, p. 171]).

LEMMA 2.3. *Let* $L \in \mathcal{B}(\mathcal{H})$ *be nonnegative definite. Then*

$$\sum_i \langle L\phi_i, \phi_i \rangle,$$

whether finite or infinite, is independent of the orthonormal basis $\{\phi_i\}$. *The summation is finite if and only if* $L \in \mathcal{B}_1(\mathcal{H})$.

Many of the operators introduced in the following section have finite-dimensional domain or range space and hence are degenerate, i.e., have finite rank. Recall that degenerate operators are necessarily trace class. The following result, which generalizes [53, Thm. 2.1, p. 240] in certain respects, will be fundamental in decomposing finite-rank operators.

LEMMA 2.4. Suppose $L_1, \cdots, L_r \in B(\mathcal{H}, \mathcal{H}')$ have finite rank. Then there exists a finite-dimensional subspace $\mathcal{M} \subset \mathcal{H}$ such that $L_i\mathcal{M}^\perp = 0$, $i = 1, \cdots, r$. Furthermore, if $\mathcal{H} = \mathcal{H}'$ then $\mathcal{M}$ can be chosen such that $L_i\mathcal{M} \subset \mathcal{M}$, $i = 1, \cdots, r$.

Proof. It suffices to consider the case $r = 1$. Writing L for L_1, note that since $\rho(L^*) < \infty$, $\mathcal{N}(L)^\perp = \mathcal{R}(L^*)$ [50, p. 155] and $\mathcal{N}(L)$ is closed, the first statement holds with $\mathcal{M} = \mathcal{N}(L)^\perp$. When $\mathcal{H} = \mathcal{H}'$ set $\mathcal{M} = \mathcal{N}(L)^\perp + \mathcal{R}(L)$ and note that $\mathcal{M}^\perp = \mathcal{N}(L) \cap \mathcal{R}(L)^\perp \subset \mathcal{N}(L)$ and $L\mathcal{M} \subset \mathcal{R}(L) \subset \mathcal{M}$. □

The following generalization of Sylvester's inequality [54, p. 66] will be used repeatedly in handling finite-rank operators.

LEMMA 2.5. *Let* $L \in \mathcal{B}(\mathcal{H}, \mathcal{H}')$ *and* $S \in \mathcal{B}(\mathcal{H}', \mathcal{H}'')$. *Then*

$$\rho(SL) \leqq \min\{\rho(S), \rho(L)\}. \tag{2.1}$$

If $\dim \mathcal{H}' = \nu < \infty$, *then*

$$\rho(S) + \rho(L) - \nu \leqq \rho(SL). \tag{2.2}$$

Proof. If either S or L does not have finite rank then (2.1) is immediate. If both S and L have finite rank then the standard arguments [54] used to prove the finite-dimensional version of (2.1) remain valid. To prove (2.2), note that Lemma 2.4 implies that there exist orthonormal bases for $\mathcal{H}$ and $\mathcal{H}'$ with respect to which L has the matrix representation $[\tilde{L}\ 0]$, where $\tilde{L} \in \mathbb{R}^{\nu \times p}$. Similarly, there exist orthonormal bases for $\mathcal{H}'$ and $\mathcal{H}''$ with respect to which S has the matrix representation $\left[\begin{smallmatrix} \tilde{S} \\ 0 \end{smallmatrix}\right]$, where $\tilde{S} \in \mathbb{R}^{q \times \nu}$. Since the two cited bases for $\mathcal{H}'$ may be different, let orthogonal $U \in \mathbb{R}^{\nu \times \nu}$ be the matrix representation (with respect to either basis for $\mathcal{H}'$) for the change in orthonormal basis [6, p. 100]. Hence SL has the matrix representation $\left[\begin{smallmatrix} \tilde{S}U\tilde{L} & 0 \\ 0 & 0 \end{smallmatrix}\right]$ and (2.2) follows from the known result [54, p. 66]. □

As in the proof of Lemma 2.5, we shall utilize the infinite-matrix representation of an operator with respect to an orthonormal basis. All matrix representations given

here will consist of real entries since the Hilbert spaces involved are real. When the orthonormal bases are specified and no confusion can arise, we shall not differentiate between an operator and its matrix representation. We shall use the infinite identity matrix I_∞ interchangeably with the identity $I_{\mathscr{H}}$ on $\mathscr{H}$.

When dealing with finite-dimensional Euclidean spaces the notation and terminology introduced above will be utilized with only minor changes. For example, bounded linear operators will be represented by matrices whose elements are determined according to fixed orthonormal bases and hence we identify $\mathbb{R}^{m\times n} = \mathscr{B}(\mathbb{R}^n, \mathbb{R}^m)$. Note that if $L \in \mathscr{B}(\mathbb{R}^n, \mathscr{H})$ and $S \in \mathscr{B}(\mathscr{H}, \mathbb{R}^m)$ then SL is an $m \times n$ matrix which is independent of any particular orthonormal basis for $\mathscr{H}$. The transposes of $x \in \mathbb{R}^n \triangleq \mathbb{R}^{n\times 1}$ and $M \in \mathbb{R}^{m\times n}$ are denoted by x^T and M^T and $M^{-T} \triangleq (M^T)^{-1}$. Let I_n denote the $n \times n$ identity matrix.

To specialize some of the above operator terminology to matrices, let $M \in \mathbb{R}^{n\times n}$. We shall say M is *nonnegative* (resp., *positive*) *diagonal* if M is diagonal with nonnegative (resp., positive) diagonal elements. M is *nonnegative* (resp., *positive*) *definite* if M is symmetric and $x^TMx \geqq 0$ (resp., $x^TMx > 0$), $x \in \mathbb{R}^n$. Recall that M is symmetric (resp., nonnegative definite, positive definite) if and only if there exists orthogonal $U \in \mathbb{R}^{n\times n}$ such that UMU^T is diagonal (resp., nonnegative diagonal, positive diagonal). M is *semisimple* [55, p. 13], or *nondefective* [56, p. 375], if M has n linearly independent eigenvectors, i.e., M has a diagonal Jordan canonical form over the complex field. M is *real* (resp., *nonnegative, positive*) *semisimple* if M is semisimple with real (resp., nonnegative, positive) eigenvalues. Note that M is real (resp., nonnegative, positive) semisimple if and only if there exists invertible $S \in \mathbb{R}^{n\times n}$ such that SMS^{-1} is diagonal (resp., nonnegative diagonal, positive diagonal). Alternatively, M is real (resp., nonnegative, positive) semisimple if and only if there exists invertible $S \in \mathbb{R}^{n\times n}$ such that SMS^{-1} is symmetric (resp., nonnegative definite, positive definite).

LEMMA 2.6. *The product of two nonnegative- (resp., positive-) definite matrices is nonnegative (resp., positive) semisimple.*

Proof. If S, $L \in \mathbb{R}^{n\times n}$ are both nonnegative (resp., positive) definite then by [55, Thm. 6.2.5, p. 123] there exists invertible $\phi \in \mathbb{R}^{n\times n}$ such that $D_S \triangleq \phi^{-1}S\phi^{-T}$ and $D_L \triangleq \phi^TL\phi$ are nonnegative (resp., positive) diagonal. Hence, $SL = \phi D_S D_L \phi^{-1}$ is nonnegative (resp., positive) semisimple, as desired. Alternatively, if either S or L is positive definite, then the result follows from $SL = L^{-1/2}(L^{1/2}SL^{1/2})L^{1/2}$ if L is positive definite or $SL = S^{1/2}(S^{1/2}LS^{1/2})S^{-1/2}$ if S is positive definite. □

3. Problem statement and the Main Theorem. We consider the following steady-state fixed-order dynamic-compensation problem. Given the dynamical system on $[0, \infty)$

$$\dot{x}(t) = Ax(t) + Bu(t) + H_1w(t), \tag{3.1}$$

$$y(t) = Cx(t) + H_2w(t), \tag{3.2}$$

design a finite-dimensional fixed-order dynamic compensator

$$\dot{x}_c(t) = A_cx_c(t) + B_cy(t), \tag{3.3}$$

$$u(t) = C_cx_c(t) \tag{3.4}$$

which minimizes the steady-state performance criterion

$$J(A_c, B_c, C_c) \triangleq \lim_{t\to\infty} \mathbb{E}[\langle R_1x(t), x(t)\rangle + u(t)^TR_2u(t)]. \tag{3.5}$$

The following data are assumed. The state $x(t)$ is an element of a real separable Hilbert space $\mathcal{H}$ and the state differential equation is interpreted in the weak sense (see, e.g., [6, pp. 229, 317]). The closed, densely defined operator $A: \mathcal{D}(A) \subset \mathcal{H} \to \mathcal{H}$ generates a strongly continuous semigroup e^{At}, $t \geqq 0$. The control $u(t) \in \mathbb{R}^m$, $B \in \mathcal{B}(\mathbb{R}^m, \mathcal{H})$ and the operator $R_1 \in \mathcal{B}_1(\mathcal{H})$ and the matrix $R_2 \in \mathbb{R}^{m \times m}$ are nonnegative definite and positive definite, respectively. $w(\cdot)$ is a zero-mean Gaussian "standard white noise process" in $L_2((0, \infty), \mathcal{H}')$ (see [6, p. 314]), where $\mathcal{H}'$ is a real separable Hilbert space, $H_1 \in \mathcal{B}_2(\mathcal{H}', \mathcal{H})$, $H_2 \in \mathcal{B}(\mathcal{H}', \mathbb{R}^l)$ and "$\mathbb{E}$" denotes expectation. We assume that $H_1 H_2^* = 0$, i.e., the disturbance and measurement noises are independent,[3] and that $V_2 \triangleq H_2 H_2^* \in \mathbb{R}^l$ is positive definite, i.e., all measurements are noisy. Note that $V_1 \triangleq H_1 H_1^* \in \mathcal{B}_1(\mathcal{H})$ is nonnegative definite and trace class.[4] The initial state $x(0)$ is Gaussian and independent of $w(\cdot)$. The observation $y(t) \in \mathbb{R}^l$ and $C \in \mathcal{B}(\mathcal{H}, \mathbb{R}^l)$. The dimension of the compensator state $x_c(t)$ is of fixed, finite order $n_c \leqq \dim \mathcal{H}$ and the optimization is performed over $A_c \in \mathbb{R}^{n_c \times n_c}$, $B_c \in \mathbb{R}^{n_c \times l}$ and $C_c \in \mathbb{R}^{m \times n_c}$.

To handle the closed-loop system (3.1)-(3.4), we introduce the augmented state space $\tilde{\mathcal{H}} \triangleq \mathcal{H} \oplus \mathbb{R}^{n_c}$ which is a real separable Hilbert space with inner product $\langle \tilde{x}_1, \tilde{x}_2 \rangle \triangleq \langle x_1, x_2 \rangle + x_{c1}^T x_{c2}$, $\tilde{x}_i \triangleq (x_i, x_{ci})$. An operator $L \in \mathcal{B}(\tilde{\mathcal{H}})$ has a "decomposition" into operators $L_1 \in \mathcal{B}(\mathcal{H})$, $L_{12} \in \mathcal{B}(\mathbb{R}^{n_c}, \mathcal{H})$, $L_{21} \in \mathcal{B}(\mathcal{H}, \mathbb{R}^{n_c})$ and $L_2 \in \mathbb{R}^{n_c \times n_c}$ in the sense that for $\tilde{x} \triangleq (x, x_c) \in \tilde{\mathcal{H}}$, $L\tilde{x} = (L_1 x + L_{12} x_c, L_{21} x + L_2 x_c)$, or, in "block" form,

$$L = \begin{bmatrix} L_1 & L_{12} \\ L_{21} & L_2 \end{bmatrix}.$$

For later use note that

$$\|L\| \leqq \|L_1\| + \|L_{12}\| + \|L_{21}\| + \|L_{22}\|$$

and

$$L^* = \begin{bmatrix} L_1^* & L_{21}^* \\ L_{12}^* & L_2^T \end{bmatrix}.$$

We can similarly construct unbounded operators in $\tilde{\mathcal{H}}$. Hence, define the closed-loop dynamics operator $\tilde{A}: \mathcal{D}(\tilde{A}) \subset \tilde{\mathcal{H}} \to \tilde{\mathcal{H}}$ on the dense domain $\mathcal{D}(\tilde{A}) \triangleq \mathcal{D}(A) \times \mathbb{R}^{n_c}$ by $\tilde{A}\tilde{x} = (Ax + BC_c x_c, B_c C x + A_c x_c)$. Since $\tilde{A}$ can be represented by

$$\tilde{A} = \begin{bmatrix} A & BC_c \\ B_c C & A_c \end{bmatrix} = \begin{bmatrix} A & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & BC_c \\ B_c C & A_c \end{bmatrix}$$

and since the closed-loop operator

$$\begin{bmatrix} A & 0 \\ 0 & 0 \end{bmatrix} : \mathcal{D}(\tilde{A}) \to \tilde{\mathcal{H}}$$

generates the strongly continuous semigroup

$$\begin{bmatrix} e^{At} & 0 \\ 0 & I_{n_c} \end{bmatrix}, \qquad t \geqq 0,$$

it follows from [50, Thm., p. 497] that $\tilde{A}$ is also closed and generates a strongly continuous semigroup $e^{\tilde{A}t} \in \mathcal{B}(\tilde{\mathcal{H}})$, $t \geqq 0$. To guarantee that J is finite and independent

[3] This assumption and its analogue, the lack of a cross-weighting term $x(t)^T R_{12} u(t)$ in (3.5), are for convenience only. See § 5.

[4] We must require that R_1 and V_1 be nuclear since covariance operators in the white noise formulation of [6] are not necessarily trace class as they are in the formulation of [4].

of initial conditions we restrict our attention to the set of admissible stabilizing compensators

$$\mathscr{A} \triangleq \{(A_c, B_c, C_c): e^{\tilde{A}t} \text{ is exponentially stable}\}.$$

Hence if $(A_c, B_c, C_c) \in \mathscr{A}$ then there exist $\alpha > 0$ and $\beta > 0$ such that

$$\|e^{\tilde{A}t}\| \leqq \alpha e^{-\beta t}, \qquad t \geqq 0. \tag{3.6}$$

Since the value of J is independent of the internal realization of the compensator, we can further restrict our attention to

$$\mathscr{A}_+ \triangleq \{(A_c, B_c, C_c) \in \mathscr{A}: (A_c, B_c) \text{ is controllable and } (C_c, A_c) \text{ is observable}\}.$$

The following lemma is required for the statement of the Main Theorem.

LEMMA 3.1. *Suppose $\hat{Q}$, $\hat{P} \in \mathscr{B}(\mathscr{H})$ have finite rank and are nonnegative definite. Then $\hat{Q}\hat{P}$ is nonnegative semisimple. Furthermore, if $\rho(\hat{Q}\hat{P}) = n_c$ then there exist G, $\Gamma \in \mathscr{B}(\mathscr{H}, \mathbb{R}^{n_c})$ and positive-semisimple $M \in \mathbb{R}^{n_c \times n_c}$ such that*

$$\hat{Q}\hat{P} = G^* M \Gamma, \tag{3.7}$$

$$\Gamma G^* = I_{n_c}. \tag{3.8}$$

Proof. By Lemma 2.4 there exists a finite-dimensional subspace $\mathscr{M} \subset \mathscr{H}$ such that $\hat{Q}\mathscr{M} \subset \mathscr{M}$, $\hat{Q}\mathscr{M}^\perp = 0$, $\hat{P}\mathscr{M} \subset \mathscr{M}$ and $\hat{P}\mathscr{M}^\perp = 0$. Hence there exists an orthonormal basis for $\mathscr{H}$ with respect to which $\hat{Q}$ and $\hat{P}$ have the infinite-matrix representations

$$\hat{Q} = \begin{bmatrix} \hat{Q}_1 & 0 \\ 0 & 0 \end{bmatrix}, \qquad \hat{P} = \begin{bmatrix} \hat{P}_1 & 0 \\ 0 & 0 \end{bmatrix},$$

where $\hat{Q}_1, \hat{P}_1 \in \mathbb{R}^{r \times r}$ are nonnegative definite and $r \triangleq \dim \mathscr{M}$. Since by Lemma 2.6 there exists invertible $\Psi \in \mathbb{R}^{r \times r}$ such that $\tilde{\Lambda} = \Psi^{-1}\hat{Q}_1\hat{P}_1\Psi$ is nonnegative diagonal, we have

$$\hat{Q}\hat{P} = \begin{bmatrix} \Psi & 0 \\ 0 & I_\infty \end{bmatrix} \begin{bmatrix} \tilde{\Lambda} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \Psi^{-1} & 0 \\ 0 & I_\infty \end{bmatrix},$$

which shows that $\hat{Q}\hat{P}$ is nonnegative semisimple. If, furthermore, $\rho(\hat{Q}\hat{P}) = n_c$ then it is clear that Ψ can be chosen (i.e., modified by an orthogonal matrix) so that

$$\tilde{\Lambda} = \begin{bmatrix} \Lambda & 0 \\ 0 & 0 \end{bmatrix},$$

where $\Lambda \in \mathbb{R}^{n_c \times n_c}$ is positive diagonal. Hence,

$$\hat{Q}\hat{P} = \begin{bmatrix} \Psi & 0 \\ 0 & I_\infty \end{bmatrix} \begin{bmatrix} \begin{bmatrix} I_{n_c} \\ 0 \end{bmatrix} \\ 0 \end{bmatrix} \Lambda \begin{bmatrix} [I_{n_c} & 0] & 0 \end{bmatrix} \begin{bmatrix} \Psi^{-1} & 0 \\ 0 & I_\infty \end{bmatrix},$$

which shows that (3.7) and (3.8) are satisfied with

$$G = \begin{bmatrix} [S^T & 0] & 0 \end{bmatrix} \begin{bmatrix} \Psi^T & 0 \\ 0 & I_\infty \end{bmatrix}, \qquad M = S^{-1}\Lambda S, \Gamma = \begin{bmatrix} [S^{-1} & 0] & 0 \end{bmatrix} \begin{bmatrix} \Psi^{-1} & 0 \\ 0 & I_\infty \end{bmatrix},$$

for all invertible $S \in \mathbb{R}^{n_c \times n_c}$. □

We shall refer to $G, \Gamma \in \mathscr{B}(\mathscr{H}, \mathbb{R}^{n_c})$ and positive-semisimple $M \in \mathbb{R}^{n_c \times n_c}$ satisfying (3.7) and (3.8) as a *(G, M, Γ)-factorization of $\hat{Q}\hat{P}$*. For convenience in stating the Main Theorem define

$$\sum \triangleq BR_2^{-1}B^*, \qquad \overline{\sum} \triangleq C^* V_2^{-1} C.$$

MAIN THEOREM. *Suppose* $(A_c, B_c, C_c) \in \mathscr{A}_+$ *solves the steady-state fixed-order dynamic-compensation problem. Then there exist nonnegative-definite* $Q, P, \hat{Q}, \hat{P} \in \mathscr{B}_1(\mathscr{H})$ *such that* A_c, B_c *and* C_c *are given by*

$$A_c = \Gamma(A - Q\bar{\Sigma} - \Sigma P)G^*, \tag{3.9}$$

$$B_c = \Gamma Q C^* V_2^{-1}, \tag{3.10}$$

$$C_c = -R_2^{-1} B^* P G^*, \tag{3.11}$$

for some (G, M, Γ)*-factorization of* $\hat{Q}\hat{P}$*, and such that, with* $\tau \triangleq G^*\Gamma$*, the following conditions are satisfied*:

$$Q: \mathscr{D}(A^*) \to \mathscr{D}(A), \qquad P: \mathscr{D}(A) \to \mathscr{D}(A^*), \tag{3.12a, b}$$

$$\hat{Q}: \mathscr{H} \to \mathscr{D}(A), \qquad \hat{P}: \mathscr{H} \to \mathscr{D}(A^*), \tag{3.13a, b}$$

$$\rho(\hat{Q}) = \rho(\hat{P}) = \rho(\hat{Q}\hat{P}) = n_c, \tag{3.14a, b, c}[5]$$

$$0 = (A - \tau Q\bar{\Sigma})Q + Q(A - \tau Q\bar{\Sigma})^* + V_1 + \tau Q\bar{\Sigma}Q\tau^*, \tag{3.15}$$

$$0 = (A - \Sigma P\tau)^* P + P(A - \Sigma P\tau) + R_1 + \tau^* P\Sigma P\tau, \tag{3.16}$$

$$0 = [(A - \Sigma P)\hat{Q} + \hat{Q}(A - \Sigma P)^* + Q\bar{\Sigma}Q]\tau^*, \tag{3.17}$$

$$0 = [(A - Q\bar{\Sigma})^* \hat{P} + \hat{P}(A - Q\bar{\Sigma}) + P\Sigma P]\tau. \tag{3.18}$$

The content of the Main Theorem is clearly a set of necessary conditions which characterize the optimal steady-state fixed-order dynamic compensator when it exists. These necessary conditions consist of a system of four operator equations including a pair of modified Riccati equations (3.15) and (3.16) and a pair of modified Lyapunov equations (3.17) and (3.18). The salient feature of these four equations is the coupling by the operator $\tau \in \mathscr{B}(\mathscr{H})$ which, because of (3.8), is idempotent, i.e., $\tau^2 = \tau$. In general, τ is an *oblique* projection and not an orthogonal projection since there is no requirement that τ be selfadjoint. Additional features of the Main Theorem will be discussed in the remainder of this section. For convenience, let G, M, Γ, τ, Q, P, $\hat{Q}$ and $\hat{P}$ be as given by the Main Theorem and define $\Lambda \triangleq \operatorname{diag}(\lambda_1, \cdots, \lambda_{n_c})$, where $\lambda_1 \geqq \lambda_2 \geqq \cdots \geqq \lambda_{n_c} > 0$ are the eigenvalues of M.

We begin by noting that if x_c is replaced by Sx_c, where $S \in \mathbb{R}^{n_c \times n_c}$ is invertible, then an "equivalent" compensator is obtained with (A_c, B_c, C_c) replaced by $(SA_cS^{-1}, SB_c, C_cS^{-1})$.

PROPOSITION 3.1. *Let* $(A_c, B_c, C_c) \in \mathscr{A}_+$. *If* $S \in \mathbb{R}^{n_c \times n_c}$ *is invertible then* $(SA_cS^{-1}, SB_c, C_cS^{-1}) \in \mathscr{A}_+$ *and*

$$J(A_c, B_c, C_c) = J(SA_cS^{-1}, SB_c, C_cS^{-1}). \tag{3.19}$$

Proof. Although the result is obvious from system-theoretic arguments, we shall prove it analytically by utilizing elements of the development in § 4. Define

$$\tilde{S} \triangleq \begin{bmatrix} I_\infty & 0 \\ 0 & S \end{bmatrix} \in \mathscr{B}(\tilde{\mathscr{H}})$$

and note that replacing (A_c, B_c, C_c) by $(SA_cS^{-1}, SB_c, C_cS^{-1})$ is equivalent to replacing $\tilde{A}$, $\tilde{V}$ and $\tilde{R}$ by $\tilde{S}\tilde{A}\tilde{S}^{-1}$, $\tilde{S}\tilde{V}\tilde{S}^*$ and $\tilde{S}^{-*}\tilde{R}\tilde{S}^{-1}$, respectively. If $\alpha, \beta > 0$ satisfy (3.6) then a straightforward application of the Hille-Yosida theorem [57, pp. 153-5] shows that

[5] (3.14a) refers to $\rho(\hat{Q}) = n_c$, etc.

the strongly continuous semigroup generated by $\tilde{S}\tilde{A}\tilde{S}^{-1}$ satisfies $\|e^{\tilde{S}\tilde{A}\tilde{S}^{-1}t}\| \leqq \|\tilde{S}\|\,\|\tilde{S}^{-1}\|\alpha\, e^{-\beta t}$, which proves the first assertion. Since $\tilde{S}\,e^{\tilde{A}t}\tilde{S}^{-1}$, $t \geqq 0$, is also a strongly continuous semigroup with generator $\tilde{S}\tilde{A}\tilde{S}^{-1}$, it follows that $\tilde{S}\,e^{\tilde{A}t}\tilde{S}^{-1} = e^{\tilde{S}\tilde{A}\tilde{S}^{-1}t}$. Hence

$$\int_0^\infty e^{\tilde{S}\tilde{A}\tilde{S}^{-1}t}(\tilde{S}\tilde{V}\tilde{S}^*)\, e^{(\tilde{S}\tilde{A}\tilde{S}^{-1})^*t}\, dt = \tilde{S}\tilde{Q}\tilde{S}^*$$

and (3.19) follows from $\operatorname{tr}\tilde{Q}\tilde{R} = \operatorname{tr}(\tilde{S}\tilde{Q}\tilde{S}^*)(\tilde{S}^{-*}\tilde{R}\tilde{S}^{-1})$. □

In view of Proposition 3.1 one would expect the Main Theorem to apply also to $(SA_cS^{-1}, SB_c, C_cS^{-1})$. Indeed, it may be noted that no claim was made as to the uniqueness of the (G, M, Γ)-factorization of $\hat{Q}\hat{P}$ used to determine A_c, B_c and C_c in (3.9)-(3.11). These observations are reconciled by the following result which shows that a transformation of the compensator state basis corresponds to the alternative factorization $\hat{Q}\hat{P} = (S^{-T}G)^T(SMS^{-1})(S\Gamma)$ and, moreover, that all (G, M, Γ)-factorizations of $\hat{Q}\hat{P}$ are related by a nonsingular transformation. Note that τ remains invariant over the class of factorizations.

PROPOSITION 3.2. *If $S \in \mathbb{R}^{n_c \times n_c}$ is invertible then $\bar{G} \triangleq S^{-T}G$, $\bar{\Gamma} \triangleq S\Gamma$ and $\bar{M} \triangleq SMS^{-1}$ satisfy*

$$\hat{Q}\hat{P} = \bar{G}^*\bar{M}\bar{\Gamma}, \tag{3.7)'}$$

$$\bar{\Gamma}\bar{G}^* = I_{n_c}. \tag{3.8)'}$$

Conversely, if $\bar{G}, \bar{\Gamma} \in \mathcal{B}(\mathcal{H}, \mathbb{R}^{n_c})$ and invertible $\bar{M} \in \mathbb{R}^{n_c \times n_c}$ satisfy (3.7)′ *and* (3.8)′, *then there exists invertible $S \in \mathbb{R}^{n_c \times n_c}$ such that $\bar{G} = S^{-T}G$, $\bar{\Gamma} = S\Gamma$ and $\bar{M} = SMS^{-1}$.*

Proof. The first part of the proposition is immediate. The second part follows by taking $S \triangleq \bar{M}^{-1}\bar{\Gamma}G^*M^{-1}$, noting $S^{-1} = M\bar{\Gamma}\bar{G}^*\bar{M}^{-1}$ and using the identities $\bar{\Gamma}G^*M\Gamma\bar{G}^* = \bar{M}$ and $M\Gamma\bar{G}^* = \Gamma\bar{G}^*\bar{M}$. □

The next result shows that there exists a similarity transformation which simultaneously diagonalizes $\hat{Q}\hat{P}$ and τ.

PROPOSITION 3.3. *There exists invertible $\Phi \in \mathcal{B}(\mathcal{H})$ such that*

$$\hat{Q} = \Phi^{-1}\begin{bmatrix} \Lambda_{\hat{Q}} & 0 \\ 0 & 0 \end{bmatrix}\Phi^{-*}, \qquad \hat{P} = \Phi^*\begin{bmatrix} \Lambda_{\hat{P}} & 0 \\ 0 & 0 \end{bmatrix}\Phi, \tag{3.20a, b}$$

$$\hat{Q}\hat{P} = \Phi^{-1}\begin{bmatrix} \Lambda & 0 \\ 0 & 0 \end{bmatrix}\Phi, \qquad \tau = \Phi^{-1}\begin{bmatrix} I_{n_c} & 0 \\ 0 & 0 \end{bmatrix}\Phi, \tag{3.21a, b}$$

where $\Lambda_{\hat{Q}}, \Lambda_{\hat{P}} \in \mathbb{R}^{n_c \times n_c}$ are positive diagonal and $\Lambda_{\hat{Q}}\Lambda_{\hat{P}} = \Lambda$. Consequently,

$$\hat{Q} = \tau\hat{Q}, \qquad \hat{P} = \hat{P}\tau. \tag{3.22a, b}$$

Proof. Proceeding as in the proof of Lemma 3.1, choose an orthonormal basis for $\mathcal{H}$ with respect to which

$$\hat{Q} = \begin{bmatrix} \hat{Q}_1 & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad \hat{P} = \begin{bmatrix} \hat{P}_1 & 0 \\ 0 & 0 \end{bmatrix},$$

where $\hat{Q}_1, \hat{P}_1 \in \mathbb{R}^{r \times r}$ are nonnegative definite. By [55, Thm. 6.2.5, p. 123], there exists invertible $\Psi \in \mathbb{R}^{r \times r}$ such that $\tilde{\Lambda}_{\hat{Q}} \triangleq \Psi\hat{Q}_1\Psi^T$ and $\tilde{\Lambda}_{\hat{P}} = \Psi^{-T}\hat{P}_1\Psi^{-1}$ are nonnegative diagonal. Because of (3.14), it is clear that Ψ can be chosen so that

$$\tilde{\Lambda}_{\hat{Q}} = \begin{bmatrix} \Lambda_{\hat{Q}} & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad \tilde{\Lambda}_{\hat{P}} = \begin{bmatrix} \Lambda_{\hat{P}} & 0 \\ 0 & 0 \end{bmatrix},$$

where $\Lambda_{\hat{Q}}, \Lambda_{\hat{P}} \in \mathbb{R}^{n_c \times n_c}$ are positive diagonal. Thus (3.20) holds with

$$\Phi \triangleq \begin{bmatrix} \Psi & 0 \\ 0 & I_\infty \end{bmatrix}.$$

From (3.20) it follows that

$$\hat{Q}\hat{P} = \Phi^{-1} \begin{bmatrix} \Lambda_Q \Lambda_P & 0 \\ 0 & 0 \end{bmatrix} \Phi.$$

Now define $\bar{G} = [I_{n_c} \ 0]\Phi^{-*}$, $\bar{M} = \Lambda_{\hat{Q}}\Lambda_{\hat{P}}$ and $\bar{\Gamma} = [I_{n_c} \ 0]\Phi$ so that (3.7)′ and (3.8)′ are satisfied. By the second part of Proposition 3.2 there exists invertible $S \in \mathbb{R}^{n_c \times n_c}$ such that $G = S^T\bar{G}$, $M = S^{-1}\bar{M}S$ and $\Gamma = S^{-1}\bar{\Gamma}$. Since M and $\bar{M}$ have the same eigenvalues, $\bar{M} = \Lambda$ (modulo an ordering of the diagonal elements) and thus (3.21a) holds. Finally, (3.21b) follows from

$$\tau = G^*\Gamma = \bar{G}^*\bar{\Gamma} = \Phi^{-1} \begin{bmatrix} I_{n_c} & 0 \\ 0 & 0 \end{bmatrix} \Phi. \qquad \square$$

Remark 3.1. Proposition 3.3 shows that $\lambda_1, \cdots, \lambda_{n_c}$ are the positive eigenvalues of $\hat{Q}\hat{P}$.

Remark 3.2. The simultaneous diagonalization in (3.20) has been effected by a *contragredient transformation* [55], [58]. For applications of this type of transformation to model reduction and realization problems see [12], [59]–[61]. Simultaneous diagonalization of operators is discussed in [53, p. 181].

The following result validates the precise handling of the unbounded operator A in (3.9), (3.17) and (3.18).

PROPOSITION 3.4. *The following relations hold*:

(3.23a, b, c) $$\rho(G) = \rho(\Gamma) = \rho(\tau) = n_c,$$

(3.24a, b) $$\tau: \mathscr{H} \to \mathscr{D}(A), \qquad \tau^*: \mathscr{H} \to \mathscr{D}(A^*),$$

(3.25a, b) $$G^*: \mathbb{R}^{n_c} \to \mathscr{D}(A), \qquad \Gamma^*: \mathbb{R}^{n_c} \to \mathscr{D}(A^*).$$

Proof. From (3.8) and (2.1) it follows that $n_c = \rho(\Gamma G^*) \leqq \min\{\rho(\Gamma), \rho(G^*)\}$. Since $\rho(\Gamma) \leqq n_c$, $\rho(G) = \rho(G^*)$ and $\rho(G) \leqq n_c$, (3.23a) and (3.23b) hold. To show (3.23c) either note (3.21b) or use (3.14a) and (3.22) to obtain

$$n_c = \rho(\hat{Q}) = \rho(\tau\hat{Q}) \leqq \rho(\tau) = \rho(G^*\Gamma) \leqq \rho(\Gamma) = n_c.$$

To prove (3.24a) note that (3.22a) implies $\mathscr{R}(\hat{Q}) \subset \mathscr{R}(\tau)$ and thus $\rho(\hat{Q}) = \rho(\tau)$ implies $\mathscr{R}(\hat{Q}) = \mathscr{R}(\tau)$, and similarly for (3.24b). Finally, (3.25) follows from (3.23), (3.24), the definition $\tau = G^*\Gamma$ and the fact that $\tau^* = \Gamma^*G$. $\square$

Since the domain of A may not be all of $\mathscr{H}$, expressions involving A require special interpretation. First note that because of the range condition (3.25a), the expression (3.9) indeed represents an $n_c \times n_c$ matrix (see, e.g., [6, p. 80]). Similarly, because of (3.25b), A_c^T is given by

(3.26) $$A_c^T = G(A^* - \bar{\Sigma}Q - P\Sigma)\Gamma^*.$$

With regard to (3.15), note that because of (3.12a), the right-hand side of (3.15) is a linear operator with domain $\mathscr{D}(A^*)$. Since $\Theta \triangleq -\tau Q\bar{\Sigma}Q - Q\bar{\Sigma}Q\tau^* + V_1 + \tau Q\bar{\Sigma}Q\tau^*$ is continuous on $\mathscr{D}(A^*)$, $AQ + QA^*$ has a continuous extension on $\mathscr{H}$ given precisely by $-\Theta$. Similar remarks apply to (3.16). Analogous domain conditions were obtained in [5] for a deterministic infinite-dimensional linear-quadratic control problem with

full-state feedback. Finally, because of (3.24) the right-hand sides of (3.17) and (3.18) denote bounded linear operators on all of $\mathscr{H}$.

It is useful to present an alternative form of the optimal projection equations (3.15)-(3.18). For convenience define the notation

$$\tau_\perp \triangleq I_{\mathscr{H}} - \tau.$$

PROPOSITION 3.5. *Equations* (3.15)-(3.18) *are equivalent, respectively, to*

$$0 = AQ + QA^* + V_1 - Q\bar{\Sigma}Q + \tau_\perp Q\bar{\Sigma}Q\tau_\perp^*, \tag{3.27}$$

$$0 = A^*P + PA + R_1 - P\Sigma P + \tau_\perp^* P\Sigma P\tau_\perp, \tag{3.28}$$

$$0 = (A - \Sigma P)\hat{Q} + \hat{Q}(A - \Sigma P)^* + Q\bar{\Sigma}Q - \tau_\perp Q\bar{\Sigma}Q\tau_\perp^*, \tag{3.29}$$

$$0 = (A - Q\bar{\Sigma})^*\hat{P} + \hat{P}(A - Q\bar{\Sigma}) + P\Sigma P - \tau_\perp^* P\Sigma P\tau_\perp. \tag{3.30}$$

Proof. The equivalence of (3.27) and (3.28) to (3.15) and (3.16) is immediate. Using (3.22a) in the form $\hat{Q} = \hat{Q}\tau^*$, we obtain (3.17) = (3.29)τ^*. Conversely, from (3.22a) and $[(A - \Sigma P)\hat{Q}]^* = \hat{Q}(A - \Sigma P)^*$ (see, e.g., [6, p. 80]) it follows that (3.29) = (3.17) + (3.17)* − τ(3.17). Similarly, (3.18) and (3.30) are equivalent. □

The form of the optimal projection equations (3.27)-(3.30) helps demonstrate the relationship between the Main Theorem and the classical *LQG* result when $\dim \mathscr{H} = n < \infty$. In this case we need only note that the (G, M, Γ)-factorization of $\hat{Q}\hat{P}$ in the "full-order" case $n_c = n$ is given by $G = \Gamma = I_n$ and $M = \hat{Q}\hat{P}$. Since $\tau = I_n$, and thus $\tau_\perp = 0$, (3.27) and (3.28) reduce to the standard observer and regulator Riccati equations and (3.9)-(3.11) yield the usual *LQG* expressions. Furthermore, note that in the full-order case

$$A_c = A + BC_c - B_cC \tag{3.31}$$

and (3.29) and (3.31) can be written as

$$0 = (A_c + B_cC)\hat{Q} + \hat{Q}(A_c + B_cC)^T + B_cV_2B_c^T, \tag{3.32}$$

$$0 = (A_c - BC_c)^T\hat{P} + \hat{P}(A_c - BC_c) + C_c^TR_2C_c. \tag{3.33}$$

Since, as is well known, the stability of $\tilde{A}$ corresponds to the stability of $A + BC_c = A_c + B_cC$ and $A - B_cC = A_c - BC_c$, it follows from standard results (e.g., [62, pp. 48, 277]) that the positive-definiteness conditions (3.14a, b) are equivalent to the assumption that (A_c, B_c, C_c) is controllable and observable.

To obtain a geometric intepretation of the optimal projection we introduce the quasi-full-state estimate

$$\hat{x}(t) \triangleq G^*x_c(t) \in \mathscr{H}$$

so that $\tau\hat{x}(t) = \hat{x}(t)$ and $x_c(t) = \Gamma\hat{x}(t)$. Now, the closed-loop system (3.1)-(3.4) can be written as

$$\dot{x}(t) = Ax(t) - B\hat{C}_c\tau\hat{x}(t) + H_1w(t), \tag{3.34}$$

$$\dot{\hat{x}}(t) = \tau(A + B\hat{C}_c - \hat{B}_cC)\tau\hat{x}(t) + \tau\hat{B}_c(Cx(t) + H_2w(t)), \tag{3.35}$$

where (3.35) is interpreted in the sense of (3.34) since $\hat{x}(t) \in \mathscr{H}$ and where

$$\hat{B}_c \triangleq QC^*V_2^{-1}, \qquad \hat{C}_c \triangleq -R_2^{-1}B^*P.$$

It can thus be seen that the geometric structure of the quasi-full-order compensator is entirely dictated by the projection τ. In particular, control inputs $\tau\hat{x}(t)$ determined by

(3.35) are contained in $\mathcal{R}(\tau)$ and sensor inputs $\tau\hat{B}_c y(t)$ are annihilated unless they are contained in $[\mathcal{N}(\tau)]^\perp = \mathcal{R}(\tau^*)$. Consequently, $\mathcal{R}(\tau)$ and $\mathcal{R}(\tau^*)$ are the control and observation subspaces, respectively, of the compensator. Since τ is not necessarily an orthogonal projection, these (finite-dimensional) subspaces may be different.

From the form of (3.35) it is tempting to suggest that the optimal fixed-order dynamic compensator can be obtained by projecting the full-order (infinite-dimensional) *LQG* compensator. However, this is generally impossible for the following simple reason. Although the expressions for A_c, B_c and C_c in (3.9)-(3.11) have the *form* of a projection of the full-order *LQG* compensator, the operators Q and P in (3.9)-(3.11) are *not* the solutions of the usual *LQG* Riccati equations but instead must be obtained by simultaneously solving all four coupled equations (3.15)-(3.18). This observation reinforces the statement made in § 1 that the optimal fixed-order dynamic compensator cannot in general be obtained by *LQG* followed by closed-loop controller reduction as in [14] and [15].

We now give an explicit characterization of the optimal projection in terms of $\hat{Q}$ and $\hat{P}$. Since $\hat{Q}\hat{P}$ has finite rank, its Drazin inverse $(\hat{Q}\hat{P})^D$ exists (see [63, Thm. 6, p. 108]) and, since $(\hat{Q}\hat{P})^2 = G^*M^2\Gamma$, and hence $\rho(\hat{Q}\hat{P})^2 = \rho(\hat{Q}\hat{P})$, the "index" of $\hat{Q}\hat{P}$ (see [63], [64]) is 1. In this case the Drazin inverse is traditionally called the group inverse and is denoted by $(\hat{Q}\hat{P})^\#$ (see, e.g., [64, p. 124] or [65]).

PROPOSITION 3.6. *The optimal projection τ is given by*

$$\tau = \hat{Q}\hat{P}(\hat{Q}\hat{P})^\#. \tag{3.36}$$

Proof. It is easy to verify that the conditions characterizing the Drazin inverse [63] for the case that $\hat{Q}\hat{P}$ has index 1 are satisfied by $G^*M^{-1}\Gamma$. Hence $(\hat{Q}\hat{P})^\# = G^*M^{-1}\Gamma$ and (3.8) implies (3.36). □

We now give an alternative characterization of the optimal projection by introducing the following notation from [51, p. 73]. For $\phi, \psi \in \mathcal{H}$ define the operator $\phi \otimes \psi \in \mathcal{B}(\mathcal{H})$ by

$$(\phi \otimes \psi)x \triangleq \langle x, \phi\rangle\psi, \qquad x \in \mathcal{H},$$

and note that $\rho(\phi \otimes \psi) = 1$ if ϕ and ψ are both nonzero and $(\phi \otimes \psi)^* = \psi \otimes \phi$. Using this notation, (3.21a) can be written as

$$\Phi\hat{Q}\hat{P}\Phi^{-1} = \sum_{i=1}^{n_c} \lambda_i \xi_i \otimes \xi_i, \tag{3.37}$$

where $\{\xi_i\}_{i=1}^\infty$ is an orthonormal basis for $\mathcal{H}$. In terms of the Riesz bases (see e.g., [52, p. 309])

$$\phi_i \triangleq \Phi^*\xi_i, \qquad \psi_i \triangleq \Phi^{-1}\xi_i, \qquad i = 1, 2, \cdots,$$

(3.37) is equivalent to

$$\hat{Q}\hat{P} = \sum_{i=1}^{n_c} \lambda_i \phi_i \otimes \psi_i, \tag{3.38}$$

which can be regarded as a specialized spectral decomposition of a semisimple operator. We emphasize that, in contrast to the singular value decomposition for compact nonnormal operators (see, e.g., [50, p. 261]), the λ_i in (3.38) are *eigenvalues* of $\hat{Q}\hat{P}$ (see Remark 3.1), not singular values. Moreover, although $\{\phi_i\}_{i=1}^\infty$ and $\{\psi_i\}_{i=1}^\infty$ are bases for $\mathcal{H}$, they are not necessarily orthogonal. They are, however, biorthonormal, i.e., $\langle \phi_i, \psi_j\rangle = \delta_{ij}$, and hence $\phi_i \otimes \psi_i$ is a rank-one projection and $(\phi_i \otimes \psi_i)(\phi_j \otimes \psi_j) = 0$, $i \neq j$.

Since τ is a rank-n_c projection, it is not surprising that τ is given precisely by

$$\tau = \sum_{i=1}^{n_c} \phi_i \otimes \psi_i. \tag{3.39}$$

The following result summarizes the above observations.

PROPOSITION 3.7. *There exist biorthonormal linearly independent sets* $\{\psi_i\}_{i=1}^{n_c} \subset \mathscr{D}(A)$ *and* $\{\phi_i\}_{i=1}^{n_c} \subset \mathscr{D}(A^*)$ *such that* (3.38) *and* (3.39) *hold. Furthermore, if the* (G, M, Γ)*-factorization of* $\hat{Q}\hat{P}$ *is chosen such that* $M = \Lambda$, *then, for all* $x \in \mathscr{H}$,

$$Gx = (\langle x, \psi_1 \rangle, \cdots, \langle x, \psi_{n_c} \rangle)^T,$$

$$\Gamma x = (\langle x, \phi_1 \rangle, \cdots, \langle x, \phi_{n_c} \rangle)^T.$$

Remark 3.3. Note that $\hat{P}\hat{Q}$ and τ^* are given by

$$\hat{P}\hat{Q} = \sum_{i=1}^{n_c} \lambda_i \psi_i \otimes \phi_i, \qquad \tau^* = \sum_{i=1}^{n_c} \psi_i \otimes \phi_i,$$

and, for all $y \triangleq (y_1, \cdots, y_{n_c})^T \in \mathbb{R}^{n_c}$, G^* and Γ^* satisfy

$$G^* y = \sum_{i=1}^{n_c} y_i \psi_i, \qquad \Gamma^* y = \sum_{i=1}^{n_c} y_i \phi_i.$$

4. Proof of the Main Theorem. We state and prove a series of lemmas which allow us to compute the Frechet derivatives of J with respect to A_c, B_c and C_c. Requiring that these derivatives vanish leads to the necessary conditions in their "primitive" form. A transformation of variables then leads to the form of the necessary conditions (3.9)-(3.18).

Let "u-lim" denote the uniform limit (i.e., limit in operator norm) for bounded linear operators [50, p. 150] and, for strongly continuous $S(t) \in \mathscr{B}(\mathscr{H})$, $t \geqq 0$, interpret the strong integral $\int_{t_1}^{t_2} S(t)\, dt$ according to $\int_{t_1}^{t_2} S(t) z\, dt$, $z \in \mathscr{H}$ [50, p. 152]. Also recall the standard fact [6, p. 186] that $(e^{At})^* = e^{A^* t}$ and similarly for $\tilde{A}$. Throughout this section let $(A_c, B_c, C_c) \in \mathscr{A}_+$ and let $\alpha, \beta > 0$ satisfy (3.6).

To begin, note that the closed-loop system (3.1)-(3.4) can be written as

$$\dot{\tilde{x}}(t) = \tilde{A}\tilde{x}(t) + \tilde{H}w(t), \tag{4.1}$$

where

$$\tilde{H} \triangleq \begin{bmatrix} H_1 \\ B_c H_2 \end{bmatrix} \in \mathscr{B}_2(\mathscr{H}' \oplus \mathbb{R}^l).$$

For convenience define the nonnegative-definite operator

$$\tilde{V} \triangleq \tilde{H}\tilde{H}^* = \begin{bmatrix} V_1 & 0 \\ 0 & B_c V_2 B_c^T \end{bmatrix} \in \mathscr{B}_1(\tilde{\mathscr{H}}).$$

In terms of the augmented state $\tilde{x}(t)$, the performance criterion (3.5) becomes

$$J(A_c, B_c, C_c) = \lim_{t \to \infty} \mathbb{E}\langle \tilde{R}\tilde{x}(t), \tilde{x}(t) \rangle, \tag{4.2}$$

where the nonnegative-definite operator $\tilde{R}$ is defined by

$$\tilde{R} \triangleq \begin{bmatrix} R_1 & 0 \\ 0 & C_c^T R_2 C_c \end{bmatrix} \in \mathscr{B}_1(\tilde{\mathscr{H}}).$$

To write (4.2) in terms of the covariance of $\tilde{x}(t)$, recall [6, p. 308] that the covariance "$\mathbb{E}[(\xi-\mathbb{E}\xi)(\xi-\mathbb{E}\xi)^*]$" of a Hilbert-space-valued weak random variable ξ is defined to be the nonnegative-definite operator S which satisfies

$$\langle Sy, z\rangle = \mathbb{E}\langle \xi-\mathbb{E}\xi, y\rangle\langle \xi-\mathbb{E}\xi, z\rangle$$

for all y, z in the Hilbert space. Hence define [6, p. 317]

$$\tilde{Q}(t) \triangleq \mathbb{E}[(\tilde{x}(t)-\mathbb{E}\tilde{x}(t))(\tilde{x}(t)-\mathbb{E}\tilde{x}(t))^*].$$

LEMMA 4.1. *$\tilde{Q} \triangleq u\text{-}\lim_{t\to\infty} \tilde{Q}(t)$ exists and is given by*

$$\tilde{Q} = \int_0^\infty e^{\tilde{A}t}\tilde{V}e^{\tilde{A}^*t}\,dt. \tag{4.3}$$

Furthermore,

$$J(A_c, B_c, C_c) = \operatorname{tr}\tilde{Q}\tilde{R}. \tag{4.4}$$

Proof. First compute (as in [6, p. 317])

$$\begin{aligned}
\langle \tilde{Q}(t)\tilde{y}, \tilde{z}\rangle &= \mathbb{E}\langle \tilde{x}(t) - e^{\tilde{A}t}\mathbb{E}\tilde{x}(0), \tilde{y}\rangle\langle \tilde{x}(t) - e^{\tilde{A}t}\mathbb{E}\tilde{x}(0), \tilde{z}\rangle \\
&= \mathbb{E}\left\langle \int_0^t e^{\tilde{A}(t-s)}\tilde{H}\tilde{w}(s)\,ds, \tilde{y}\right\rangle\left\langle \int_0^t e^{\tilde{A}(t-\sigma)}\tilde{H}\tilde{w}(\sigma)\,d\sigma, \tilde{z}\right\rangle \\
&\quad + \langle \tilde{Q}(0)\, e^{\tilde{A}^*t}\tilde{y}, e^{\tilde{A}^*t}\tilde{z}\rangle \\
&= \mathbb{E}\int_0^t\int_0^t \langle \tilde{w}(s), \tilde{H}^*\, e^{\tilde{A}^*(t-s)}\tilde{y}\rangle\langle \tilde{w}(\sigma), \tilde{H}^*\, e^{\tilde{A}^*(t-\sigma)}\tilde{z}\rangle\,ds\,d\sigma \\
&\quad + \langle e^{\tilde{A}t}\tilde{Q}(0)\, e^{\tilde{A}^*t}\tilde{y}, \tilde{z}\rangle \\
&= \int_0^t \langle e^{\tilde{A}(t-s)}\tilde{V}e^{\tilde{A}^*(t-s)}\tilde{y}, \tilde{z}\rangle\,ds + \langle e^{\tilde{A}t}\tilde{Q}(0)\, e^{\tilde{A}^*t}\tilde{y}, \tilde{z}\rangle,
\end{aligned}$$

which shows that $\tilde{Q}(t)$ is given by

$$\tilde{Q}(t) = e^{\tilde{A}t}\tilde{Q}(0)\, e^{\tilde{A}^*t} + \int_0^t e^{\tilde{A}s}\tilde{V}e^{\tilde{A}^*s}\,ds.$$

Clearly, (4.3) makes sense as a strong integral since

$$\|\tilde{Q}\| \leqq \int_0^\infty \|e^{\tilde{A}t}\tilde{V}e^{\tilde{A}^*t}\|\,dt \leqq \alpha^2\|\tilde{V}\|\int_0^\infty e^{-2\beta t}\,dt < \infty.$$

To demonstrate uniform convergence it need only be noted that

$$\begin{aligned}
\|\tilde{Q} - \tilde{Q}(t)\| &= \sup_{\|\tilde{y}\|=1}\|(\tilde{Q}-\tilde{Q}(t))\tilde{y}\| \\
&= \sup_{\|\tilde{y}\|=1}\left\|\int_t^\infty e^{\tilde{A}s}\tilde{V}e^{\tilde{A}^*s}\tilde{y}\,ds - e^{\tilde{A}t}\tilde{Q}(0)\, e^{\tilde{A}^*t}\tilde{y}\right\| \\
&\leqq \int_t^\infty \|e^{\tilde{A}s}\tilde{V}e^{\tilde{A}^*s}\|\,ds + \|e^{\tilde{A}t}\tilde{Q}(0)\, e^{\tilde{A}^*t}\| \\
&\leqq \tfrac{1}{2}\alpha^2\|\tilde{V}\|\beta^{-1}\, e^{-2\beta t} + \|\tilde{Q}(0)\|\, e^{-2\beta t}.
\end{aligned}$$

Next, let $\{\phi_i\}_{i=1}^{\infty}$ be an orthonormal basis for $\tilde{\mathscr{H}}$ and use Parseval's equality to obtain

$$J(A_c, B_c, C_c) = \lim_{t\to\infty} \mathbb{E}\|\tilde{R}^{1/2}\tilde{x}(t)\|^2 = \lim_{t\to\infty} \mathbb{E} \sum_{i=1}^{\infty} \langle \tilde{R}^{1/2}\tilde{x}(t), \phi_i\rangle^2.$$

Since

$$f_n(t) \triangleq \sum_{i=1}^{n} \langle \tilde{R}^{1/2}\tilde{x}(t), \phi_i\rangle^2, \qquad t \geqq 0,$$

is nonnegative for each n and is increasing in n for each t with limit $\langle \tilde{R}\tilde{x}(t), \tilde{x}(t)\rangle$, monotone convergence permits expectation-limit interchange. Hence using $\mathbb{E}\tilde{x}(t) = e^{\tilde{A}t}\mathbb{E}\tilde{x}(0)$ we have

$$\begin{aligned} J(A_c, B_c, C_c) &= \lim_{t\to\infty} \sum_{i=1}^{\infty} \mathbb{E}\langle \tilde{x}(t), \tilde{R}^{1/2}\phi_i\rangle^2 \\ &= \lim_{t\to\infty} \sum_{i=1}^{\infty} [\langle \tilde{Q}(t)\tilde{R}^{1/2}\phi_i, \tilde{R}^{1/2}\phi_i\rangle + \langle e^{\tilde{A}t}\mathbb{E}\tilde{x}(0), \tilde{R}^{1/2}\phi_i\rangle^2] \\ &= \lim_{t\to\infty} \{\operatorname{tr}[\tilde{R}^{1/2}\tilde{Q}(t)\tilde{R}^{1/2}] + \|\tilde{R}^{1/2} e^{\tilde{A}t}\mathbb{E}\tilde{x}(0)\|^2\} \end{aligned}$$

which by Corollary 2.1 yields (4.4). □

We shall also require the "dual" of $\tilde{Q}$ given by

$$\tilde{P} = \int_0^{\infty} e^{\tilde{A}^* t}\tilde{R}\, e^{\tilde{A}t}\, dt. \tag{4.5}$$

Since $\tilde{V}$ and $\tilde{R}$ are nonnegative definite it is readily seen that $\tilde{Q}$ and $\tilde{P}$ are also nonnegative definite.

LEMMA 4.2. $\tilde{Q}, \tilde{P} \in \mathscr{B}_1(\tilde{\mathscr{H}})$.

Proof. It suffices to consider $\tilde{Q}$ only since the situation for $\tilde{P}$ is exactly analogous. Since $\tilde{Q}$ is nonnegative definite, Lemma 2.3 can be used. Letting $\{\phi_i\}_{i=1}^{\infty}$ be an orthonormal basis for $\tilde{\mathscr{H}}$, we have

$$\begin{aligned} \operatorname{tr}\tilde{Q} &= \sum_{i=1}^{\infty} \langle \tilde{Q}\phi_i, \phi_i\rangle = \sum_{i=1}^{\infty} \left\langle \int_0^{\infty} e^{\tilde{A}t}\tilde{V}e^{\tilde{A}^* t}\phi_i\, dt, \phi_i \right\rangle \\ &= \lim_{n\to\infty} \int_0^{\infty} \sum_{i=1}^{n} \langle \tilde{V}e^{\tilde{A}^* t}\phi_i, e^{\tilde{A}^* t}\phi_i\rangle\, dt. \end{aligned}$$

Let $f_n(t)$ denote the above integrand. Since $\tilde{V}$ is nonnegative definite, $\{f_n(\cdot)\}$ is a monotonically increasing sequence of nonnegative functions such that $f_n(t) \to \operatorname{tr} e^{\tilde{A}t}\tilde{V}e^{\tilde{A}^* t}$, $t \geqq 0$. Hence, by monotone convergence and Lemma 2.2,

$$\begin{aligned} \operatorname{tr}\tilde{Q} &= \int_0^{\infty} \operatorname{tr}[e^{\tilde{A}t}\tilde{V}e^{\tilde{A}^* t}]\, dt \\ &= \int_0^{\infty} \|e^{\tilde{A}t}\tilde{V}e^{\tilde{A}^* t}\|_1\, dt \leqq \alpha^2\|\tilde{V}\|_1 \int_0^{\infty} e^{-2\beta t}\, dt < \infty. \end{aligned}$$

□

LEMMA 4.3. *With $\tilde{Q}$ and $\tilde{P}$ given by* (4.3) *and* (4.5) *it follows that*

$$\operatorname{tr}\tilde{Q}\tilde{R} = \operatorname{tr}\tilde{V}\tilde{P}. \tag{4.6}$$

Proof. For any orthonormal basis $\{\phi_i\}_{i=1}^{\infty}$ of $\tilde{\mathscr{H}}$ we have

$$\operatorname{tr} \tilde{Q}\tilde{R} = \operatorname{tr} \tilde{R}\tilde{Q} = \sum_{i=1}^{\infty} \left\langle \tilde{R} \int_0^{\infty} e^{\tilde{A}t} \tilde{V} e^{\tilde{A}^* t} \phi_i \, dt, \phi_i \right\rangle$$

$$= \lim_{n\to\infty} \int_0^{\infty} \sum_{i=1}^{n} \langle \tilde{R} e^{\tilde{A}t} \tilde{V} e^{\tilde{A}^* t} \phi_i, \phi_i \rangle \, dt.$$

Letting $f_n(t)$ denote the above integrand it follows that $f_n(t) \to \operatorname{tr} \tilde{R} e^{\tilde{A}t} \tilde{V} e^{\tilde{A}^* t}$, $t \geqq 0$, and

$$|f_n(t)| \leqq \sum_{i=1}^{\infty} |\langle e^{\tilde{A}t} \tilde{V} e^{\tilde{A}^* t} \phi_i, \tilde{R}\phi_i \rangle| \leqq \alpha^2 \|\tilde{V}\| \, e^{-2\beta t} \sum_{i=1}^{\infty} \|\tilde{R}\phi_i\|.$$

If $\{\phi_i\}_{i=1}^{\infty}$ is chosen to be the set of orthonormal eigenvectors of $\tilde{R}$ then Lemma 2.1 implies $\sum_{i=1}^{\infty} \|\tilde{R}\phi_i\| = \|\tilde{R}\|_1$ and thus $|f_n(t)|$ is bounded on $[0, \infty)$ by an integrable function. Hence by dominated convergence,

$$\operatorname{tr} \tilde{Q}\tilde{R} = \int_0^{\infty} \operatorname{tr} [\tilde{R} e^{\tilde{A}t} \tilde{V} e^{\tilde{A}^* t}] \, dt = \int_0^{\infty} \operatorname{tr} [e^{\tilde{A}^* t} \tilde{R} e^{\tilde{A}t} \tilde{V}] \, dt = \int_0^{\infty} \sum_{i=1}^{\infty} \langle \tilde{V}\phi_i, e^{\tilde{A}^* t} \tilde{R} e^{\tilde{A}t} \rangle \, dt.$$

And again using dominated convergence,

$$\operatorname{tr} \tilde{Q}\tilde{R} = \sum_{i=1}^{\infty} \int_0^{\infty} \langle \tilde{V}\phi_i, e^{\tilde{A}^* t} \tilde{R} e^{\tilde{A}t} \phi_i \rangle \, dt = \sum_{i=1}^{\infty} \left\langle \tilde{V}\phi_i, \int_0^{\infty} e^{\tilde{A}^* t} \tilde{R} e^{\tilde{A}t} \phi_i \, dt \right\rangle = \operatorname{tr} \tilde{V}\tilde{P}. \qquad \square$$

The next result is important in that it allows us to treat $\tilde{Q}$ and $\tilde{P}$ as solutions of dual algebraic Lyapunov equations. For a similar result involving groups rather than semigroups see [50, pp. 555–557].

LEMMA 4.4. *$\tilde{Q}$ is given by* (4.3) *if and only if $\tilde{Q} \in \mathscr{B}(\tilde{\mathscr{H}})$ satisfies*

$$\tilde{Q}\colon \mathscr{D}(\tilde{A}^*) \to \mathscr{D}(\tilde{A}), \tag{4.7}$$

$$0 = \tilde{A}\tilde{Q} + \tilde{Q}\tilde{A}^* + \tilde{V}, \tag{4.8}$$

where (4.8) *holds in the sense discussed in* § 3. *Furthermore, $\tilde{P}$ is given by* (4.5) *if and only if $\tilde{P} \in \mathscr{B}(\tilde{\mathscr{H}})$ satisfies*

$$\tilde{P}\colon \mathscr{D}(\tilde{A}) \to \mathscr{D}(\tilde{A}^*), \tag{4.9}$$

$$0 = \tilde{A}^*\tilde{P} + \tilde{P}\tilde{A} + \tilde{R}. \tag{4.10}$$

Proof. We consider $\tilde{Q}$ only. To prove necessity let $t' > 0$. Then for all $t \in [0, t')$ and $\tilde{x} \in \mathscr{D}(\tilde{A}^*)$ we can write

$$e^{\tilde{A}t} \tilde{Q} e^{\tilde{A}^* t'} \tilde{x} = \int_0^{\infty} e^{\tilde{A}(t+s)} \tilde{V} e^{\tilde{A}^*(t'+s)} \tilde{x} \, ds$$

$$= \int_t^{\infty} e^{\tilde{A}\sigma} \tilde{V} e^{\tilde{A}^*\sigma} \, e^{\tilde{A}^*(t'-t)} \tilde{x} \, d\sigma.$$

Hence,

$$\frac{d}{dt} e^{\tilde{A}t} \tilde{Q} e^{\tilde{A}^* t'} \tilde{x} = -\int_t^{\infty} e^{\tilde{A}\sigma} \tilde{V} e^{\tilde{A}^*\sigma} \, e^{\tilde{A}^*(t'-t)} \tilde{A}^* \tilde{x} \, d\sigma - e^{\tilde{A}t} \tilde{V} e^{\tilde{A}^* t'} \tilde{x}, \tag{4.11}$$

which shows that $e^{\tilde{A}t} \tilde{Q} e^{\tilde{A}^* t'}$ is strongly differentiable with respect to t for all $t \in [0, t')$. In particular, setting $t = 0$ it follows that $\tilde{Q} e^{\tilde{A}^* t'} \tilde{x} \in \mathscr{D}(\tilde{A})$ for all $\tilde{x} \in \mathscr{D}(\tilde{A}^*)$ (see, e.g., [6, p. 173] or [50, p. 485]). Performing the differentiation on the left-hand side of

(4.11) and setting $t=0$ yields

$$\tilde{A}\tilde{Q}\,e^{\tilde{A}^*t'}\tilde{x} = -\int_0^\infty e^{\tilde{A}\sigma}\tilde{V}\,e^{\tilde{A}^*\sigma}\,e^{\tilde{A}^*t'}\tilde{A}^*\tilde{x}\,d\sigma - \tilde{V}\,e^{\tilde{A}^*t'}\tilde{x}. \tag{4.12}$$

Now fix $\tilde{x}\in\mathcal{D}(\tilde{A}^*)$. Then for $\{t_i\}_{i=1}^\infty$, $t_i>0$, $t_i\to 0$, we have

$$\tilde{Q}\,e^{\tilde{A}^*t_i}\tilde{x}\in\mathcal{D}(\tilde{A}), \qquad i=1,2,3,\cdots,$$
$$\tilde{Q}\,e^{\tilde{A}^*t_i}\tilde{x}\xrightarrow[i\to\infty]{}\tilde{Q}\tilde{x}.$$

Now consider the sequence $\{\tilde{A}\tilde{Q}\,e^{\tilde{A}^*t_i}\tilde{x}\}_{i=1}^\infty$. Letting $t'=t_i$ in (4.12) and using dominated convergence to interchange limit and integration ($\tilde{A}^*\tilde{x}$ is a fixed element of $\tilde{\mathscr{H}}$), it follows that

$$\lim_{i\to\infty}\tilde{A}\tilde{Q}\,e^{\tilde{A}^*t_i}\tilde{x} = -\int_0^\infty e^{\tilde{A}\sigma}\tilde{V}\,e^{\tilde{A}^*\sigma}\tilde{A}^*\tilde{x}\,d\sigma - \tilde{V}\tilde{x}. \tag{4.13}$$

Since $\tilde{A}$ is closed, $\tilde{Q}\tilde{x}\in\mathcal{D}(\tilde{A})$. This proves (4.7). Also, since $\tilde{A}$ is closed we have

$$\lim_{i\to\infty}\tilde{A}\tilde{Q}\,e^{\tilde{A}^*t_i}\tilde{x} = \tilde{A}\tilde{Q}\tilde{x},$$

which with (4.13) implies

$$\tilde{A}\tilde{Q}\tilde{x} = -\tilde{Q}\tilde{A}^*\tilde{x} - \tilde{V}\tilde{x},$$

and hence

$$(\tilde{A}\tilde{Q}+\tilde{Q}\tilde{A}^*+\tilde{V})\tilde{x}=0, \qquad \tilde{x}\in\mathcal{D}(\tilde{A}^*),$$

as desired.

To prove sufficiency let $\tilde{x}\in\mathcal{D}(\tilde{A})$. Then $e^{\tilde{A}^*t}\tilde{x}\in\mathcal{D}(\tilde{A}^*)$, $t\geqq 0$, and hence

$$\frac{d}{dt}\,e^{\tilde{A}t}\tilde{Q}\,e^{\tilde{A}^*t}\tilde{x} = e^{\tilde{A}t}(\tilde{A}\tilde{Q}+\tilde{Q}\tilde{A}^*)\,e^{\tilde{A}^*t}\tilde{x}.$$

Thus

$$e^{\tilde{A}t}\tilde{Q}\,e^{\tilde{A}^*t}\tilde{x} - \tilde{Q}\tilde{x} = \int_0^t e^{\tilde{A}s}(\tilde{A}\tilde{Q}+\tilde{Q}\tilde{A}^*)\,e^{\tilde{A}^*s}\tilde{x}\,ds, \qquad \tilde{x}\in\mathcal{D}(\tilde{A}^*).$$

Extending $\tilde{A}\tilde{Q}+\tilde{Q}\tilde{A}^*$ to all of $\tilde{\mathscr{H}}$ we obtain

$$e^{\tilde{A}t}\tilde{Q}\,e^{\tilde{A}^*t}\tilde{x} - \tilde{Q}\tilde{x} = -\int_0^t e^{\tilde{A}s}\tilde{V}\,e^{\tilde{A}^*s}\tilde{x}\,ds, \qquad \tilde{x}\in\tilde{\mathscr{H}}.$$

Letting $t\to\infty$ yields (4.3). □

We now introduce some notation which will prove to be most convenient in the following results. For $(A_c', B_c', C_c')\in\mathbb{R}^{n_c\times n_c}\times\mathbb{R}^{n_c\times l}\times\mathbb{R}^{m\times n_c}$ define

$$\delta_{A_c}\triangleq A_c'-A_c, \qquad \delta_{B_c}\triangleq B_c'-B_c, \qquad \delta_{C_c}\triangleq C_c'-C_c$$

and

$$\|(\delta_{A_c},\delta_{B_c},\delta_{C_c})\|\triangleq\|\delta_{A_c}\|+\|\delta_{B_c}\|+\|\delta_{C_c}\|.$$

Furthermore, let $\tilde{A}'$, $\tilde{V}'$ and $\tilde{R}'$ denote $\tilde{A}$, $\tilde{V}$ and $\tilde{R}$ with (A_c, B_c, C_c) replaced by

(A_c', B_c', C_c') and define

$$\delta_{\tilde{A}} \triangleq \tilde{A}' - \tilde{A} = \begin{bmatrix} 0 & B\delta_{C_c} \\ \delta_{B_c} C & \delta_{A_c} \end{bmatrix},$$

$$\delta_{\tilde{V}} \triangleq \tilde{V}' - \tilde{V} = \begin{bmatrix} 0 & 0 \\ 0 & B_c V_2 \delta_{B_c}^T + \delta_{B_c} V_2 B_c^T + \delta_{B_c} V_2 \delta_{B_c}^T \end{bmatrix},$$

$$\delta_{\tilde{R}} \triangleq \tilde{R}' - \tilde{R} = \begin{bmatrix} 0 & 0 \\ 0 & C_c^T R_2 \delta_{C_c} + \delta_{C_c}^T R_2 C_c + \delta_{C_c}^T R_2 \delta_{C_c} \end{bmatrix}.$$

We shall also write $\tilde{Q}'$, $\tilde{P}'$ for $\tilde{Q}$, $\tilde{P}$ as given by (4.3) and (4.5) with $\tilde{A}$, $\tilde{V}$, $\tilde{R}$ replaced by $\tilde{A}'$, $\tilde{V}'$, $\tilde{R}'$ and define

$$\delta_{\tilde{Q}} \triangleq \tilde{Q}' - \tilde{Q}, \qquad \delta_{\tilde{P}} \triangleq \tilde{P}' - \tilde{P}.$$

LEMMA 4.5. $\mathscr{A}$ is open.

Proof. Let $(A_c, B_c, C_c) \in \mathscr{A}$ be arbitrary and consider the open set

$$N \triangleq \{(A_c', B_c', C_c') \in \mathbb{R}^{n_c \times n_c} \times \mathbb{R}^{n_c \times l} \times \mathbb{R}^{m \times n_c} : \|(\delta_{A_c}, \delta_{B_c}, \delta_{C_c})\| < \beta/2\alpha\gamma\}, \tag{4.14}$$

where $\gamma \triangleq \max\{1, \|B\|, \|C\|\}$. Then, since $\tilde{A}' = \tilde{A} + \delta_{\tilde{A}}$ and $\delta_{\tilde{A}} \in \mathscr{B}(\tilde{\mathscr{H}})$ it follows from Theorem 2.1, p. 497 of [50], that for all $(A_c', B_c', C_c') \in N$ and $t \geqq 0$,

$$\|e^{\tilde{A}'t}\| \leqq \alpha\, e^{(-\beta + \alpha\|\delta_{\tilde{A}}\|)t} \leqq \alpha\, e^{-\beta t/2}.$$

Hence, $N \subset \mathscr{A}$, as desired. □

LEMMA 4.6. *There exists $c > 0$ such that*

$$\|\delta_{\tilde{Q}}\| \leqq c\|(\delta_{A_c}, \delta_{B_c}, \delta_{C_c})\|, \tag{4.15}$$

$$\|\delta_{\tilde{P}}\| \leqq c\|(\delta_{A_c}, \delta_{B_c}, \delta_{C_c})\|, \tag{4.16}$$

for all $(A_c', B_c', C_c') \in N$, where $N \subset \mathscr{A}$ is the open neighborhood of (A_c, B_c, C_c) defined by (4.14).

Proof. We consider (4.15) only. Since $\|e^{\tilde{A}'t}\| \leqq \alpha e^{-\beta t/2}$, $t \geqq 0$, $(A_c', B_c', C_c') \in N$, it follows that

$$\begin{aligned} \|\delta_{\tilde{Q}}\| &\leqq \int_0^\infty \|e^{\tilde{A}'t}\tilde{V}'e^{\tilde{A}'^*t} - e^{\tilde{A}t}\tilde{V}e^{\tilde{A}^*t}\|\, dt \\ &\leqq \int_0^\infty \{\|e^{\tilde{A}'t}\|\,\|\tilde{V}'\|\,\|e^{\tilde{A}'^*t} - e^{\tilde{A}^*t}\| + \|e^{\tilde{A}'t}\|\,\|\delta_{\tilde{V}}\|\,\|e^{\tilde{A}^*t}\| + \|e^{\tilde{A}'t} - e^{\tilde{A}t}\|\,\|\tilde{V}\|\,\|e^{\tilde{A}^*t}\|\}\, dt \\ &\leqq \alpha(\|\tilde{V}\| + \|\delta_{\tilde{V}}\|) \int_0^\infty \|e^{(\tilde{A}^* + \delta_{\tilde{A}}^*)t} - e^{\tilde{A}^*t}\|\, e^{-\beta t/2}\, dt \qquad (4.17) \\ &\quad + \alpha^2\|\delta_{\tilde{V}}\| \int_0^\infty e^{-3\beta t/2}\, dt + \alpha\|\tilde{V}\| \int_0^\infty \|e^{(\tilde{A} + \delta_{\tilde{A}})t} - e^{\tilde{A}t}\|\, e^{-\beta t/2}\, dt \\ &= \alpha(2\|\tilde{V}\| + \|\delta_{\tilde{V}}\|) \int_0^\infty \|e^{(\tilde{A} + \delta_{\tilde{A}})t} - e^{\tilde{A}t}\| e^{-\beta t/2}\, dt + \frac{2\alpha^2}{3\beta}\|\delta_{\tilde{V}}\|. \end{aligned}$$

From [50, p. 497], it follows that the perturbed semigroup $e^{(\tilde{A} + \delta_{\tilde{A}})t}$ has an expansion

$$e^{(\tilde{A} + \delta_{\tilde{A}})t} = e^{\tilde{A}t} + \sum_{i=1}^{\infty} U_i(t), \qquad t \geqq 0,$$

where $U_i(t) \in \mathcal{B}(\tilde{\mathcal{H}})$, $t \geq 0$, satisfy the estimates

$$\|U_i(t)\| \leq \alpha^{i+1} \|\delta_{\tilde{A}}\|^i e^{-\beta t} t^i / i!.$$

Hence, for all $(A_c', B_c', C_c') \in N$,

$$\|e^{(\tilde{A}+\delta_{\tilde{A}})t} - e^{\tilde{A}t}\| \leq \sum_{i=1}^{\infty} \|U_i(t)\| \leq \alpha e^{-\beta t}[e^{\alpha\|\delta_{\tilde{A}}\|t} - 1]. \tag{4.18}$$

From (4.17), (4.18) and the relations $\|\delta_{\tilde{A}}\| \leq \gamma\|(\delta_{A_c}, \delta_{B_c}, \delta_{C_c})\| < \beta/2\alpha$ and

$$\int_0^{\infty} [e^{\alpha\|\delta_{\tilde{A}}\|t} - 1]\, e^{-3\beta t/2}\, dt < \frac{\alpha\gamma}{3\beta^2} \|(\delta_{A_c}, \delta_{B_c}, \delta_{C_c})\|$$

it follows that

$$\|\delta_{\tilde{Q}}\| \leq \frac{2\alpha^3\gamma}{3\beta^2}(2\|\tilde{V}\| + \|\delta_{\tilde{V}}\|)\|(\delta_{A_c}, \delta_{B_c}, \delta_{C_c})\|$$

$$+ \frac{2\alpha^2}{3\beta}(2\|B_c V_2\|\, \|\delta_{B_c}\| + \|V_2\|\, \|\delta_{B_c}\|^2),$$

which yields (4.15). □

Since $\tilde{Q}, \tilde{P} \in \mathcal{B}(\tilde{\mathcal{H}})$ we can write

$$\tilde{Q} = \begin{bmatrix} Q_1 & Q_{12} \\ Q_{12}^* & Q_2 \end{bmatrix}, \qquad \tilde{P} = \begin{bmatrix} P_1 & P_{12} \\ P_{12}^* & P_2 \end{bmatrix},$$

where $Q_1 \in \mathcal{B}(\mathcal{H})$, $Q_{12} \in \mathcal{B}(\mathbb{R}^{n_c}, \mathcal{H})$, $Q_2 \in \mathbb{R}^{n_c \times n_c}$ and similarly for P_1, P_{12} and P_2. Note that Q_1, Q_2, P_1 and P_2 are nonnegative definite. Also, define the notation

$$\tilde{P}\tilde{Q} = \begin{bmatrix} Z_1 & Z_{12} \\ Z_{21} & Z_2 \end{bmatrix},$$

where

$$Z_1 \triangleq P_1 Q_1 + P_{12} Q_{12}^*, \qquad Z_{12} \triangleq P_1 Q_{12} + P_{12} Q_2,$$
$$Z_{21} \triangleq P_{12}^* Q_1 + P_2 Q_{12}^*, \qquad Z_2 \triangleq P_{12}^* Q_{12} + P_2 Q_2,$$

and, for $(A_c', B_c', C_c') \in \mathcal{A}$, let

$$\delta_J(\delta_{A_c}, \delta_{B_c}, \delta_{C_c}) \triangleq J(A_c', B_c', C_c') - J(A_c, B_c, C_c).$$

LEMMA 4.7. *Let* $(A_c', B_c', C_c') \in \mathcal{A}$. *Then*

$$\delta_J(\delta_{A_c}, \delta_{B_c}, \delta_{C_c}) = \mathcal{L}(\delta_{A_c}, \delta_{B_c}, \delta_{C_c}) + o(\|(\delta_{A_c}, \delta_{B_c}, \delta_{C_c})\|), \tag{4.19}$$

where

$$\mathcal{L}(\delta_{A_c}, \delta_{B_c}, \delta_{C_c}) \triangleq 2\,\mathrm{tr}\,[Z_2 \delta_{A_c}] + 2\,\mathrm{tr}\,[(V_2 B_c^T P_2 + C Z_{21}^*)\delta_{B_c}] + 2\,\mathrm{tr}\,[Q_2 C_c^T R_2 + Z_{12}^* B)\delta_{C_c}] \tag{4.20}$$

and

$$\lim_{(\delta_{A_c}, \delta_{B_c}, \delta_{C_c}) \to 0} \|(\delta_{A_c}, \delta_{B_c}, \delta_{C_c})\|^{-1} o(\|(\delta_{A_c}, \delta_{B_c}, \delta_{C_c})\|) = 0. \tag{4.21}$$

Proof. Combining (4.8) and (4.10) with (4.6), J can be written as

$$J(A_c, B_c, C_c) = \mathrm{tr}\,[\tilde{Q}\tilde{R} + \tilde{P}\tilde{V}] + \tfrac{1}{2}\,\mathrm{tr}\,[\tilde{Q}\,\mathrm{cl}\,(\tilde{A}^*\tilde{P} + \tilde{P}\tilde{A}) + \tilde{P}\,\mathrm{cl}\,(\tilde{A}\tilde{Q} + \tilde{Q}\tilde{A}^*)],$$

and likewise for (A_c', B_c', C_c'), where "cl" denotes closure (i.e., extension) of a bounded operator to all of $\mathscr{H}$. Now using the identity

$$\operatorname{tr}[\tilde{Q}'\tilde{R}'+\tilde{P}'\tilde{V}']-\operatorname{tr}[\tilde{Q}\tilde{R}+\tilde{P}\tilde{V}]=\operatorname{tr}[\tilde{Q}\delta_{\tilde{R}}+\tilde{P}\delta_{\tilde{V}}]+\operatorname{tr}[\delta_{\tilde{Q}}\tilde{R}'+\delta_{\tilde{P}}\tilde{V}']$$

we can compute

$$\begin{aligned}\delta_J(\delta_{A_c}, \delta_{B_c}, \delta_{C_c}) &= \operatorname{tr}[\tilde{Q}\delta_{\tilde{R}}+\tilde{P}\delta_{\tilde{V}}]+\tfrac{1}{2}\operatorname{tr}[\tilde{Q}\ \operatorname{cl}(\tilde{A}^*(\tilde{P}+\delta_{\tilde{P}})+(\tilde{P}+\delta_{\tilde{P}})\tilde{A}')]\\ &\quad+\tfrac{1}{2}\operatorname{tr}[\delta_{\tilde{Q}}\ \operatorname{cl}(\tilde{A}'^*\tilde{P}'+\tilde{P}'\tilde{A}')]\\ &\quad+\tfrac{1}{2}\operatorname{tr}[\tilde{P}\ \operatorname{cl}(\tilde{A}'(\tilde{Q}+\delta_{\tilde{Q}})+(\tilde{Q}+\delta_{\tilde{Q}})\tilde{A}'^*)]\\ &\quad+\tfrac{1}{2}\operatorname{tr}[\delta_{\tilde{P}}\ \operatorname{cl}(\tilde{A}'\tilde{Q}'+\tilde{Q}'\tilde{A}'^*)]\\ &\quad-\tfrac{1}{2}\operatorname{tr}[\tilde{Q}\ \operatorname{cl}(\tilde{A}^*\tilde{P}+\tilde{P}\tilde{A})+\tilde{P}\ \operatorname{cl}(\tilde{A}\tilde{Q}+\tilde{Q}\tilde{A}^*)]\\ &\quad+\operatorname{tr}[\delta_{\tilde{Q}}\tilde{R}'+\delta_{\tilde{P}}\tilde{V}'].\end{aligned}$$

Using $\tilde{A}'=\tilde{A}+\delta_{\tilde{A}}$ and combining the second, fourth and sixth terms yields

$$\delta_J(\delta_{A_c}, \delta_{B_c}, \delta_{C_c})=\Lambda+\Omega,$$

where

$$\begin{aligned}\Lambda &\triangleq \operatorname{tr}[\tilde{Q}\delta_{\tilde{R}}+\tilde{P}\delta_{\tilde{V}}]+\tfrac{1}{2}\operatorname{tr}[\tilde{Q}(\delta_{\tilde{A}}^*\tilde{P}+\tilde{P}\delta_{\tilde{A}})+\tilde{P}(\delta_{\tilde{A}}\tilde{Q}+\tilde{Q}\delta_{\tilde{A}}^*)]\\ &=\operatorname{tr}[\tilde{Q}\delta_{\tilde{R}}+\tilde{P}\delta_{\tilde{V}}]+2\operatorname{tr}[\delta_{\tilde{A}}\tilde{Q}\tilde{P}]\end{aligned}$$

and

$$\begin{aligned}\Omega &\triangleq \tfrac{1}{2}\operatorname{tr}[\tilde{Q}\ \operatorname{cl}(\tilde{A}'^*\delta_{\tilde{P}}+\delta_{\tilde{P}}\tilde{A}')+\tilde{P}\ \operatorname{cl}(\tilde{A}'\delta_{\tilde{Q}}+\delta_{\tilde{Q}}\tilde{A}'^*)]\\ &\quad+\tfrac{1}{2}\operatorname{tr}[\delta_{\tilde{Q}}\ \operatorname{cl}(\tilde{A}'^*\tilde{P}'+\tilde{P}'\tilde{A}')+\delta_{\tilde{P}}\ \operatorname{cl}(\tilde{A}'\tilde{Q}'+\tilde{Q}'\tilde{A}'^*)]+\operatorname{tr}[\delta_{\tilde{Q}}\tilde{R}'+\delta_{\tilde{P}}\tilde{V}'].\end{aligned}$$

Computing

$$\operatorname{tr}[\tilde{Q}\delta_{\tilde{R}}+\tilde{P}\delta_{\tilde{V}}]=2\operatorname{tr}[V_2B_c^TP_2\delta_{B_c}]+2\operatorname{tr}[Q_2C_c^TR_2\delta_{C_c}]+\operatorname{tr}[P_2\delta_{B_c}V_2\delta_{B_c}^T+Q_2\delta_{C_c}^TR_2\delta_{C_c}]$$

and

$$2\operatorname{tr}[\delta_{\tilde{A}}\tilde{Q}\tilde{P}]=2\operatorname{tr}[Z_2\delta_{A_c}]+2\operatorname{tr}[CZ_{21}^*\delta_{B_c}]+2\operatorname{tr}[Z_{12}^*B\delta_{C_c}]$$

and retaining first-order terms, we obtain (4.20).

To evaluate Ω, use (4.8) and (4.10) to replace $\tilde{R}'$ and $\tilde{V}'$ in the last term in Ω and write $\tilde{A}'=\tilde{A}+\delta_{\tilde{A}}$, to obtain

$$\begin{aligned}\Omega &= \tfrac{1}{2}\operatorname{tr}[\tilde{Q}\ \operatorname{cl}(\tilde{A}^*\delta_{\tilde{P}}+\delta_{\tilde{P}}\tilde{A})+\tilde{P}\ \operatorname{cl}(\tilde{A}\delta_{\tilde{Q}}+\delta_{\tilde{Q}}\tilde{A}^*)]\\ &\quad+\tfrac{1}{2}\operatorname{tr}[\tilde{Q}(\delta_{\tilde{A}}^*\delta_{\tilde{P}}+\delta_{\tilde{P}}\delta_{\tilde{A}})+\tilde{P}(\delta_{\tilde{A}}\delta_{\tilde{Q}}+\delta_{\tilde{Q}}\delta_{\tilde{A}}^*)]\\ &\quad-\tfrac{1}{2}\operatorname{tr}[\delta_{\tilde{Q}}\ \operatorname{cl}(\tilde{A}'^*\tilde{P}'+\tilde{P}'\tilde{A}')+\delta_{\tilde{P}}\ \operatorname{cl}(\tilde{A}'\tilde{Q}'+\tilde{Q}'\tilde{A}'^*)].\end{aligned}\tag{4.22}$$

Next we note that

$$\operatorname{tr}[\tilde{Q}\ \operatorname{cl}(\tilde{A}^*\delta_{\tilde{P}}+\delta_{\tilde{P}}\tilde{A}^*)]=\operatorname{tr}[\delta_{\tilde{P}}\ \operatorname{cl}(\tilde{A}\tilde{Q}+\tilde{Q}\tilde{A}^*)].\tag{4.23}$$

To see this we observe that by arguments similar to those used in the proof of Lemma 4.4 and the fact that $\delta_{\tilde{P}}\colon \mathscr{D}(\tilde{A})\to\mathscr{D}(\tilde{A}^*)$ it follows that

$$\delta_{\tilde{P}}=-\int_0^\infty e^{\tilde{A}^*t}\ \operatorname{cl}(\tilde{A}^*\delta_{\tilde{P}}+\delta_{\tilde{P}}\tilde{A})\ e^{\tilde{A}t}\,dt.$$

Now, using the technique of Lemma 4.3 with the role of $\tilde{R}$ played by $-\operatorname{cl}(\tilde{A}^*\delta_{\tilde{P}}+\delta_{\tilde{P}}\tilde{A})$,

we see that

$$\operatorname{tr}[\tilde{Q}\,\mathrm{cl}\,(\tilde{A}^*\delta_{\tilde{P}}+\delta_{\tilde{P}}\tilde{A})]=-\operatorname{tr}[\delta_{\tilde{P}}\tilde{V}]=\operatorname{tr}[\delta_{\tilde{P}}\,\mathrm{cl}\,(\tilde{A}\tilde{Q}+\tilde{Q}\tilde{A}^*)].$$

Similarly, it can be shown that

$$\operatorname{tr}[\tilde{P}\,\mathrm{cl}\,(\tilde{A}\delta_{\tilde{Q}}+\delta_{\tilde{Q}}\tilde{A}^*)]=\operatorname{tr}[\delta_{\tilde{Q}}\,\mathrm{cl}\,(\tilde{A}^*\tilde{P}+\tilde{P}\tilde{A})]. \tag{4.24}$$

Now substitute (4.23) and (4.24) into (4.22) and rearrange the second term in (4.22) so that

$$\begin{aligned}\Omega&=\tfrac{1}{2}\operatorname{tr}[\delta_{\tilde{Q}}\,\mathrm{cl}\,(\tilde{A}^*\tilde{P}+\tilde{P}\tilde{A})+\delta_{\tilde{P}}\,\mathrm{cl}\,(\tilde{A}\tilde{Q}+\tilde{Q}\tilde{A}^*)]\\&\quad+\tfrac{1}{2}\operatorname{tr}[\delta_{\tilde{Q}}(\delta_{\tilde{A}}^*\tilde{P}+\tilde{P}\delta_{\tilde{A}})+\delta_{\tilde{P}}(\delta_{\tilde{A}}\tilde{Q}+\tilde{Q}\delta_{\tilde{A}}^*)]\\&\quad-\tfrac{1}{2}\operatorname{tr}[\delta_{\tilde{Q}}\,\mathrm{cl}\,(\tilde{A}'^*\tilde{P}'+\tilde{P}'\tilde{A}')+\delta_{\tilde{P}}\,\mathrm{cl}\,(\tilde{A}'\tilde{Q}'+\tilde{Q}'\tilde{A}'^*)]\\&=-\tfrac{1}{2}\operatorname{tr}[\delta_{\tilde{Q}}\,\mathrm{cl}\,(\tilde{A}'^*\delta_{\tilde{P}}+\delta_{\tilde{P}}\tilde{A}')+\delta_{\tilde{P}}\,\mathrm{cl}\,(\tilde{A}'\delta_{\tilde{Q}}+\delta_{\tilde{Q}}\tilde{A}'^*)].\end{aligned}$$

Using (4.8) to obtain

$$0=\tilde{A}'\delta_{\tilde{Q}}+\delta_{\tilde{Q}}\tilde{A}'^*+\delta_{\tilde{A}}\tilde{Q}+\tilde{Q}\delta_{\tilde{A}}^*+\delta_{\tilde{V}}$$

and (4.10) to obtain a similar relation involving $\tilde{P}$, we have

$$\Omega=\operatorname{tr}[\delta_{\tilde{Q}}(\delta_{\tilde{A}}^*\tilde{P}+\tilde{P}\delta_{\tilde{A}}+\delta_{\tilde{R}})]+\operatorname{tr}[\delta_{\tilde{P}}\tilde{Q}+\tilde{Q}\delta_{\tilde{A}}^*+\delta_{\tilde{V}})].$$

Restricting (A_c', B_c', C_c') to N (see (4.14)), using Lemma 4.6 and noting that $\delta_{\tilde{A}}$ and $\delta_{\tilde{R}}$ have finite rank, it follows that there exists $c_1>0$ such that

$$\|\Omega\|\leqq c_1\|(\delta_{A_c},\delta_{B_c},\delta_{C_c})\|^2. \tag{4.25}$$

Combining Ω with the second-order terms in Λ yields the desired result. □

LEMMA 4.8. $\mathscr{A}_+$ is open.

Proof. From the "generic" property of controllability and observability [62, p. 44] there exists an open neighborhood of (A_c, B_c, C_c) each of whose elements is minimal. Combining this fact with Lemma 4.5 yields the desired result. □

LEMMA 4.9. *Q_2 and P_2 are positive definite.*

Proof. First note that expanding the $\mathbb{R}^{n_c\times n_c}$-component of the Lyapunov equation (4.8) yields (4.50) below. By a minor extension of results from [66] or [67], (4.50) can be rewritten as

$$0=(A_c+B_cCQ_{12}Q_2^+)Q_2+Q_2(A_c+B_cCQ_{12}Q_2^+)^T+B_cV_2B_c^T,$$

where Q_2^+ is the Moore-Penrose or Drazin generalized inverse of Q_2. Next note that since (A_c, B_c) is controllable then so is $(A_c+B_cCQ_{12}Q_2^+, B_cV_2^{1/2})$. Now, since Q_2 and $B_cV_2B_c^T$ are nonnegative definite, it follows from [62, Lemma 12.2] that Q_2 is positive definite. Similar arguments show that P_2 is positive definite. □

Having established Lemmas 4.1-4.9, we can now proceed with the proof of the Main Theorem. Let $(A_c, B_c, C_c)\in\mathscr{A}_+$ be as in the Main Theorem and consider (4.19) with (A_c', B_c', C_c') confined to $\mathscr{A}_+$. Because $\mathscr{L}\colon \mathbb{R}^{n_c\times n_c}\times\mathbb{R}^{n_c\times l}\times\mathbb{R}^{m\times n_c}\to\mathbb{R}$ is a bounded linear functional and $\mathscr{A}_+$ is open, the convergence in (4.21) implies that $\mathscr{L}$ is precisely the Frechet derivative of J with respect to (A_c, B_c, C_c). Since $\mathscr{A}_+$ is open, the optimality of (A_c, B_c, C_c) implies

$$\mathscr{L}(\delta_{A_c},\delta_{B_c},\delta_{C_c})=0 \tag{4.26}$$

for all $(\delta_{A_c}, \delta_{B_c}, \delta_{C_c})$. Clearly, (4.26) is equivalent to

$$Z_2=0, \tag{4.27}$$

(4.28) $$V_2 B_c^T P_2 + C Z_{21}^* = 0,$$

(4.29) $$Q_2 C_c^T R_2 + Z_{12}^* B = 0.$$

Thus, B_c and C_c are given by

(4.30) $$B_c = -P_2^{-1} Z_{21} C^* V_2^{-1},$$

(4.31) $$C_c = -R_2^{-1} B^* Z_{12} Q_2^{-1}.$$

Although B_c and C_c are now determined in terms of $\tilde{Q}$ and $\tilde{P}$, A_c remains to be found. Moreover, $\tilde{Q}$ and $\tilde{P}$ themselves depend (via (4.8) and (4.10)) on B_c and C_c. Hence our task now is to consolidate and simplify (4.7)-(4.10), (4.27), (4.30) and (4.31) to obtain the more tractable conditions (3.9)-(3.18). To this end let us define new variables

(4.32a, b) $$Q \triangleq Q_1 - Q_{12} Q_2^{-1} Q_{12}^*, \qquad P \triangleq P_1 - P_{12} P_2^{-1} P_{12}^*,$$

(4.33a, b) $$\hat{Q} \triangleq Q_{12} Q_2^{-1} Q_{12}^*, \qquad \hat{P} \triangleq P_{12} P_2^{-1} P_{12}^*.$$

Clearly, $\hat{Q}$ and $\hat{P}$ are nonnegative definite and have finite rank. Since by Lemma 4.2 $\tilde{Q}, \tilde{P} \in \mathscr{B}_1(\tilde{\mathscr{H}})$, it can be seen that $Q_1, P_1 \in \mathscr{B}_1(\mathscr{H})$, which implies $Q, P \in \mathscr{B}_1(\mathscr{H})$. To show that Q and P are nonnegative definite, note that Q is the $\mathscr{B}(\mathscr{H})$-component of the nonnegative-definite operator $\mathscr{Q}\tilde{Q}\mathscr{Q}^* \in \mathscr{B}(\tilde{\mathscr{H}})$, where

$$\mathscr{Q} \triangleq \begin{bmatrix} I_{\mathscr{H}} & -Q_{12} Q_2^{-1} \\ 0 & -I_{n_c} \end{bmatrix}.$$

Similarly, P is nonnegative definite.

From the domain conditions (4.7) and (4.9) it follows that

(4.34a, b) $$Q_1 : \mathscr{D}(A^*) \to \mathscr{D}(A), \qquad P_1 : \mathscr{D}(A) \to \mathscr{D}(A^*),$$

(4.35a, b) $$Q_{12} : \mathbb{R}^{n_c} \to \mathscr{D}(A), \qquad P_{12} : \mathbb{R}^{n_c} \to \mathscr{D}(A^*),$$

which lead to (3.12) and (3.13).

Next note that (4.27) is equivalent to (3.8) with

(4.36a, b) $$G \triangleq Q_2^{-1} Q_{12}^*, \qquad \Gamma \triangleq -P_2^{-1} P_{12}^*$$

and that (3.7) holds with

(4.37) $$M \triangleq Q_2 P_2.$$

Since Q_2 and P_2 are positive definite, Lemma 2.6 implies that M is positive semisimple. We can also define $\tau = G^* \Gamma$ which, by (3.8) satisfies $\tau^2 = \tau$. It is helpful to note the identities

(4.38a, b) $$\hat{Q} = Q_{12} G = G^* Q_{12}^*, \qquad \hat{P} = -P_{12} \Gamma = -\Gamma^* P_{12}^*,$$

(4.39a, b) $$\hat{Q} = G^* Q_2 G, \qquad \hat{P} = \Gamma^* P_2 \Gamma,$$

(4.40a, b) $$\tau G^* = G^*, \qquad \Gamma \tau = \Gamma,$$

(4.41a, b) $$\hat{Q} = \tau \hat{Q}, \qquad \hat{P} = \hat{P} \tau,$$

(4.42) $$\hat{Q} \hat{P} = -Q_{12} P_{12}^*.$$

From (3.8) and (2.1) it follows that

(4.43a, b) $$\rho(G) = \rho(\Gamma) = n_c,$$

(4.44a, b) $$\rho(Q_{12}) = \rho(P_{12}) = n_c.$$

Hence, (2.2) and (4.38) imply $n_c = \rho(Q_{12}) + \rho(G) - n_c \leqq \rho(\hat{Q}) \leqq \rho(Q_{12}) = n_c$, which yields (3.14a). Similarly, (3.14b) holds and (3.14c) follows from (2.2) and (4.42).

Using (4.38) and (4.39), the components of $\tilde{Q}$ and $\tilde{P}$ can be written in terms of G, Γ, Q, P, $\hat{Q}$ and $\hat{P}$ as

$$Q_1 = Q + \hat{Q}, \qquad P_1 = P + \hat{P}, \tag{4.45}$$

$$Q_{12} = \hat{Q}\Gamma^*, \qquad P_{12} = -\hat{P}G^*, \tag{4.46}$$

$$Q_2 = \Gamma\hat{Q}\Gamma^*, \qquad P_2 = G\hat{P}G^*. \tag{4.47}$$

Now (3.10) and (3.11) can be obtained by substituting (4.45)-(4.47) into (4.30) and (4.31).

Expanding the $\mathscr{B}(\mathscr{H})$, $\mathscr{B}(\mathbb{R}^{n_c}, \mathscr{H})$ and $\mathbb{R}^{n_c \times n_c}$ components of (4.8) and (4.10) yields

$$0 = AQ_1 + Q_1A^* + BC_cQ_{12}^* + Q_{12}(BC_c)^* + V_1, \tag{4.48}$$

$$0 = AQ_{12} + Q_{12}A_c^T + BC_cQ_2 + Q_1(B_cC)^*, \tag{4.49}$$

$$0 = A_cQ_2 + Q_2A_c^T + B_cCQ_{12} + Q_{12}^*(B_cC)^* + B_cV_2B_c^T, \tag{4.50}$$

$$0 = A^*P_1 + P_1A + (B_cC)^*P_{12}^* + P_{12}B_cC + R_1, \tag{4.51}$$

$$0 = P_{12}A_c + A^*P_{12} + (B_cC)^*P_2 + P_1BC_c, \tag{4.52}$$

$$0 = A_c^TP_2 + P_2A_c + (BC_c)^*P_{12} + P_{12}^*BC_c + C_c^TR_2C_c. \tag{4.53}$$

Substituting (4.45)-(4.47) into (4.48)-(4.53), using the identities

$$B_cC = \Gamma Q\bar{\Sigma}, \qquad BC_c = -\Sigma PG^*,$$

$$B_cV_2B_c^T = \Gamma Q\bar{\Sigma}Q\Gamma^*, \qquad C_c^TR_2C_c = GP\Sigma PG^*,$$

and defining

$$A_Q \triangleq A - Q\bar{\Sigma}, \qquad A_P \triangleq A - \Sigma P,$$

we obtain

$$0 = AQ + QA^* + A_P\hat{Q} + \hat{Q}A_P + V_1, \tag{4.54}$$

$$0 = [A_P\hat{Q} + Q\bar{\Sigma}Q + \hat{Q}(\Gamma^*A_c^TG + \bar{\Sigma}Q)]\Gamma^*, \tag{4.55}$$

$$0 = \Gamma[G^*A_c\Gamma\hat{Q} + Q\bar{\Sigma}\hat{Q} + Q\bar{\Sigma}Q + \hat{Q}(\Gamma^*A_c^TG + \bar{\Sigma}Q)]\Gamma^*, \tag{4.56}$$

$$0 = A^*P + PA + A_Q^*\hat{P} + \hat{P}A_Q + R_1, \tag{4.57}$$

$$0 = -[A_Q^*\hat{P} + P\Sigma P + \hat{P}(G^*A_c\Gamma + \Sigma P)]G^*, \tag{4.58}$$

$$0 = G[\Gamma^*A_c^TG\hat{P} + P\Sigma\hat{P} + P\Sigma P + \hat{P}(G^*A_c\Gamma + \Sigma P)]G^*. \tag{4.59}$$

We are now in a position to determine A_c by computing (4.56) $-\Gamma$(4.55) which yields (3.9). Alternatively, A_c can be obtained by computing (4.59) $+ G$(4.58). As mentioned in § 3, (3.9) is valid since $G^*: \mathbb{R}^{n_c} \to \mathscr{D}(A)$ and A_c^T is given by (3.26).

Next we substitute the expressions for A_c and A_c^T into (4.55), (4.56), (4.58) and (4.59) and compute the relations (4.55)G, G^*(4.56)G, $-$(4.58)Γ and Γ^*(4.59)Γ to obtain, respectively,

$$0 = [A_P\hat{Q} + \hat{Q}A_P^* + Q\bar{\Sigma}Q]\tau^*, \tag{4.60}$$

$$0 = \tau[A_P\hat{Q} + \hat{Q}A_P^* + Q\bar{\Sigma}Q]\tau^*, \tag{4.61}$$

$$0 = [A_Q^*\hat{P} + \hat{P}A_Q + P\Sigma P]\tau, \tag{4.62}$$

$$0 = \tau^*[A_Q^*\hat{P} + \hat{P}A_Q + P\Sigma P]\tau. \tag{4.63}$$

Note that (4.60)-(4.63) are *equivalent* to (4.55), (4.56), (4.58) and (4.59) since G and Γ have full rank. Since $(4.61) = \tau(4.60)$ and $(4.63) = \tau^*(4.62)$, (4.61) and (4.63) are superfluous and can be omitted. Thus we have derived (3.17) and (3.18).

To obtain (3.15) and (3.16) we need only compute the relations $(4.54) + \tau(4.60) - (4.60) - (4.60)^*$ and $(4.57) + \tau^*(4.62) - (4.62) - (4.62)^*$ and use (4.41).

Finally, to show that the preceding development entails no loss of generality in the optimality conditions we now use (3.9)-(3.18) to obtain (4.7)-(4.10) and (4.27)-(4.29). Let A_c, B_c, C_c, G, Γ, τ, Q, P, $\hat{Q}$, $\hat{P}$ be as in the theorem statement and define Q_1, Q_{12}, Q_2, P_1, P_{12}, P_2 by (4.45)-(4.47). Note that (3.12) and (3.13) imply (4.34) and (4.35) and hence (4.7) and (4.9). Using (3.8), (3.10), (3.11) and (3.22) it is easy to verify (4.27)-(4.29). Finally, substitute (4.32), (4.33) and (4.36) into (3.15)-(3.18), reverse the steps taken earlier in the proof and use (3.9)-(3.11) to obtain (4.8) and (4.10), which completes the proof. □

5. Concluding remarks. This paper has considered the problem of quadratically optimal, steady-state, fixed-order dynamic compensation for linear infinite-dimensional systems. The Main Theorem presents the stationarity conditions of the optimization problem in a highly simplified and rigorous form. The "optimal projection equations" (3.15)-(3.18) (or, equivalently, (3.27)-(3.30)) of the Main Theorem reveal the essential structure of the first-order necessary conditions and display the central role played by the optimal projection τ. The relationship of the Main Theorem to the standard finite-dimensional steady-state *LQG* problem can be demonstrated by replacing τ with the identity matrix and noting that (3.27) and (3.28) reduce immediately to the familiar pair of operator Riccati equations and that (3.29) and (3.30) yield the controllability and observability gramians of the controller.

Inasmuch as the Main Theorem is a fundamental generalization of classical steady-state *LQG* theory, a number of issues must be reexamined. Hence, in conclusion we should like to point out some possible extensions of the Main Theorem along with directions for further research.

1. *Sufficiency theory.* Although sufficient conditions for the existence of an optimal compensator were not investigated in this paper, auxiliary conditions based upon the structure of (3.15)-(3.18) could perhaps be imposed upon Q, P, $\hat{Q}$ and $\hat{P}$ to single out the global optimum from amongst the local minima. This would be similar to the situation in *LQG* theory where, under stabilizability and detectability hypotheses, optimal stabilizing Q and P are identified as the unique nonnegative-definite solutions of the pair of algebraic Riccati equations.

2. *Stabilizability.* Just as in the full-order *LQG* problem, one would expect a natural relationship between the structure of the optimal solution and stabilizability/detectability hypotheses. The results of [41], [42] and [68] could serve as a starting point in this regard.

3. *Numerical algorithms.* In practical situations, the distributed parameter system would be replaced by a high-order discretized model for which the matrix version (rather than the operator version) of the optimal projection equations could be solved numerically. A numerical algorithm for solving the matrix version of the optimal projection equations has been developed in [32] and [34]. The proposed computational scheme is fundamentally quite different from gradient search algorithms [17], [18], [21], [22], [24], [25], [28], [30] in that it operates through direct solution of the optimal projection equations by iterative refinement of the optimal projection.

4. *Convergence.* One of the principal uses for the optimal projection equations will be to understand the relationship between fixed-order dynamic-compensator

designs which are optimal with respect to approximate models and the optimal fixed-order dynamic compensator for the distributed parameter system itself. By considering a sequence of nth-order approximate models which converge to the distributed parameter system, conditions would be sought guaranteeing that the sequence of fixed-order compensators based on each approximate model approach the optimal dynamic compensator based upon the distributed parameter system (see [38]-[40]). This approach is analogous to the convergence results obtained in [7], [8] with the major difference being that the optimal projection equations permit the order of the compensator to remain fixed in accordance with real-world implementation constraints whereas in [7]-[9] the order of the compensator increases without bound.

5. *Unbounded control and observation.* An important generalization of the problem considered in this paper involves the case in which the input and output operators B and C are unbounded. The mathematical details for this problem are considerably more complex (see, e.g., [69]).

6. Singular observation noise/singular control weighting. As pointed out in [22], [33], [36] the assumptions of nonsingular control weighting and nonsingular observation noise preclude the use of direct output feedback as in

$$u(t) = C_c x_c(t) + D_c y(t) \tag{5.1}$$

since J is undefined unless

$$\operatorname{tr}[D_c^T R_2 D_c V_2] = 0 (\Leftrightarrow R_2 D_c V_2 = 0).$$

Although with due attention to (5.1) direct output feedback can be used in the singular case, the nature of the problem forebodes all of the difficulties associated with the singular *LQG* problem. Note that the deterministic output feedback problem [70], when viewed in this context, is highly singular.

7. *Discrete-time system/discrete-time compensator.* Digital implementation can be modelled by a discrete-time compensator with control of a continuous-time system facilitated by sampling and reconstruction devices. See [71], [73] for results in this direction.

8. *Cross weighting/correlated disturbance and observation noise.* This extension is straightforward and entirely analogous to the *LQG* case (see, e.g., [18, p. 351]).

Acknowledgments. We wish to thank Ardeth P. Grant for excellent and careful typing of the original manuscript version of this paper. Word processing and revision support was provided by Harris Corporation, GASD.

REFERENCES

[1] M. J. BALAS, *Toward a more practical control theory for distributed parameter systems*, in Control and Dynamic Systems, C. T. Leondes, ed., Advances in Theory and Applications, 19, Academic Press, New York, 1982.

[2] M. ATHANS, *Toward a practical theory of distributed parameter systems*, IEEE Trans. Automat. Control., AC-15 (1970), pp. 245-247.

[3] S. A. REIBLE, *Acoustoelectric convolver technology for spread-spectrum communications*, IEEE Trans. Microwave Theory Tech., MTT-29 (1981), pp. 463-473.

[4] R. F. CURTAIN AND A. J. PRITCHARD, *Infinite Dimensional Linear Systems Theory*, Springer-Verlag, New York, 1978.

[5] J. S. GIBSON, *The Riccati integral equations for optimal control problems on Hilbert spaces*, this Journal, 17 (1979), pp. 537-565.

[6] A. V. BALAKRISHNAN, *Applied Functional Analysis*, Springer-Verlag, New York, 1981.
[7] J. S. GIBSON, *An analysis of optimal modal regulation: convergence and stability*, this Journal, 19 (1981), pp. 686-707.
[8] ———, *Linear-quadratic optimal control of hereditary differential systems: infinite dimensional Riccati equations and numerical approximations*, this Journal, 21 (1983), pp. 95-139.
[8a] H. T. BANKS AND K. KUNISCH, *The linear regulator problem for parabolic systems*, this Journal, 22 (1984), pp. 684-698.
[9] H. T. BANKS, I. G. ROSEN AND K. ITO, *A spline based technique for computing Riccati operators and feedback controls in regulator problems for delay equations*, SIAM J. Sci. Stat. Comput., 5 (1984), pp. 830-855.
[10] M. AOKI, *Control of large-scale dynamic systems by aggregation*, IEEE Trans. Automat. Control, AC-13 (1968), pp. 246-253.
[11] D. A. WILSON, *Optimum solution of model-reduction problems*, Proc. IEE, 117 (1970), pp. 1161-1165.
[12] B. C. MOORE, *Principal component analysis in linear systems: controllability, observability, and model reduction*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 17-32.
[13] R. E. SKELTON AND A. YOUSUFF, *Component cost analysis of large scale systems*, in Control and Dynamic Systems, C. T. Leondes, ed., Academic Press, New York, 1982.
[14] E. A. JONCKHEERE AND L. M. SILVERMAN, *A new set of invariants for linear systems—application to reduced-order compensator design*, IEEE Trans. Automat. Control, AC-28 (1983), pp. 953-964.
[15] A. YOUSUFF AND R. E. SKELTON, *Controller reduction by component cost analysis*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 520-530.
[16] T. L. JOHNSON AND M. ATHANS, *On the design of optimal constrained dynamic compensators for linear constant systems*, IEEE Trans. Automat. Control, AC-15 (1970), pp. 658-660.
[17] W. S. LEVINE, T. L. JOHNSON AND M. ATHANS, *Optimal limited state variable feedback controllers for linear systems*, IEEE Trans. Automat. Control, AC-16 (1971), pp. 785-793.
[18] K. KWAKERNAAK AND R. SIVAN, *Linear Optimal Control Systems*, Wiley-Interscience, New York, 1972.
[19] D. B. ROM AND P. E. SARACHIK, *The design of optimal compensators for linear constant systems with inaccessible states*, IEEE Trans. Automat. Control, AC-18 (1973), pp. 509-512.
[20] M. SIDAR AND B.-Z. KURTARAN, *Optimal low-order controllers for linear stochastic systems*, Int. J. Control, 22 (1975), pp. 377-387.
[21] J. M. MENDEL AND J. FEATHER, *On the design of optimal time-invariant compensators for linear stochastic time-invariant systems*, IEEE Trans. Automat. Control, AC-20 (1975), pp. 653-657.
[22] S. BASUTHAKUR AND C. H. KNAPP, *Optimal constant controllers for stochastic linear systems*, IEEE Trans. Automat. Control, AC-20 (1975), pp. 664-666.
[23] R. B. ASHER AND J. C. DURRETT, *Linear discrete stochastic control with a reduced-order dynamic compensator*, IEEE Trans. Automat. Control, AC-21 (1976), pp. 626-627.
[24] W. J. NAEIJE AND O. H. BOSGRA, *The design of dynamic compensators for linear multivariable systems*, 1977 IFAC, Fredericton, New Brunswick, Canada, pp. 205-212.
[25] H. R. SIRISENA AND S. S. CHOI, *Design of optimal constrained dynamic compensators for non-stationary linear stochastic systems*, Int. J. Contr., 25 (1977), pp. 513-524.
[26] P. J. BLANVILLAIN AND T. L. JOHNSON, *Specific-optimal control with a dual minimal-order observer-based compensator*, Int. J. Contr., 28 (1978), pp. 277-294.
[27] ———, *Invariants of optimal minimal-order observer-based compensators*, IEEE Trans. Automat. Control, AC-23 (1978), pp. 473-474.
[28] C. J. WENK AND C. H. KNAPP, *Parameter optimization in linear systems with arbitrarily constrained controller structure*, IEEE Trans. Automat. Control, AC-25 (1980), pp. 496-500.
[29] J. O'REILLY, *Optimal low-order feedback controllers for linear discrete-time systems*, in Control and Dynamic Systems 16. C. T. Leondes, ed., Academic Press, New York, 1980.
[30] D. P. LOOZE AND N. R. SANDELL, JR., *Gradient calculations for linear quadratic fixed control structure problems*, IEEE Trans. Automat. Control, AC-25 (1980), pp. 285-288.
[31] D. C. HYLAND, *Optimality conditions for fixed-order dynamic compensation of flexible spacecraft with uncertain parameters*, AIAA 20th Aerospace Sciences Mtg., paper 82-0312, Orlando, FL, Jan. 1982.
[32] ———, *The optimal projection approach to fixed-order compensation: Numerical methods and illustrative results*, AIAA 21st Aerospace Sciences Mtg., paper 83-0303, Reno, NV, Jan. 1983.
[33] D. C. HYLAND AND D. S. BERNSTEIN, *Explicit optimality conditions for fixed-order dynamic compensation*, Proc. IEEE Conference on Decision and Control, San Antonio, TX, Dec. 1983, pp. 161-165.
[34] D. C. HYLAND, *Comparison of various controller-reduction methods: Suboptimal versus optimal projection*, Proc. AIAA Dynamics Specialists Conference, Palm Springs, CA, May 1984, pp. 382-389.

[35] D. S. BERNSTEIN AND D. C. HYLAND, *The optimal projection equations for fixed-order dynamic compensation of distributed parameter systems*, Proc. AIAA Dynamics Specialists Conference, Palm Springs, CA, May 1984, pp. 396-400.

[36] ———, *The optimal projection equations for fixed-order dynamic compensation*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 1034-1037.

[37] ———, *The optimal projection approach to model reduction and the relationship between the methods of Wilson and Moore*, Proc. IEEE Conference on Decision and Control, Las Vegas, NV, Dec. 1984, pp. 120-126.

[37a] ———, *The optimal projection equations for model reduction and the relationships among the methods of Wilson, Skelton and Moore*, to appear.

[38] T. L. JOHNSON, *Optimization of low order compensators for infinite dimensional systems*, Proc. 9th IFIP Symposium on Optimization Techniques, Warsaw, Poland, September 1979, pp. 394-401.

[39] R. K. PEARSON, *Optimal fixed-form compensators for large space structures*, in ACOSS SIX (Active Control of Space Structures), RADC-TR-81-289, Final Technical Report, Rome Air Development Center, Griffiss AFB, New York, 1981.

[40] ———, *Optimal velocity feedback control of flexible structures*, Ph.D. dissertation, Dept. Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, 1982.

[41] R. F. CURTAIN, *Compensators for infinite-dimensional linear systems*, J. Franklin Inst., 315 (1983), pp. 331-346.

[42] J. M. SCHUMACHER, *A direct approach to compensator design for distributed parameter systems*, this Journal, 21 (1983), pp. 823-836.

[43] D. L. RUSSELL, *Linear stabilization of the linear oscillator in Hilbert space*, J. Math. Anal. Appl., 25 (1969), pp. 663-675.

[44] ———, *Decay rates for weakly damped systems in Hilbert space obtained with control theoretic methods*, J. Differential Equations, 19 (1975), pp. 344-370.

[45] M. J. BALAS, *Modal control of certain flexible dynamic systems*, this Journal, 16 (1978), pp. 450-462.

[46] ———, *Feedback control of flexible systems*, IEEE Trans. Automat. Control, AC-23 (1978), pp. 673-679.

[47] J. S. GIBSON, *A note on stabilization of infinite dimensional linear oscillators by compact feedback*, this Journal, 18 (1980), pp. 311-316.

[48] M. J. BALAS, *Trends in large space structure control theory: Fondest hopes, wildest dreams*, IEEE Trans. Automat. Control, AC-24 (1982), pp. 522-535.

[49] T. L. JOHNSON, *Progress in modelling and control of flexible spacecraft*, J. Franklin Inst., 315 (1983), pp. 495-520.

[49a] M. J. BALAS, *Feedback control of dissipative hyperbolic distributed parameter systems with finite dimensional controllers*, J. Math. Anal. Appl., 98 (1984), pp. 1-24.

[50] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.

[51] J. R. RINGROSE, *Compact Non-Self-Adjoint Operators*, Van Nostrand Reinhold, London, 1971.

[52] I. C. GOHBERG AND M. G. KREIN, *Introduction to the Theory of Linear Nonselfadjoint Operators*, Translations of Mathematical Monographs, Vol. 18, American Mathematical Society, Providence, RI, 1966.

[52a] M. S. BRODSKII, *Triangular and Jordan Representations of Linear Operators*, Translations of Mathematical Monographs, Vol. 32, American Mathematical Society, Providence, RI, 1971.

[53] I. GOHBERG AND S. GOLDBERG, *Basic Operator Theory*, Birkhauser, Boston, 1981.

[54] F. R. GANTMACHER, *The Theory of Matrices*, Vol. I, Chelsea, New York, 1977.

[55] C. R. RAO AND S. K. MITRA, *Generalized Inverse of Matrices and Its Applications*, John Wiley, New York, 1971.

[56] B. NOBLE AND J. W. DANIEL, *Applied Linear Algebra*, Second edition, Prentice-Hall, Englewood Cliffs, NJ, 1977.

[57] R. F. CURTAIN AND A. J. PRITCHARD, *Functional Analysis in Modern Applied Mathematics*, Academic Press, London, 1977.

[58] S. CHAKRABARTI, B. B. BATTACHARYYA AND M. N. S. SWAMY, *On simultaneous diagonalization of a collection of hermitian matrices*, Matrix and Tensor Quart., 29 (1978), pp. 35-54.

[59] C. T. MULLIS AND R. A. ROBERTS, *Synthesis of minimum roundoff noise fixed point digital filters*, IEEE Trans. Circ. Syst., CAS-23 (1976), pp. 551-562.

[60] A. J. LAUB, *Computation of balancing transformation*, Proc. 1980 Joint Automation Control Conference, San Francisco, CA, Aug. 1980.

[61] E. JONCKHEERE, *Open-loop and closed loop approximations of linear systems and associated balanced realizations*, 1982 Symposium on Circuits and Systems, Rome, May 1982.

[62] W. M. WONHAM, *Linear Multivariable Control: A Geometric Approach*, Springer-Verlag, New York, 1974.

[63] D. C. LAY, *Spectral properties of generalized inverses of linear operators*, SIAM J. Appl. Math., 29 (1975), pp. 103–109.

[64] S. L. CAMPBELL AND C. D. MEYER, JR., *Generalized Inverses of Linear Transformations*, Pitman, London, 1979.

[65] P. ROBERT, *On the group-inverse of a linear transformation*, J. Math. Anal. Appl., 22 (1968), pp. 658–669.

[66] A. ALBERT, *Conditions for positive and nonnegative definiteness in terms of pseudo inverse*, SIAM J. Appl. Math., 17 (1969), pp. 434–440.

[67] E. KREINDLER AND A. JAMESON, *Conditions for nonnegativeness of partitioned matrices*, IEEE Trans. Automat. Control, AC-17 (1972), pp. 147–148.

[68] C. N. NETT, C. A. JACOBSON AND M. J. BALAS, *Fractional representation theory: Robustness results with applications to finite dimensional control of a class of linear distributed systems*, IEEE Conference on Decision and Control, San Antonio, TX, Dec. 1983.

[69] R. F. CURTAIN, *Finite-dimensional compensators for parabolic distributed systems with unbounded control and observation*, this Journal, 22 (1984), pp. 255–276.

[70] W. S. LEVINE AND M. ATHANS, *On the determination of the optimal constant output feedback gains for linear multivariable systems*, IEEE Trans. Automat. Control, AC-15 (1970), pp. 44–48.

[71] M. J. BALAS, *The structure of discrete-time finite-dimensional control of distributed parameter systems*, Proc. IEEE International Large Scale Systems Symposium, Virginia Beach, VA, 1982.

[72] M. S. GHAVSI AND J. J. KELLY, *Introduction to Distributed-Parameter Networks with Application to Integrated Circuits*, Holt, Rinehart and Winston, New York, 1968.

[73] D. S. BERNSTEIN, L. D. DAVIS AND D. C. HYLAND, *The optimal projection equations for reduced-order, discrete-time modelling, extimation and control*, submitted for publication.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 1, January 1986

AN EXAMPLE ON THE EFFECT OF TIME DELAYS IN BOUNDARY FEEDBACK STABILIZATION OF WAVE EQUATIONS*

R. DATKO†, J. LAGNESE†‡ AND M. P. POLIS§‡

Abstract. This note is concerned with the effect of time delays in boundary feedback stabilization schemes for wave equations. The question to be addressed is whether such delays can destabilize a system which is uniformly asymptotically stable in the absence of delays.

Key words. boundary stabilization, delay systems

AMS(MOS) subject classification. 93D15

This note is concerned with the effect of time delays in boundary feedback stabilization schemes for wave equations. The question to be addressed is whether such delays can destabilize a system which is uniformly asymptotically stable in the absence of delays. It will be shown by example that for "almost" arbitrary delays such destabilization can indeed occur in certain otherwise stable boundary feedback schemes for both undamped and damped wave equations. Although the examples involve only one spatial dimension, it is to be expected that a similar phenomenon occurs in higher dimensions. This suggests that certain boundary stabilization schemes which have been proposed for various classes of hyperbolic systems [1]-[2], [5]-[7] may not be robust to the small delays which might well occur in computing feedback controls.

Consider the equation

$$u_{tt} - u_{xx} + 2au_t + a^2u = 0, \qquad 0 < x < 1, \quad t > 0, \tag{1}$$

with boundary conditions

$$u(0, t) = 0, \qquad t > 0, \tag{2}$$

$$u_x(1, t) = -ku_t(1, t), \qquad t > 0, \tag{3}$$

where $a \geqq 0$ and $k \geqq 0$. The question to be treated is the effect on the stability of (1)-(3) of a time delay in the right side of (3) when $k > 0$. If $k = 0$ and $a > 0$ Datko [3] has shown that the system (1)-(3) will be destabilized by a time delay in the velocity term of (1).

When $a > 0$, (1) contains both viscous damping and a restoring force proportional to displacement. If $k > 0$, (3) represents boundary damping in the system. It is known that (1)-(3) is uniformly asymptotically stable as long as $a^2 + k^2 > 0$ (see e.g. [1]) and that

$$E(u, t) \leqq C e^{-\alpha t} E(u, 0) \tag{4}$$

for some positive constant α, where E is the usual energy functional for (1) defined by

$$E(u, t) = \int_0^1 (u_t^2 + u_x^2 + a^2u^2)\, dx.$$

* Received by the editors November 12, 1984, and in revised form April 1, 1985.
† Department of Mathematics, Georgetown University, Washington, DC 20057.
‡ Currently on leave to the National Science Foundation, Washington, DC 20550.
§ Department of Electrical Engineering, Ecole Polytechnique de Montreal, Montreal, Quebec, Canada.

A necessary condition for (4) is that spectrum of (1)-(3) lie in a half-plane $\operatorname{Re}\omega \leqq \text{Const.} < 0$. The spectrum can be exhibited by setting $u = e^{\omega t}\phi(x)$. The eigenvalues ω and eigenfunctions ϕ are then determined by the problem

$$\phi''(x) - (a+\omega)^2\phi(x) = 0, \qquad 0 < x < 1, \tag{5}$$

$$\phi(0) = 0, \qquad \phi'(1) + k\omega\phi(1) = 0. \tag{6}$$

From (5), (6), the eigenvalues are the solutions of

$$e^{2(\omega+a)} = \frac{(k-1)\omega - a}{(k+1)\omega + a}. \tag{7}$$

If $a = 0$ and $k \neq 1$, (7) can be solved explicitly to yield

$$\operatorname{Re}\omega = \frac{1}{2}\log\left|\frac{k-1}{k+1}\right| < 0.$$

When $a = 0$ and $k = 1$, all solutions of (1)-(3) can be shown to vanish identically for $t > 2$. When $a \neq 0$ the solutions ω_n of (7) can be shown to lie in a half-plane $\operatorname{Re}\omega \leqq \alpha < 0$ and to satisfy $|\operatorname{Im}\omega_n| \to +\infty$.

Now let $\varepsilon > 0$ and suppose the boundary condition (3) is replaced by

$$u_x(1, t) = -ku_t(1, t-\varepsilon), \qquad t > \varepsilon. \tag{8}$$

The following result will be established.

THEOREM. *Let $K = e^{-2a}$. The system* (1), (2), (8) *has the following stability properties*:

(i) *If $0 < k < (1-K)/(1+K)$, for each $\varepsilon > 0$ there exists $\beta(\varepsilon) > 0$ such that the spectrum of the system lies in* $\operatorname{Re}\omega \leqq -\beta$.

(ii) *If $k = (1-K)/(1+K)$, for each $\varepsilon > 0$ the spectrum lies in* $\operatorname{Re}\omega < 0$, *but there is a countably dense set R in $(0, \infty)$ such that for each ε in R there is a sequence $\{\omega_n\}$ in the spectrum such that*

$$\lim_{n\to\infty} \operatorname{Re}\omega_n = 0.$$

(iii) *If $k > (1-K)/(1+K)$, there is dense open set D in $(0, \infty)$ such that for each ε in D the system admits exponentially unstable solutions.*

To prove the theorem, two lemmas will be needed. The first lemma is a special case of [4, Lemma 2.3].

LEMMA 1. *Let*

$$\begin{aligned} h(\varepsilon, \omega) &= \omega[1 + K e^{-2\omega} + k e^{-\varepsilon\omega}(1 - K e^{-2\omega})] + a(1 + K e^{-2\omega}) \\ &= \omega f(\varepsilon, \omega) + a(1 + K e^{-2\omega}). \end{aligned} \tag{9}$$

If, for fixed ε, $f(\varepsilon, \omega)$ has a zero at $\omega_0 = \xi_0 + i\eta_0$, then for any $\delta > 0$ the vertical strip $\{\omega: \xi_0 - \delta < \operatorname{Re}\omega < \xi_0 + \delta\}$ has an infinite number of zeros of both $f(\varepsilon, \omega)$ and $h(\varepsilon, \omega)$.

LEMMA 2. *Let $K = e^{-2a}$ and $k > (1-K)/(1+K)$. Then there exists for each such k an open dense set, $\mathcal{D}$, in $(0, \infty)$ such that for every ε in $\mathcal{D}$, $f(\varepsilon, \omega) = 0$ has at least one solution with* $\operatorname{Re}\omega > 0$.

Proof of Lemma 2. Let K and k satisfying the hypotheses of the lemma be fixed. Consider the mapping from the complex ω-plane into the complex ε-plane defined implicitly by

$$k = \frac{1 + K e^{-2\omega}}{K e^{-2\omega} - 1} e^{\varepsilon\omega}. \tag{10}$$

For $\omega \neq 0$ solutions of equation (10) can be found among the infinite family of meromorphic functions given by the equations

$$\varepsilon = \frac{1}{\omega}\left[\log k + \log\left(\frac{K e^{-2\omega} - 1}{K e^{-2\omega} + 1}\right) + 2m\pi i\right], \tag{11}$$

where log is the principal value of the logarithm (see e.g. [9]) and m is a positive integer.

Now let n be a fixed positive integer and

$$\omega = \omega_1 + \frac{(2n+1)}{2}\pi i, \qquad \omega_1 > 0. \tag{12}$$

Then ε in (11) has the form

$$\varepsilon = \frac{\log k + \log\left((K e^{-2\omega_1} + 1)/(1 - K e^{-2\omega_1})\right) + (2m+1)\pi i}{\omega_1 + ((2n+1)/2)\pi i}. \tag{13}$$

When ω_1 satisfies the equation

$$\omega_1 = \frac{2n+1}{2(2m+1)}\left[\log k + \log\left(\frac{K e^{-2\omega_1} + 1}{1 - K e^{-2\omega_1}}\right)\right], \tag{14}$$

equation (13) satisfies

$$\varepsilon = \frac{2(2m+1)}{2n+1}. \tag{15}$$

Equation (14) always has a solution for some $\omega_1 > 0$. To see this, notice that because $k > (1-K)/(1+K)$ the right side of (14) is positive for $\omega_1 = 0$ and as ω_1 tends to infinity the right side tends to $(2n+1)/(2(2m+1)) \log k$.

Next observe that points of the form (15) are dense on $(0, \infty)$. Furthermore because of the open mapping property of meromorphic functions (see e.g. [9, p. 116] about each point of the form (15) there is an open interval, with $2(2m+1)/(2n+1)$ as its center which is contained in the image, under the mapping (11), of some open ball in $\operatorname{Re} \omega > 0$. This completes the proof of the lemma.

Remark. It is tempting, and indeed it is probably correct, to state that the dense set $\mathcal{D}$ in Lemma 2 is $(0, \infty)$. However the proof of the lemma does not justify this statement.

Proof of the theorem. The spectrum of (1), (2), (8) is determined by the eigenvalue problem consisting of equation (5) together with the boundary condition $\phi(0) = 0$, $\phi'(1) + k\omega e^{-\varepsilon\omega}\phi(1) = 0$. The equation for the eigenvalues is therefore obtained by replacing k in (7) by $k e^{-\varepsilon\omega}$. The result is equivalent to

$$h(\varepsilon, \omega) = 0 \tag{16}$$

where h is defined by (9) and $K = e^{-2a}$.

(i) Suppose $k \leqq (1-K)/(1+K)$ and $\varepsilon > 0$. Since $\omega = 0$ is not a solution of (16), that equation can be rewritten as

$$1 = k e^{-\varepsilon\omega}\frac{K e^{-2\omega} - 1}{K e^{-2\omega} + 1} - \frac{a}{\omega}. \tag{17}$$

If (17) has a solution $\omega = \xi + i\eta$ with $\xi > 0$, then

$$1 = k \operatorname{Re}\left[e^{-\varepsilon\omega}\frac{K e^{-2\omega} - 1}{K e^{-2\omega} + 1}\right] - \frac{a\xi}{\xi^2 + \eta^2} < k\frac{1 + K e^{-2\xi}}{1 - K e^{-2\xi}} < k\frac{1+K}{1-K} \leqq 1 \tag{18}$$

which is a contradiction. Thus (16) has no solution with $\operatorname{Re}\omega > 0$, for any $\varepsilon > 0$, provided $k \leqq (1-K)/(1+K)$. Also, if $k < (1-K)/(1+K)$ the string of inequalities in (18) again leads to a contradiction whenever $\xi = \operatorname{Re}\omega \geqq 0$, so that all zeros of (16) must satisfy $\operatorname{Re}\omega < 0$ when $k < (1-K)/(1+K)$. Moreover, in this case it is not possible for a sequence of such zeros to accumulate at the imaginary axis. For suppose there were a sequence $\omega_n = \xi_n + i\eta_n$ of zeros such that $\lim \xi_n = 0$, $\underline{\lim}\,|\eta_n| > 0$. From (17)

$$1 \leqq \overline{\lim}\left\{k \operatorname{Re}\left[e^{-\varepsilon\omega_n}\frac{K e^{-2\omega_n}-1}{K e^{-2\omega_n}+1}\right] - \frac{a\xi_n}{\xi_n^2+\eta_n^2}\right\}$$

$$\leqq \lim k\frac{1+K e^{-2\xi_n}}{1-K e^{-2\xi_n}} = k\frac{1+K}{1-K} < 1,$$

a contradiction. Thus for $\varepsilon > 0$ and $k < (1-K)/(1+K)$, the spectrum of (1), (2), (8) must lie in a half-plane $\operatorname{Re}\omega \leqq -\beta$, $\beta > 0$.

(ii) From the proof of (i), if $k = (1-K)/(1+K)$ and $\varepsilon > 0$ the spectrum lies in $\operatorname{Re}\omega \leqq 0$. If (16) has a zero $\omega = i\eta$, $\eta \neq 0$, then from (17)

$$1 + \frac{a}{i\eta} = k e^{-i\varepsilon\eta}\left(\frac{K e^{-2i\eta}-1}{K e^{-2i\eta}+1}\right). \tag{19}$$

Taking the modulus of each side of (19) results in the contradiction

$$1 + \frac{a^2}{\eta^2} = k^2\frac{1+K^2-2K\cos 2\eta}{1+K^2+2K\cos 2\eta} \leqq k^2\left(\frac{1+K}{1-K}\right)^2 = 1.$$

However, it is easily seen that

$$f\left(\frac{2(2m+1)}{2n+1}, \frac{(2n+1)\pi i}{2}\right) = 0$$

where f is defined in (9) and m, n are arbitrary. Thus, by Lemma 1, given any $\delta > 0$ the vertical strip $\{\omega: -\delta < \operatorname{Re}\omega < 0\}$ contains an infinite number of points of the spectrum of (1), (2), (8).

(iii) Let $k > (1-K)/(1+K)$. By Lemmas 1 and 2 it follows that for each such k there is an open dense set, $\mathcal{D}$, in $(0, \infty)$ such that (16) is satisfied for each ε in $\mathcal{D}$ and some ω with $\operatorname{Re}\omega > 0$. This completes the proof.

Remark. One possible interpretation of the destabilizing effect of arbitrarily small time delays is that a delay ε will excite a high frequency mode (i.e., a mode with frequency $\approx 1/\varepsilon$) by causing the control force to be in phase rather than out of phase with the velocity of the mode in question. That is, time delays cause phase shifts in the control force which can have the effect of exciting rather than damping the high frequency modes of the system.

REFERENCES

[1] G. CHEN, *Energy decay estimates and exact boundary value controllability for the wave equation in a bounded domain*, J. Math. Pures Appl., 58 (1979), pp. 249-274.

[2] ———, *A note on boundary stabilization of the wave equation*, this Journal, 19 (1981), pp. 106-113.

[3] R. DATKO, *Representation of solutions and stability of linear differential-difference equations in a Banach space*, J. Differential Equations, 29 (1978), pp. 105-166.

[4] ———, *A procedure for determination of the exponential stability of certain differential-difference equations*, Quart. Appl. Math., 36 (1978), pp. 279-292.

[5] J. LAGNESE, *Decay of solutions of wave equations in a bounded region with boundary dissipation*, J. Differential Equations, 50 (1983), pp. 163–182.
[6] ———, *Boundary stabilization of linear elastodynamic systems*, this Journal, 21 (1983), pp. 968–984.
[7] J. P. QUINN AND D. L. RUSSELL, *Asymptotic stability and energy decay rates for solutions of hyperbolic equations with boundary damping*, Proc. Royal Soc. Edinburgh, 77A (1977), pp. 97–127.
[8] S. SAKS AND A. ZYGMUND, *Analytic Functions*, Elsevier Publishing Company, Amsterdam, 1971.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 1, January, 1986

008

OPTIMAL INTERPOLATION WITH CONVEX SPLINES OF SECOND DEGREE*

LAKSHMAN S. THAKUR†

Abstract. Estimation of a convex function interpolating its known values and satisfying certain smoothness properties is needed in some applications and has been investigated in many studies from various perspectives. Without the convexity assumption, Karlin's theorem characterizes the solution to the problem studied here while under convexity, the analogous result is due to Smith and Ward. The aim of this paper is to give a convex programming characterization that can be used to calculate an optimal spline which solves a given problem of degree 2. Specifically, our aim is to determine a smooth ($f, f^{(1)}$ absolutely continuous) convex spline f of second degree, which interpolates $(r+2)$ given points in $[a, b]$, and minimizes the Tchebycheff norm $\|f^{(2)}\|_\infty$, with $f^{(2)}$ essentially bounded in $[a, b]$. The convexity of the formulation enables us to calculate an optimal spline using widely available computer routines for nonlinear optimization. The approach is illustrated by providing the convex programming formulations and the computer-obtained optimal solutions for two numerical examples.

Key words. convex programming, function estimation, spline functions, convex functions, convex splines, optimal interpolation

AMS(MOS) subject classifications. Primary 26A51; secondary 41A15, 90C25

1. Introduction.

Let $F_\infty^{(n)}[a, b]$ be a subset of the real Sobolev space

$$W_\infty^{(n)}[a, b] = \{f \in C^{(n-1)}[a, b] \mid f^{(n-1)} \text{ abs.cont.}; f^{(n)} \in L_\infty[a, b]\}$$

defined by

$$F_\infty^{(n)}[a, b] = \{f \in W_\infty^{(n)}[a, b] \mid f \text{ convex}; f(x_i) = y_i, i = 1, \cdots, n+r\}$$

where $\{x_i\}_1^{n+r}, \{y_i\}_1^{n+r} \in R^{n+r}$; $r \geqq 1$; $\{x_i\}_1^{n+r}$ is a strictly increasing sequence with $x_1 = a$, $x_{n+r} = b$, $x_i \in [a, b]$ for all i. Then the problem of *degree n* can be defined as

$$\text{Minimize } \{\|f^{(n)}\|_\infty : f \in F_\infty^{(n)}[a, b]\}. \tag{1}$$

Without the convexity condition on the functions, the work of Karlin [23], [24] and others [9], [14], [15] has established the existence of a perfect spline solution of (1). Under the convexity constraint, Smith and Ward [43] have given a natural analogue (given in § 3, for $n = 2$) of Karlin's theorem.

For $n = 2$, the case treated in this paper, Iliev and Pollul [22] have independently used a similar approach and have shown that the problem has a quadratic spline solution, characterized by the existence of a core interval [15] where all solutions must be the same and where the second derivative of the solution is the positive part of a perfect spline. Though not giving any computational results, they also state briefly an algorithm to find an optimal solution of (1). However, it uses their Lemma 3 and Lemma 4 [22, pp. 52-53]—("with Lemma 3a), 3b) define . . . , with the help of Lemma 4 . . . try to decrease . . ." [22, p. 55])—which involve 21 different graphs given in their descriptions, several graphs having 3-4 different cases in turn. Thus some 20-30 different situations may have to be considered, and the algorithm, as given, does not seem to be easily implementable.

* Received by the editors August 20, 1981, and in final revised form April 29, 1985.

† Yale School of Management, Yale University, New Haven, Connecticut 06520. On leave from Department of Management Science, Shippensburg University, Shippensburg, Pennsylvania 17257.

Two further remarks about this work also deserve mention. First, their Theorem 3, giving the number of knots of the quadratic spline solution, (a) is not the tightest $(r+2)$ knots, vs. $(r-1)$ given in Theorem 2 here); and (b) does not assert the "perfectness" of the spline, as is done in our theorem. When "perfectness" is enforced in Remark 6, the knot count jumps to a much larger value: $(2r+2)$. Theorem 2 here combines the tightest knot count with "perfectness" and presents a much stronger result. Second, [22] considers the data which is *strictly* convex ($d_1 < d_2 < \cdots < d_{r+1}$, where $d_i = (y_{i+1} - y_i)/(x_{i+1} - x_i)$, $i = 1, \cdots, r+1$) whereas we assume only nondecreasing d_i's: $d_1 \leqq d_2 \leqq \cdots \leqq d_{r+1}$ (see Lemma 2 and Theorem 1). Inclusion of this data may require some changes in the analysis of [22].

In this paper, for degree $n = 2$, a concrete nonlinear convex programming formulation of the problem is given, enabling us to determine an optimal convex spline by the use of easily available computer-based optimization techniques [26], [42], [48].

There is a great deal of other work on related problems. Therefore, a brief review may be appropriate in order to differentiate the results presented here from this work. The main differences arise from the numerous combinations of differences in (i) problem definition, (ii) the norms considered, (iii) the emphasis on existence and characterization rather than computation, (iv) the emphasis on approximation rather than interpolation, (v) the lack of the convexity requirement, or (vi) the methods used in the computational scheme. We will refer to the most pertinent literature and point out these differences.

In [49] optimal convex splines are characterized but the norm considered is L_2-norm, and the suggested quadratic programming approach, based mainly on [25] (and [39], [46], [47]), is for the L_2-norm. The joint work of Passow, Roulier, and McAllister [30], [32], [38] (and the references there) on the shape preserving, monotone, and convex interpolating splines, deals with the existence and construction of such splines with given, prechosen smoothness, with no attempt to minimize $\|f^{(2)}\|_\infty$, as is the case with our problem. Regarding the lower order derivatives, in [40] Roulier and McAllister present estimates of $\max_{x_1 \leqq x \leqq x_{r+2}} |f^{(j)}(x) - s^{(j)}(x)|$, $j = 0, 1$, where s is the spline produced by the algorithm in [32]. Such results as in [34], [35], [36], [37], and [50] require the monotonicity of the spline, but not its convexity. Therefore, they are related but different. In [2], [3], [16], and [41], where conditions on derivatives (which can be used to imply convexity) are considered, the emphasis is on the question of the existence of certain polynomials (not splines) and on the degree of approximation.

Among the papers on "optimal recovery" of smooth functions, parallel results given in [6], [17], and [33] deal with the related problem of finding the smallest interval of possible values of $f(x_0)$, when a point x_0, the function values $f(x_1), \cdots, f(x_m)$, and a bound on $\|f^{(k)}\|_\infty$ $(k \leqq m)$ is given. The computational routines for optimal recovery (based on Newton's method) are given in [10], and [11]. Spline computations involve numerical evaluation of B-splines, which can be carried out by the use of divided differences, or by a more stable, efficient (and essentially the same) method presented in [7], [8], and [27].

Thus, many related problems have been studied in the literature. The use of mathematical programming techniques in the study of an optimal spline is also not new. For example, linear programming is used for best spline function approximation in [13], and for a discrete approximation problem with constraints in [1]. The Kuhn-Tucker conditions are used in [28], cubic and bicubic spline interpolation is studied via dynamic programming in [4], [5], and the use of quadratic programming in [25] has already been mentioned.

Finally, the literature dealing with the application of an optimal convex spline should be mentioned. It is easy to show (use [44, p. 706]) that if $\hat{f}$ is the piecewise linear function obtained by connecting the adjacent points $\{(x_i, y_i)\}_1^{r+2}$, then a solution f^* of (1) for $n=2$ gives the following bound E_{f^*} on the function error

$$\max_{a\leqq x\leqq b} |f^*(x)-\hat{f}(x)| \leqq E_{f^*} = K_{f^*}\delta^2/8$$

where $K_{f^*} = \|f^{*(2)}\|_\infty$, $\delta = \max_{1\leqq i\leqq r+1}(x_{i+1}-x_i)$. And, since we obviously have $E_{f^*} \leqq E_f \ \forall f \in F_\infty^{(2)}[a, b]$, one specific use of such a solution f^* is in investigating minimal error bounds on piecewise linear approximations in the error analysis of convex separable programs [44], [45]. But the major impetus to this paper, besides the perfect spline literature cited above, is provided by the same underlying motivations of such other work as: a quadratic programming formulation for obtaining a nonnegative, nondecreasing, or convex *piecewise linear* function minimizing the *least-square* norm [12]; its recent generalization to n variables via a variant of generalized programming formulation [18]; and an algorithm where the function is assumed, in addition, to have a polynomial form [21].

2. Preliminary analysis. If we define $G_\infty^{(n)}[a, b]$ for degree n, and p given points $\{x_i\}_1^p$, $\{y_i\}_1^p \in R^p$, $p \geqq n+1$ as

$$G_\infty^{(n)}[a, b] = \Big\{ g(x) \in W_\infty^{(n)}[a, b] \,\Big|\, g(x) \text{ nondecreasing in } [a, b];$$
$$\int_{x_i}^{x_{i+1}} g(x)\,dx = (y_{i+1}-y_i),\ i=1, \cdots, p-1 \Big\},$$

then it is easy to verify that for $n=2$, a solution of (1), if it exists for the given data, can be obtained by integrating the solution of

$$\text{Minimize } \{\|g^{(1)}\|_\infty : g(x) \in G_\infty^{(1)}[a, b]\}. \tag{2}$$

If $g^*(x)$ is the solution of (2), a solution $f^*(x)$ of (1) for $n=2$ is given by $f^*(x) = \int_{x_1}^x g^*(x)\,dx + y_1$. In addition, $\|f^{*(2)}\|_\infty = \|g^{*(1)}\|_\infty$. Now problem (2) is simpler than (1), in that it can be attacked directly as shown below. We begin by obtaining the optimal solution of a problem closely related to (2) in a single interval, i.e., for $p=2$.

LEMMA 1. *For degree $n=1$ and $p=2$ points $\{x_i\}_1^2$, $\{y_i\}_1^2 \in R^2$, $x_1 < x_2$, let*

$$\Omega = \{h(x) \in G_\infty^{(1)}[a, b] \,|\, h(x_1) = z_1,\ h(x_2) = z_2\}. \tag{3}$$

Then: (i) *There exists a nondecreasing continuous function $g(x) \in \Omega$ iff* (a) $(x_2-x_1)z_1 = (y_2-y_1) = (x_2-x_1)z_2$, *if $z_2 = z_1$, or* (b) $(x_2-x_1)z_1 < (y_2-y_1) < (x_2-x_1)z_2$ *if $z_2 > z_1$.*

(ii) *The value of* $\min\{\|g^{(1)}\|_\infty : g(x) \in \Omega\}$ *is given by* $\|g^{*(1)}\|_\infty$ *defined as follows* (*when* (i)(a) *or* (i)(b) *above is satisfied*):

$$\|g^{*(1)}\|_\infty = \begin{cases} 0 & \text{if } s_1 = s_2 = 0, \\ (s_1+s_2)^2/4s_1 & \text{if } 0 < s_1 \leqq s_2, \\ (s_1+s_2)^2/4s_2 & \text{if } s_1 \geqq s_2 > 0, \end{cases} \tag{4}$$

where $\Delta x = x_2-x_1$, $\Delta y = y_2-y_1$, $\Delta z = z_2-z_1$, $s_1 = 2[(\Delta y/\Delta x)-z_1]/\Delta x$, $s_2 = 2[z_2-(\Delta y/\Delta x)]/\Delta x$.

Proof. (i) *Necessity.* Due to the the nondecreasing property, we cannot have $z_2 < z_1$. If $z_2 = z_1$, the nondecreasing property implies that $g(x)$ is constant in $[x_1, x_2]$: $g(x) = z_1 = z_2$, and (a) follows immediately. For $z_2 > z_1$, let us take an arbitrary $g(x) \in$

$G_\infty^{(1)}[a, b]$. Since $g(x)$ is nondecreasing, we have $\min\{g(x): x \in [x_1, x_2]\} = g(x_1) = z_1$, and we get $(y_2 - y_1) = \int_{x_1}^{x_2} g(x)\,dx \geqq \int_{x_1}^{x_2} z_1\,dx = (x_2 - x_1)z_1$. But the center *equality* is impossible since it leads to a contradiction as follows. $\int_{x_1}^{x_2} g(x)\,dx = \int_{x_1}^{x_2} z_1\,dx$ implies $\int_{x_1}^{x_2} (g(x) - z_1)\,dx = 0$, and since $g(x)$ is nondecreasing (thus $(g(x) - z_1) \geqq 0$), this means $g(x) - z_1 = 0$ a.e. in $[x_1, x_2]$. Now by the continuity of $g(x)$, this implies that $g(x)$ is constant in $[x_1, x_2]$, giving $g(x_1) = z_1 = g(x_2) = z_2$, which is contrary to our assumption $z_2 > z_1$. Thus $(y_2 - y_1) > (x_2 - x_1)z_1$, the left inequality in (b). Similarly, considering $\max\{g(x): x \in [x_1, x_2]\} = g(x_2) = z_2$, we can show the right inequality: $(y_2 - y_1) < (x_2 - x_1)z_2$, completing the proof of (b).

Sufficiency. (a), the case for $z_2 = z_1$, is obvious since the function $g^*(x) = (y_2 - y_1)/(x_2 - x_1)$, $x_1 \leqq x \leqq x_2$ satisfies all the requirements. For (b), the $z_2 > z_1$ case, it is lengthy but easy to construct a nondecreasing continuous function $g^*(x)$, given below, and to check that it belongs to Ω.

If $s_1 < s_2$:

$$g^*(x) = \begin{cases} z_1, & x_1 \leqq x \leqq \tilde{x}, \\ l_1(x), & \tilde{x} \leqq x \leqq x_2, \end{cases} \tag{5}$$

where $l_1(x)$ is defined by $l_1(\tilde{x}) = z_1$, $l_1(x_2) = z_2$, $l_1^{(1)}(x) = \Delta z/(x_2 - \tilde{x})\ \forall x \in R$, and $\tilde{x} = x_2 - [2\Delta x s_1/(s_1 + s_2)]$.

If $s_1 \geqq s_2$:

$$g^*(x) = \begin{cases} l_2(x), & x_1 \leqq x \leqq \tilde{x}, \\ z_2, & \tilde{x} \leqq x \leqq x_2, \end{cases} \tag{6}$$

where $l_2(x)$ is defined by $l_2(x_1) = z_1$, $l_2(\tilde{x}) = z_2$, $l_2^{(1)}(x) = \Delta z/(\tilde{x} - x_1)\ \forall x \in R$, and $\tilde{x} = (2x_2 - x_1) - [2\Delta x s_1/(s_1 + s_2)]$.

(ii) By noting that $\Delta z/(x_2 - \tilde{x}) = (s_1 + s_2)^2/4s_1$, $\Delta z/(\tilde{x} - x_1) = (s_1 + s_2)^2/4s_2$, we can check directly that $\|g^{*(1)}\|_\infty$ is given by (4) for the functions defined by (5) and (6). What remains to be shown is that $\|g^{*(1)}\|_\infty$ is $\min\{\|g^{(1)}\|_\infty: g(x) \in \Omega\}$. There are two cases to be considered.

Case $s_1 < s_2$. Note that for this case $g^*(x)$ is given by (5). Take any $g(x) \neq g^*(x)$, $g(x) \in \Omega$. Then either $g(x) = g^*(x)$ in $[x_1, \tilde{x}]$, or $g(x) \neq g^*(x)$ in $[x_1, \tilde{x}]$. First, let $g(x) = g^*(x)$ in $[x_1, \tilde{x}]$. Considering points $(\tilde{x}, z_1)$, (x_2, z_2) through which both the functions pass, we see that $\|g^{(1)}(x)\|_\infty \geqq \|g^{*(1)}(x)\|_\infty$ in $(\tilde{x}, x_2)$, because $g^{*(1)}(x) = [(z_2 - z_1)/(x_2 - \tilde{x})]$, the minimum possible $\|f^{(1)}\|_\infty$ for any function f passing through $(\tilde{x}, z_1)$, (x_2, z_2). Now let $g(x) \neq g^*(x)$ in $[x_1, \tilde{x}]$. Since (i) g and g^* are nondecreasing, (ii) they are continuous, (iii) $g(x) = g^*(x) = z_1$, the minimum value of $g(x)$, $g^*(x)$ in $[x_1, x_2]$, and (iv) $g^*(x) = z_1$, $\forall x \in [x_1, \tilde{x}]$, we must have some $\hat{x} \in (x_1, \tilde{x})$ such that $g(x) > g^*(x)\ \forall x \in [\hat{x}, \tilde{x}]$ and in particular $g(\tilde{x}) > g^*(\tilde{x})$. This gives $\int_{x_1}^{\tilde{x}} g(x)\,dx > \int_{x_1}^{\tilde{x}} g^*(x)\,dx$, implying that $\int_{\tilde{x}}^{x_2} g(x)\,dx < \int_{\tilde{x}}^{x_2} g^*(x)\,dx$, which shows that there exists a $c \in (\tilde{x}, x_2)$ such that $g(c) < g^*(x)$. Thus, at $\tilde{x}$, $g(\tilde{x}) > g^*(\tilde{x})$, and at c, $g(c) < g^*(c)$, which together imply that there is a $d \in (\tilde{x}, c) \subset (\tilde{x}, x_2)$ such that $g(d) = g^*(d)$. Now considering points $(d, g^*(d))$, (x_2, z_2) through which both the functions pass, we see that $\|g^{(1)}(x)\|_\infty \geqq \|g^{*(1)}(x)\|_\infty$ in $(\tilde{x}, x_2)$ because $g^{*(1)}(x) = (z_2 - z_1)/(x_2 - \tilde{x}) = (z_2 - g^*(d))/(x_2 - d)$, the minimum possible $\|f^{(1)}\|_\infty$ for any function f passing through $(d, g^*(d))$, (x_2, z_2). Thus for $s_1 < s_2$, $\|g^{*(1)}\|_\infty \leqq \|g^{(1)}\|_\infty$ for any $g(x) \in \Omega$. The case $s_1 \geqq s_2$ can be proved similarly.

The following lemma considers several intervals together. Since the result is intuitive, for brevity, we will omit its proof, which involves many seemingly unavoidable details. The basic arguments used are clear: since d_i's are the first divided differences of the data, condition (a) is true if and only if $g(x) \in G_\infty^1[a, b]$ is nondecreasing, and

any i $(1 \leqq i \leqq r-2)$ such that $d_i = d_{i+1} \neq d_{i+2} = d_{i+3}$, forces a discontinuity of $g(x)$ at x_{i+2}, justifying condition (b). Note that in Lemma 2, there is no attempt to minimize $\|g^{(1)}\|_\infty$. It only considers the existence of a nondecreasing, continuous function in $G_\infty^{(1)}[a, b]$. For a related result for strictly convex data $(d_i < d_{i+1},\ i = 1, \cdots, r+1)$ see [36, Thm. 1], [38, Thm. A].

LEMMA 2. *For $n = 1$ and $p = (r+2)$ points $\{x_i\}_1^{r+2}$, $\{y_i\}_1^{r+2} \in R^{r+2}$, $x_i < x_{i+1}$, $1 \leqq i \leqq r+1$, there exists a nondecreasing, continuous function $g(x) \in G_\infty^{(1)}[a, b]$ iff* (a) *the sequence $\{d_i = (y_{i+1} - y_i)/(x_{i+1} - x_i)\}_1^{r+1}$ is nondecreasing, and* (b) *there is no i $(1 \leqq i \leqq r-2)$ for which $d_i = d_{i+1} \neq d_{i+2} = d_{i+3}$ holds.*

For the following main results, we will employ similar notation, especially the s_i's of Lemma 1 and $z_{1,i}$, $z_{2,i+1}$, where for each interval $[x_i, x_{i+1}]$, $1 \leqq i \leqq r+1$, $z_{1,i}$, $z_{2,i+1}$ represent the values corresponding to the z_1, z_2 values of Lemma 1, satisfying conditions (i)(a) or (b) given *there*. Theorem 1 combines the ideas of Lemma 1 (minimization of $\|g^{(1)}\|_\infty$ in a single interval), and Lemma 2 (problem data over several intervals) to provide a framework for solving problem (1) for $n = 2$.

3. Convex programming characterization: the main results.

THEOREM 1. *For $n = 2$ and $(r+2)$ given points $\{x_i\}_1^{r+2}$, $\{y_i\}_1^{r+2} \in R^{r+2}$, $x_i < x_{i+1}$, $i = 1, \cdots, r+1$; if $\{d_i = (y_{i+1} - y_i)/(x_{i+1} - x_i)\}_1^{r+1}$ is a nondecreasing sequence such that there is no i $(1 \leqq i \leqq r-2)$ for which $d_i = d_{i+1} \neq d_{i+2} = d_{i+3}$ holds then a solution $f^*(x)$ of* (1) *exists and can be found by solving the following convex programming problem* (7). *Problem* (7) *has a linear objective function, $(r+1)$ variables $(t_0, t_1, \cdots, t_r)$, $2r$ main constraints* ((a), (b), (c)), *and r lower and r upper bounds on the variables* (d). *The optimal value of $\|f^{*(2)}\|_\infty$ is given by the optimal value of t_0, and the optimal convex function can be constructed from the optimal values of t_i, $i = 1, \cdots, r$ (as shown in Corollary* 1):

Minimize t_0

Subject to:

$$
\begin{aligned}
&\text{(a)}\quad t_0 - t_1 \geqq 0, \qquad t_0 - (\hat{a}_r k_r - \hat{b}_r t_r) \geqq 0,\\
&\text{(b)}\quad t_0 - [(\hat{a}_i k_i - \hat{b}_i t_i + t_{i+1})^2 / 4(\hat{a}_i k_i - \hat{b}_i t_i)] \geqq 0,\\
&\text{(c)}\quad t_0 - [(\hat{a}_i k_i - \hat{b}_i t_i + t_{i+1})^2 / 4 t_{i+1}] \geqq 0,\\
&\text{(d)}\quad (\hat{a}_i / \hat{b}_i) k_i \geqq t_i \geqq 0, \qquad i = 1, \cdots, r.
\end{aligned}
\qquad \left.\begin{aligned}\\ \\ \\ \end{aligned}\right\} \quad i = 1, \cdots, r-1, \tag{7}
$$

The constants $\hat{a}_i$, $\hat{b}_i$ and k_i, $i = 1, \cdots, r$ are calculated from the data as follows:

$$
\begin{aligned}
&\hat{a}_i = (\Delta x_i + \Delta x_{i+1}) / \Delta x_{i+1}, \qquad \hat{b}_i = (\Delta x_i / \Delta x_{i+1}),\\
&k_i = 2(\Delta y_i \Delta x_{i+1} - \Delta x_i \Delta y_{i+1}) / [(x_{i+1}^2 - x_i^2)\Delta x_{i+1} - (x_{i+2}^2 - x_{i+1}^2)\Delta x_i],\\
&\Delta x_i = x_{i+1} - x_i, \qquad \Delta y_i = y_{i+1} - y_i.
\end{aligned}
$$

Proof. Lemma 2 implies that a convex function $f \in F_\infty^{(2)}[a, b]$ through the $(r+2)$ given points exists. As noted before, if $g^*(x)$ is the solution of (2), $\|f^{*(2)}\|_\infty = \|g^{*(1)}\|_\infty$. Therefore, the optimal value $\|f^{*(2)}\|_\infty$ can be determined by finding the optimal values of $\{z_{1,i}, z_{2,i+1}\}_1^{r+1}$ such that $\max\{\|g_i^{(1)}\|_\infty : 1 \leqq i \leqq r+1\}$ is minimized under the constraints

$$
\begin{aligned}
&(x_{i+1} - x_i) z_{1,i} \leqq y_{i+1} - y_i \leqq (x_{i+1} - x_i) z_{2,i+1},\\
&(z_{2,i+1} \geqq z_{1,i}),
\end{aligned}
\qquad \left.\begin{aligned}\\ \end{aligned}\right\} \quad i = 1, \cdots, r+1, \tag{8}
$$

where $\|g_i^{(1)}\|_\infty$ in the ith interval is given by (4) of Lemma 1. Note that the first constraint

implies the conditions (i)(a) if $z_{2,i+1} = z_{1,i}$, or (i)(b) if $z_{2,i+1} > z_{1,i}$, of Lemma 1. This insures that $\{z_{1,i}, z_{2,i+1}\}_1^{r+1}$ values chosen will be such that we have an appropriate convex function within each interval $[x_i, x_{i+1}]$, $1 \leqq i \leqq r+1$. The second condition insures that when these intervals are considered together this convexity extends to the entire domain $[x_1, x_{r+2}]$. Thus, a solution of (1) can be obtained by solving:

$$\text{Minimize } t_0 = \max_{i=1,3,5,\cdots,2r+1} [\max \{((s_i + s_{i+1})^2/4s_i), ((s_i + s_{i+1})^2/4s_{i+1})\}]$$

under the above constraints. Note that for the minimization of $\|g_1^{(1)}\|_\infty$, an optimal value of $z_{2,2}$ implies that $s_1 = s_2 = \|g_1^{(1)}\|_\infty$ in the *first* interval. Similarly, we have $s_{2r+2} = s_{2r+1} = \|g_{r+1}^{(1)}\|_\infty$ in the *last* interval. Hence, eliminating s_2, s_{2r+2} and renumbering the s_i's ($[(s_1, s_2), s_3, \cdots, (s_{2r+1}, s_{2r+2})] \equiv [s_1, s_2, \cdots, s_{2r}]$), the objective function can be written as:

$$\text{Minimize } t_0 = \max_{i=2,4,\cdots,2r-2} [\max \{s_1, ((s_i + s_{i+1})^2/4s_i), ((s_i + s_{i+1})^2/4s_{i+1}), s_{2r}\}].$$

Now we can directly check by their definitions that

$$s_i \Delta x_{j(i)} + s_{i+1} \Delta x_{j(i)+1} = (\Delta x_{j(i)} + \Delta x_{j(i)+1}) k_{j(i)}, \qquad i = 1, 3, \cdots, 2r-1, \quad j(i) = (i+1)/2, \tag{9}$$

or

$$s_{i+1} = \hat{a}_{j(i)} k_{j(i)} - \hat{b}_{j(i)} s_i, \qquad i = 1, 3, \cdots, 2r-1,$$

where $k_{j(i)}$ is a constant representing $f^{(2)}$ of *the* second degree polynomial through three consecutive points $(x_{j(i)}, y_{j(i)})$, $(x_{j(i)+1}, y_{j(i)+1})$, $(x_{j(i)+2}, y_{j(i)+2})$. We can also check that constraints (8) hold iff $s_i \geqq 0$, $i = 1, \cdots, 2r$.

Now if we give the name t_i for the variable $s_{(2i-1)}$, $i = 1, \cdots, r$ and use (9) to eliminate $s_2, s_4, \cdots, s_{2r}$, we can write the problem (after some simplification) as

$$\begin{aligned} &\text{Minimize } t_0 = \max_{1 \leqq i \leqq r-1} [\max \{t_1, ((\hat{a}_i k_i - \hat{b}_i t_i + t_{i+1})^2/4(\hat{a}_i k_i - \hat{b}_i t_i)), \\ &\qquad ((\hat{a}_i k_i - \hat{b}_i t_i + t_{i+1})^2/4t_{i+1}), (\hat{a}_r k_r - \hat{b}_r t_r)\}] \\ &\text{Subject to: } \hat{a}_i k_i \geqq \hat{b}_i t_i \geqq 0 \qquad i = 1, \cdots, r. \end{aligned} \tag{10}$$

Problem (10) can now be written in the form (7) given in the theorem.

To show that a minimizing problem given by (7) is convex, we need to demonstrate that its constraints are concave and the objective is convex [29]. Being linear, the objective of (7) is convex, and constraints (a) and (d) are concave. Therefore, we have only to show that (b) and (c) are concave for all $i = 1, \cdots, r-1$. Since the form of the constraints (b) and (c) is the same for all i, it is, obviously, sufficient to show concavity for a single value of i. If H is the Hessian of constraint (b) for $i = 1$, we see that

$$H = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \dfrac{-\hat{b}_1^2 t_2^2}{2(\hat{a}_1 k_1 - \hat{b}_1 t_1)^3} & \dfrac{-\hat{b}_1 t_2}{2(\hat{a}_1 k_1 - \hat{b}_1 t_1)^2} \\ 0 & \dfrac{-\hat{b}_1 t_2}{2(\hat{a}_1 k_1 - \hat{b}_1 t_1)^2} & \dfrac{-1}{2(\hat{a}_1 k_1 - \hat{b}_1 t_1)} \end{pmatrix}.$$

Let $G = -H$. Then it is easy to see that G is symmetric, all its diagonal elements are nonnegative, and all its leading principal determinants are also nonnegative for all

values of t_2, $(\hat{a}_1 k_1 - \hat{b}_1 t_1) \geqq 0$. Thus G is semipositive definite, implying that function $(-b)$ is convex, and therefore, (b) is concave for $i=1$. Similarly, the Hessian for (c), $i=1$, is

$$H = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \dfrac{-\hat{b}_1^2}{2t_2} & \dfrac{-\hat{b}_1(\hat{a}_1 k_1 - \hat{b}_1 t_1)}{2t_2^2} \\ 0 & \dfrac{-\hat{b}_1(\hat{a}_1 k_1 - \hat{b}_1 t_1)}{2t_2^2} & \dfrac{-(\hat{a}_1 k_1 - \hat{b}_1 t_1)^2}{2t_2^3} \end{pmatrix},$$

which implies that function (c) is also concave for $i=1$. This completes the proof of the theorem.

It should be noted that the convexity of this formulation is very valuable for numerical computations, since it implies that any local minimum of the problem is also its global minimum.

We can use the following corollary to calculate an optimal convex spline $f^*(x)$. In view of Lemma 1 and 2, the corollary is evident by noting that t_i^*'s here correspond to the s_i's of Lemma 1.

COROLLARY 1. *Let optimal values $z_{1,i}^*$, $z_{2,i+1}^*$, $i=1, \cdots, r+1$ be given in terms of the optimal values of the variables t_i^*, $i=1, \cdots, r$ of Theorem 1 as follows*:

$$\left.\begin{aligned} z_{1,1}^* &= (\bar{\Delta} y_2/\bar{\Delta} x_2) - t_1^*(\bar{\Delta} x_2/2), \\ z_{1,i}^* &= (\bar{\Delta} y_i/\bar{\Delta} x_i) + t_{i-1}^*(\bar{\Delta} x_i/2), \\ z_{2,i}^* &= z_{1,i}^*, \end{aligned}\right\} \quad i=2, \cdots, r+1,$$
$$z_{2,r+2}^* = (\bar{\Delta} y_{r+2}/\bar{\Delta} x_{r+2}) - t_r^*(\bar{\Delta} x_{r+1}/2) + k_r[(\bar{\Delta} x_{r+2} + \bar{\Delta} x_{r+1})/2], \tag{11}$$

where $\bar{\Delta} y_i = (y_i - y_{i-1})$, $\bar{\Delta} x_i = (x_i - x_{i-1})$, $i=2, \cdots, r+2$. Let g^* be defined in each interval $[x_i, x_{i+1}]$, $i=1, \cdots, r+1$, by (5) and (6) (of Lemma 1) with $z_1 = z_{1,i}^*$, $z_2 = z_{2,i+1}^*$. Then an optimal convex spline, solving the problem (1) for $n=2$, is given by $f^*(x) = \int_{x_1}^{x} g^*(x)\, dx + y_1$.

Now we mention two common special cases. If $\{d_i = (y_{i+1} - y_i)/(x_{i+1} - x_i)\}_1^{r+1}$ is *strictly increasing*, obviously, there is no i such that $d_i = d_{i+1} \neq d_{i+2} = d_{i+3}$ $(1 \leqq i \leqq r-2)$, so that a solution of $f^*(x)$ of (1) exists and can be found by solving the convex programming problem (7) and using Corollary 1. For *equidistant* data, it is easy to check that $\hat{a}_i$ and $\hat{b}_i$ values simplify to $\hat{a}_i = 2$, $\hat{b}_i = 1$, $i=1, \cdots, r$, in formulation (7).

The above results help find any solution $f^* \in F_\infty^{(2)}[a, b]$ which minimizes $\|f^{*(2)}\|_\infty$. There is no emphasis on determining the maximum number of knots f^* may need, or on finding an f^* analogous to the perfect spline of Karlin's theorem [23] which characterizes f^* without the convexity requirement on f^*. Under convexity, it is natural that a perfect spline f^* (which has by definition a single value for $|f^{*(2)}|$ with $f^{*(2)}$ having opposite signs on the different sides of a knot) should be replaced by a spline where $f^{*(2)}$ takes only two values, either $f^{*(2)} > 0$ or 0, with different values on the different sides of a knot. For completeness, the following theorem is given which states the result analogous to Karlin's theorem for $n=2$. It is due to Smith and Ward [43] who prove it for general n. We give it in our notation for $n=2$.

THEOREM 2 (Smith and Ward). *For $n=2$ and $(r+2)$ points $\{x_i\}_1^{r+2}, \{y_i\}_1^{r+2} \in R^{r+2}$, $x_i < x_{i+1}$, $i=1, \cdots, r+1$, if there is some $f(x) \in F_\infty^{(2)}[a, b]$, then there exist a solution f^* of (1) and values $\xi_1, \cdots, \xi_k$ with $k \leqq r-1$, such that $a = x_1 = \xi_0 < \xi_1 < \cdots < \xi_k < \xi_{k+1} =$*

$x_{r+2} = b$, *and either*

$$f^{*(2)}(x) = \begin{cases} \|f^{*(2)}\|_\infty & \text{for } x \in (\xi_{2i}, \xi_{2i+1}), \quad (2i+1) \leqq k+1, \\ 0 & \text{for } x \in (\xi_{2i+1}, \xi_{2i+2}), \quad (2i+2) \leqq k+1, \end{cases}$$

or

$$f^{*(2)}(x) = \begin{cases} 0 & \text{for } x \in (\xi_{2i}, \xi_{2i+1}), \quad (2i+1) \leqq k+1, \\ \|f^{*(2)}\|_\infty & \text{for } x \in (\xi_{2i+1}, \xi_{2i+2}), \quad (2i+2) \leqq k+1, \end{cases}$$

for $i = 0, \cdots, i(k)$, *where* $i(k) = k/2$ *if* k *is even, and* $i(k) = (k-1)/2$ *if it is odd.*

Proof. We will prove the theorem by showing that if g^* is any solution of (1) for $n = 2$ with $\|g^{*(2)}\| = \beta$, then, as described in the statement, there is also a "perfect" convex spline solution f^* of (1) having less than or equal to $(r-1)$ knots with $\|f^{*(2)}\|_\infty = \beta$. This is done by considering the following related problem:

(1)′ Minimize $\{\|f^{(2)}\|_\infty : f \in W_\infty^{(2)}[a, b], f(x_i) = y_i - (\beta/2)(x_i^2/2), i = 1, \cdots, r+2\}$,

whose perfect spline solution can be used to obtain $f^*(x)$. Note that problem (1)′ does not have the convexity requirement, therefore by Karlin's theorem [23] there is a perfect spline solution F^* of (1)′ with number of knots $\leqq (r-1)$. Now we can show that (i) $\|F^{*(2)}\|_\infty = \beta/2$, and (ii) then the desired f^* is given by $f^*(x) = F^*(x) + (\beta/2)(x^2/2)$.

(i) Let $G^*(x) = g^*(x) - (\beta/2)(x^2/2)$. Then $G^*(x) \in W_\infty^{(2)}[a, b]$, it interpolates the data of problem (1)′: $G^*(x_i) = g^*(x_i) - (\beta/2)(x_i^2/2) = y_i - (\beta/2)(x_i^2/2)$, and since $G^{*(2)}(x) = g^{*(2)}(x) - \beta/2$, we get $\|G^{*(2)}\|_\infty = \|g^{*(2)}\|_\infty - \beta/2 = \beta/2$. Thus, $G^*(x)$ would be a solution of (1)′ if $\|F^{*(2)}\|_\infty = \beta/2$, that is, $\|F^{*(2)}\|_\infty$ is not *less than* $\beta/2$. We will prove this by showing that $\|F^{*(2)}\|_\infty < \beta/2$ leads to the contradiction that g^*, an assumed solution of (1), is not a solution of (1). Let $\|F^{*(2)}\|_\infty < \beta/2 = \beta/2 - \varepsilon$, where $(\beta/2) \geqq \varepsilon > 0$. Consider $h(x) = F^*(x) + (\beta/2)(x^2/2)$. Then $h(x) \in W_\infty^{(2)}[a, b]$, it interpolates the data of problem (1): $h(x_i) = F^*(x_i) + (\beta/2)(x_i^2/2) = y_i - (\beta/2)(x_i^2/2) + (\beta/2)(x_i^2/2) = y_i$, and since $h^{(2)}(x) = F^{*(2)} + \beta/2$, we get $h^{(2)}(x) \geqq 0$ (hence $h(x)$ is convex), and $\|h^{(2)}\|_\infty = \|F^{*(2)}\|_\infty + \beta/2 = \beta/2 - \varepsilon + \beta/2 = \beta - \varepsilon$. Thus $h(x)$ is a solution of (1) with $\|h^{(2)}\|_\infty < \beta$, which contradicts the fact that $g^*(x)$ is a solution of (1) with $\|g^{*(2)}\|_\infty = \beta$. Therefore, $\|F^{*(2)}\|_\infty = \beta/2$.

(ii) As shown for $h(x)$ above, $f^*(x)$ defined by $f^*(x) = F^*(x) + (\beta/2)(x^2/2)$ is a solution of (1). Now since $f^{*(2)}(x) = F^{*(2)}(x) + \beta/2$ and $F^*(x)$ is a perfect ($F^{*(2)}(x) = \beta/2$, or $(-\beta/2)$) spline of degree 2 with number of knots $\leqq (r-1)$, we see that $f^*(x)$ has the same number of knots with $f^{*(2)}(x) = 0$ or β. This proves the theorem.

4. Numerical examples and a computer implementation. Since the capability of available computer software to find a local minimum far exceeds that of finding a global one, convexity of the formulation is very helpful in actual calculations. Here, we provide two examples for illustration and initial numerical validation. Each example is such that its optimal value $\|f^{*(2)}\|_\infty$ can also be directly found. This allows a comparison with the optimal value obtained by computer routines using our formulation.

(i) Consider a 4-point problem, $p = 4$, $n = 2$, $r = 2$:

$$\{x_i\}_1^4 \equiv \{0, 1, 2, 3\}, \qquad \{y_i\}_1^4 \equiv \{0, 10, 28, 54\}.$$

We can calculate $k_1 = k_2 = 8$, using the easily derived expression

$$k_i = 2\frac{(y_i - y_{i+1})(x_{i+1} - x_{i+2}) - (y_{i+1} - y_{i+2})(x_i - x_{i+1})}{(x_i^2 - x_{i+1}^2)(x_{i+1} - x_{i+2}) - (x_{i+1}^2 - x_{i+2}^2)(x_i - x_{i+1})}. \tag{12}$$

Recall that the constant k_i represents $f^{(2)}$ of the second degree polynomial through these consecutive points: (x_i, y_i), (x_{i+1}, y_{i+1}), and (x_{i+2}, y_{i+2}). Since the points lie on a second degree polynomial $g(x) = 4x^2 + 6x$, it is obvious that the optimal value $\|f^{*(2)}\|_\infty$ is $g^{(2)} = 8$. The formulation (7) with $\hat{a}_i = 2$, $\hat{b}_i = 1$ for equidistant data gives

Minimize t_0

Subject to:

$$
\begin{aligned}
& g_1 \equiv t_0 - t_1 \geqq 0, \\
& g_2 \equiv t_0 - [(16 - t_1 + t_2)^2/4(16 - t_1)] \geqq 0, \\
& g_3 \equiv t_0 - [(16 - t_1 + t_2)^2/4t_2] \geqq 0, \\
& g_4 \equiv t_0 - (16 - t_2) \geqq 0, \\
& g_5 \equiv 16 - t_1 \geqq 0, \\
& g_6 \equiv 16 - t_2 \geqq 0, \\
& g_7 \equiv t_1 \geqq 0, \\
& g_8 \equiv t_2 \geqq 0.
\end{aligned} \tag{13}
$$

It is easy to check that (13) has an optimal solution at $(t_0^* = 8 = \|f^{*(2)}\|_\infty$, $t_1^* = 8$, $t_2^* = 8)$. This can be done (as follows) by verifying that the Kuhn-Tucker conditions sufficient for a solution to be an optimal solution of a convex program are satisfied at $(t_0 = 8, t_1 = 8, t_2 = 8)$ with dual variable values $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = 1/4$, $\lambda_5 = \lambda_6 = \lambda_7 = \lambda_8 = 0$. Substituting $t_0 = t_1 = t_2 = 8$, we see that this is a feasible solution since all the constraints g_1 through g_8 are satisfied. Since $g_1 = g_2 = g_3 = g_4 = 0$, and $g_5 = g_6 = g_7 = g_8 > 0$ at this solution, we have the complementary slackness conditions: $g_i\lambda_i = 0$ for $i = 1, \cdots, 8$. Finally using * to denote the value at $(t_0 = t_1 = t_2 = 8)$, we see that the following gradient conditions are also satisfied:

$$
\begin{aligned}
\frac{\partial t_0^*}{\partial t_0} &= \sum_{i=1}^{8} \lambda_i \frac{\partial g_i^*}{\partial t_0} \equiv 1 = \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4, \\
\frac{\partial t_0^*}{\partial t_1} &= \sum_{i=1}^{8} \lambda_i \frac{\partial g_i^*}{\partial t_1} \equiv 0 = -\lambda_1 + \lambda_3, \\
\frac{\partial t_0^*}{\partial t_2} &= \sum_{i=1}^{8} \lambda_i \frac{\partial g_i^*}{\partial t_2} \equiv 0 = -\lambda_2 + \lambda_4.
\end{aligned}
$$

Thus our formulation solves the problem, and its optimal value t_0^* gives the optimal value of $\|f^{*(2)}\|_\infty$.

Now using a computer routine for constrained optimization given in [26, pp. 386-398, Constrained Rosenbrock HILL ALGORITHM] we obtain the same solution. Several initial (starting) solutions were tried, all converging to the optimal values shown in Table 1.

(ii) We take a 7-point problem for our second example, $p = 7$, $n = 2$, $r = 5$:

$$
\begin{aligned}
\{x_i\}_1^7 &\equiv \{0, 1, 2, 3, 4, 5, 6\}, \\
\{y_i\}_1^7 &\equiv \{16, 12, 10, 9.5, 10, 12.5, 18\}.
\end{aligned}
$$

We can calculate $k_1 = 2$, $k_2 = 1.5$, $k_3 = 1$, $k_4 = 2$, $k_5 = 3$ as before. It is easy to verify, this time geometrically, that there *is* a convex function $f^* \in F_\infty^{(2)}[a, b]$ such that $\|f^{*(2)}\|_\infty = 3$. This implies that f^* is an optimal solution with $\|f^{*(2)}\|_\infty = 3$, since for

TABLE 1

Initial solution		Optimal values		
t_1	t_2	t_0^*	t_1^*	t_2^*
1.0	1.0	8.0000000	8.000009	7.999994
5.0	5.0	8.0000010	8.000009	7.999993
12.0	12.0	8.0000000	7.999999	8.000001

any $f \in F_\infty^{(2)}[a, b]$, obviously, $\|f^{(2)}\|_\infty \geqq \max_{1 \leqq i \leqq 5} \{k_i\} = 3$. Thus, the formulation, with 6 variables in this case, must have optimal $t_0^* = 3$. Note that optimal values of $t_1, \cdots, t_5$ are not unique for this example.

The computer routine mentioned above (Rosenbrock HILL ALGORITHM) again gives the desired optimal values shown in Table 2 from several initial solutions. For brevity, explicit formulation is not given for this problem.

TABLE 2

Initial solution					Optimal values					
t_1	t_2	t_3	t_4	t_5	t_0^*	t_1^*	t_2^*	t_3^*	t_4^*	t_5^*
1.0	1.0	1.0	1.0	1.0	3.0000696	2.280090	7.855989	1.009406	1.040345	2.999885
3.5	2.5	1.5	3.5	5.5	3.0000019	2.455462	8.606190	1.711703	1.007163	2.999999
3.8	2.8	1.8	3.8	5.8	3.0005379	2.418691	1.069808	1.230089	1.112060	2.999463

Acknowledgments. I am indebted to the referees for extremely detailed and constructive criticism which has resulted in substantial improvement in the present version of the paper. I also express my gratitude to Professor Philip W. Smith for his kind hospitality, valuable discussion, and Iliev and Pollul, and Smith and Ward references.

REFERENCES

[1] R. D. ARMSTRONG AND J. W. HULTZ (1977), *An algorithm for a restricted discrete approximation problem in the L_1 norm*, SIAM J. Numer. Anal., 14, pp. 555-565.

[2] R. K. BEATSON (1978), *Jackson-type theorems for approximation with Hermite-Birkhoff interpolatory side conditions*, J. Approx. Theory, 22, pp. 95-104.

[3] ——— (1980), *On the degree of approximation with Hermite interpolatory side conditions*, J. Approx. Theory, 28, pp. 197-206.

[4] R. BELLMAN, B. G. KASHEF AND R. VASUDEVAN (1972), *Splines via dynamic programming*, J. Math. Anal. Appl., 38, pp. 471-479.

[5] ——— (1973), *Dynamic programming and bicubic spline interpolation*, J. Math. Anal. Appl., 44, pp. 160-174.

[6] B. D. BOJANOV (1974), *Optimal methods of interpolation in $W^{(r)}L_q(M; a, b)$*, C.R. Acad. Bulgare Sci., 27, pp. 885-888.

[7] M. G. COX (1972), *The numerical evaluation of B-splines*, J. Inst. Math. Appl., 10, pp. 134-149.

[8] C. DE BOOR (1972), *On calculating with B-splines*, J. Approx. Theory, 6, pp. 50-62.

[9] ——— (1974), *A remark concerning perfect splines*, Bull. Amer. Math. Soc., 80, pp. 724-727.

[10] ——— (1976), *Computational aspects of optimal recovery*, in Optimal Estimation in Approximation Theory, Proc. International Symposium, Freudenstadt, 1976, C. A. Micchelli and T. J. Rivlin, eds., pp. 69-91. Plenum, New York, 1977.

[11] ——— (1977), *Package for calculating with B-splines*, SIAM J. Numer. Anal., 14, pp. 441-472.

[12] W. DENT (1973), *A note on least squares fitting of functions constrained to be either nonnegative, nondecreasing, or convex*, Management Sci., 20, pp. 130-132.
[13] R. E. ESCH AND W. L. EASTMAN (1969), *Computational methods for best spline function approximation*, J. Approx. Theory, 2, pp. 85-96.
[14] S. D. FISHER AND J. W. JEROME (1974), *Perfect spline solutions to L_∞ extremal problems*, J. Approx. Theory, 12, pp. 78-90.
[15] ——— (1974), *Existence, characterization and essential uniqueness of L_∞ extremal problems*, Trans. Amer. Math. Soc., 187, pp. 391-404.
[16] W. T. FORD AND J. A. ROULIER (1974), *On interpolation and approximation by polynomials with monotone derivatives*, J. Approx. Theory, 10, pp. 123-130.
[17] P. W. GAFFNEY AND M. J. D. POWELL (1976), *Optimal interpolation*, in Numerical Analysis, G. A. Watson, eds., Lecture Notes in Mathematics 506, Springer-Verlag, Heidelberg.
[18] C. A. HOLLOWAY (1979), *On the estimation of convex functions*, Oper. Res., 27, pp. 401-407.
[19] V. HORNUNG (1978), *Monotone spline-interpolation*, in Numerische Methoden der Approximations Theorie 4, L. Collatz, C. Meinardus and H. Werner, eds., Birkhauser, Basel, Stuttgart.
[20] ——— (1980), *Interpolation by smooth functions under restrictions on the derivatives*, J. Approx. Theory, 28, pp. 124-128.
[21] D. J. HUDSON (1969), *Least squares fitting of a polynomial constrained to be either nonnegative, nondecreasing, or convex*, J. Roy. Statist. Soc. B, 31, pp. 113-118.
[22] G. ILIEV AND W. POLLUL (1984), *Convex interpolation with minimal L_∞-norm of the second derivative*, Math. Z., 186, pp. 49-56.
[23] S. KARLIN (1973), *Some variational problems on certain Sobolev spaces and perfect splines*, Bull. Amer. Math. Soc., 206, pp. 25-66.
[24] ——— (1975), *Interpolation properties of generalized perfect splines and the solutions of certain extremal problems* I, Trans. Amer. Math. Soc., 79, pp. 124-128.
[25] G. S. KIMELDORF AND G. WAHBA (1971), *Some results on Tchebycheffian spline functions*, J. Math. Anal. Appl., 33, pp. 82-95.
[26] J. L. KUESTER AND J. H. MIZE (1973), *Optimization Techniques With Fortran*, McGraw-Hill, New York.
[27] T. LYCHE AND L. L. SCHUMAKER (1973), *Computation of smoothing and interpolating natural splines via local bases*, SIAM J. Numer. Anal., 10, pp. 1027-1038.
[28] O. L. MANGASARIAN AND L. L. SCHUMAKER (1971), *Discrete splines via mathematical programming*, SIAM J. Control, 9, pp. 174-183.
[29] O. L. MANGASARIAN (1979), *Nonlinear Programming*, Robert E. Krieger Publishing, Huntington, NY.
[30] D. F. MCALLISTER, E. PASSOW AND J. A. ROULIER (1977), *Algorithms for computing shape preserving spline interpolations to data*, Math. Comp., 31, pp. 717-725.
[31] D. F. MCALLISTER AND J. A. ROULIER (1978), *Interpolation by convex quadratic splines*, Math. Comp., 32, pp. 1154-1162.
[32] ——— (1981), *An algorithm for computing a shape-preserving osculatory quadratic spline*, ACM Trans. Math. Software, 3, pp. 331-347.
[33] C. A. MICCHELLI, T. J. RIVLIN AND S. WINOGRAD (1976), *The optimal recovery of smooth functions*, Numer. Math., 26, pp. 191-200.
[34] E. PASSOW (1974), *Piecewise monotone spline interpolation*, J. Approx. Theory, 12, pp. 240-241.
[35] ——— (1976), *An improved estimate of the degree of monotone interpolation*, J. Approx. Theory, 17, pp. 115-118.
[36] ——— (1977), *Monotone quadratic spline interpolation*, J. Approx. Theory, 19, pp. 143-147.
[37] E. PASSOW AND L. RAYMON (1975), *The degree of piecewise monotone interpolation*, Proc. Amer. Math. Soc., 2, pp. 409-412.
[38] E. PASSOW AND J. A. ROULIER (1977), *Monotone and convex spline interpolation*, SIAM J. Numer. Anal., 14, pp. 904-909.
[39] K. RITTER (1969), *Splines and quadratic programming*, Conference on Approximations, Univ. Wisconsin, Madison.
[40] J. A. ROULIER AND D. F. MCALLISTER (1980), *Approximation by convex quadratic splines*, in Approximation Theory III, E. W. Chenney, ed., Academic Press, New York.
[41] Z. RUBINSTEIN (1970), *On polynomial δ-type functions and approximation by monotone polynomials*, J. Approx. Theory, 3, pp. 1-6.
[42] K. SCHITTKOWSKI (1980), *Nonlinear Programming Codes—Information, Tests, Performance*, Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, New York.
[43] P. W. SMITH AND J. D. WARD (1984), *Constrained L_∞ approximation* (personal communication).
[44] L. S. THAKUR (1978), *Error analysis for convex separable programs: the piecewise linear approximation and the bounds on the optimal objective value*, SIAM J. Appl. Math., 34, pp. 704-714.

[45] L. S. THAKUR (1980), *Error analysis for convex separable programs: bounds on optimal and dual optimal solutions*, J. Math. Anal. Appl., 75, pp. 486-496.

[46] G. WAHBA (1973), *On the minimization of a quadratic functional subject to a continuous family of linear inequality constraints*, this Journal, 11, pp. 64-79.

[47] ——— (1978), *Improper priors, spline smoothing and problem of guarding against model errors in regression*, Tech. Report No. 508, Dept. Statistics, Univ. Wisconsin, Madison.

[48] A. D. WAREN AND L. S. LASDON (1979), *The state of nonlinear programming software*, Oper. Res., 27, pp. 431-456.

[49] I. W. WRIGHT AND E. J. WEGMAN (1980), *Isotonic, convex, and related splines*, Ann. Statist., 8, pp. 1023-1035.

[50] S. W. YOUNG (1967), *Piecewise monotone polynomial interpolation*, Bull. Amer. Math. Soc., 73, pp. 642-643.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 2, March 1986

ON THE REGULARITY OF THE KUHN-TUCKER CURVE*

A. L. DONTCHEV† AND H. TH. JONGEN‡

Abstract. We consider twice continuously differentiable finite dimensional optimization problems, depending on a real parameter. Besides a discussion of (local) Lipschitz continuity of the Kuhn-Tucker curve, we present conditions under which the Kuhn-Tucker curve is piecewise continuously differentiable. Throughout the paper we assume the linear independence of the gradients of the binding constraint functions. We use Kojima's concept of strongly stable Kuhn-Tucker points and present a new equivalent formulation of this concept.

Key words. Kuhn-Tucker curve, Lipschitz continuity, piecewise continuous differentiability, strongly stable Kuhn-Tucker points, critical points

1. Introduction. Let $C^2(R^n, R)$ denote the space of real valued twice continuously differentiable functions on the n-dimensional Euclidean space R^n. By z we denote a vector in R^{n+1}, and z will always be partitioned as $z=(x, t)$, where $x\in R^n$, $t\in R$. The real number t will be considered as a parameter. Let I, J be finite index sets, $I=\{1, \cdots, m\}$, $J=\{1, \cdots, s\}$, and let $f, h_i, g_j \in C^2(R^{n+1}, R)$, $i\in I, j\in J$.

We define

$$M(t)=\{x\in R^n \mid h_i(x, t)=0,\ g_j(x, t)\geqq 0,\ i\in I, j\in J\}, \tag{1.1}$$

$$J_0(z)=\{j\in J \mid g_j(z)=0\}. \tag{1.2}$$

For every $t\in R$ we have the following optimization problem $P(t)$:

$$P(t):\quad \text{Minimize } f(\cdot, t) \text{ on } M(t). \tag{1.3}$$

In the sequel we denote by $D\phi(z)$ the $(n+1)$-row vector of the first partial derivatives of ϕ evaluated at z, and by $D^2\phi(z)$ the $(n+1)\times(n+1)$-matrix of the second partial derivatives. Similarly, $D_x\phi(z)$, $D_t\phi(z)$, $D_{xt}\phi(z)$ will correspond to the partial derivatives with respect to x and t.

For an $r\times q$-matrix A, the set Ker A will be:

$$\text{Ker } A=\{\xi\in R^q \mid A\xi=0\}. \tag{1.4}$$

If B is a symmetric $n\times n$-matrix and L a linear subspace of R^n, then by $B_{|L}$ we mean some matrix of the family $\mathcal{V}$, $\mathcal{V}=\{V^TBV \mid V$ is a matrix with n rows, whose columns form a basis for $L\}$. In view of Sylvester's theorem (cf. [12]) the number of negative (resp. zero, positive) eigenvalues of V^TBV does not depend on the incidental choice of V. We say that $B_{|L}$ is (non)singular (resp. positive) (semi-) definite if V^TBV is so, where $V^TBV\in\mathcal{V}$.

In this paper we are concerned with a *local* analysis thereby assuming the linear independence of the gradients of the binding constraint functions. In this spirit we simply assume throughout the whole paper that the following Condition A is satisfied.

Condition A. The set $\{D_xh_i(z), D_xg_j(z), i\in I, j\in J_0(z)\}$ is linearly independent at all $z=(x, t)$ for which $x\in M(t)$.

* Received by the editors March 6, 1984, and in revised form November 21, 1984.
† Institute of Mathematics, Bulgarian Academy of Sciences, Sofia, Bulgaria.
‡ Department of Applied Mathematics, Twente University of Technology, Enschede, the Netherlands.

DEFINITION 1.1 [8]. A point $\bar{x}$ is called a *critical point* for $P(\bar{t})$ if $\bar{x} \in M(\bar{t})$ and if there exist (unique) real numbers $\bar{\lambda}_i$, $\bar{\mu}_j$, $i \in I$, $j \in J_0(\bar{z})$ satisfying

$$D_x f = \sum_{i \in I} \bar{\lambda}_i D_x h_i + \sum_{j \in J_0(\bar{z})} \bar{\mu}_j D_x g_{j|z=\bar{z}}. \tag{1.5}$$

The numbers $\bar{\lambda}_i$, $\bar{\mu}_j$ are called Lagrange parameters. A critical point $\bar{x}$ is a *Kuhn–Tucker point* (KT-point) if $\bar{\mu}_j \geqq 0$ for all $j \in J_0(\bar{z})$.

A critical point $\bar{x}$ is called *nondegenerate* if the following two conditions hold:

ND1. $\bar{\mu}_j \neq 0, j \in J_0(\bar{z})$.

ND2. $D_x^2 L(\bar{z})_{|T}$ is nonsingular,

where the (Lagrange) function L and the linear subspace $T \subset R^n$ are defined as follows:

$$L(z) = f(z) - \sum_{i \in I} \bar{\lambda}_i h_i(z) - \sum_{j \in J_0(\bar{z})} \bar{\mu}_j g_j(z), \tag{1.6}$$

$$T = \bigcap_{i \in I} \operatorname{Ker} D_x h_i(\bar{z}) \cap \bigcap_{j \in J_0(\bar{z})} \operatorname{Ker} D_x g_j(\bar{z}). \tag{1.7}$$

A point $\bar{z} = (\bar{x}, \bar{t})$ is called a (nondegenerate) critical point if $\bar{x}$ is a (nondegenerate) critical point for $P(\bar{t})$. Let $\Sigma \subset R^{n+1}$ denote the set of critical points and Σ_{KT} be the subset of Σ consisting of the KT-points.

Remark 1.1. If one would like to drop Condition A, then there are two possibilities. Either one imposes a weaker constraint qualification guaranteeing that local minima are Kuhn–Tucker points (for example the Mangasarian–Fromowitz constraint qualification, as used by Kojima [9] and Kojima and Hirabayashi [11]). Or one studies generic one-parameter families, but then the whole concept of critical points has to be weakened (cf. [6], [7]). □

If in some neighborhood of $\bar{z} = (\bar{x}, \bar{t})$ the Kuhn–Tucker set Σ_{KT} can be parametrized by t, t ranging in some interval, then we will speak about "the Kuhn–Tucker curve." In this paper we discuss the regularity properties of the KT-curve (local Lipschitz continuity, piecewise C^1-differentiability). We emphasize that this kind of investigation is important for sensitivity analysis, which is basic in optimization. Therefore, we feel that every new insight in this area has its own relevance. A basic tool in sensitivity analysis is the idea of implicit function theorems. This was already recognized by Fiacco and so, let us start with the following profound result (a similar result holds for all nondegenerate critical points, cf. [8]).

THEOREM 1.1 (Fiacco [1]). *Let $\bar{z} = (\bar{x}, \bar{t})$ be a nondegenerate critical point, $\bar{\mu}_j \geqq 0$ for all $j \in J_0(\bar{z})$ and $D_x^2 L(\bar{z})_{|T}$ positive definite. Then in some open neighborhood $\mathcal{O}$ of $\bar{z}$ the Kuhn–Tucker curve depends C^1 on the parameter t, $\Sigma_{KT} \cap \mathcal{O} = \{(x(t), t), t \in (a, b)\}$. Every $x(t)$ is a strict local minimum for $P(t)$ and the corresponding Lagrange parameters $\lambda_i(t)$, $\mu_j(t)$ depend C^1 on t.*

The paper is organized as follows. In § 2 we apply a beautiful abstract theorem of Hager in order to obtain Lipschitz continuity for a curve which is continuously selected from a finite number of Lipschitz continuous curves. In § 3 we prove an equivalent formulation for Kojima's concept of strong stability. In § 4 we show that local Lipschitz continuity of the Kuhn–Tucker curve is a direct consequence of Kojima's continuity result and a continuous curve selection (here the result of § 2 is used). Finally, in § 5 we deduce fairly weak conditions for piecewise C^1-differentiability of the Kuhn–Tucker curve. Under these conditions it might be possible to adapt appropriately numerical schemes, as used in solving ordinary differential equations, for tracing the Kuhn–Tucker curve. Such an analysis would be relevant especially in optimal control computations. The latter fact will be the basis for future research.

2. On the Lipschitz continuity of a continuous selection. The idea of the result in this section is based on Hager's theorem which we will state first. Let $\mathcal{S}$ be a Banach space, $\mathcal{D}$ be a convex subset of a Banach space, and $z: \mathcal{D} \to \mathcal{S}$ be continuous. Moreover, let $c: \mathcal{D} \to 2^{\{1,\cdots,m\}}$ = power set of $\{1, \cdots, m\}$ have the following property:

(2.1) If $\{d_k\} \subset \mathcal{D}$, $d_k \to d \in \mathcal{D}$ as $k \to \infty$, and $I \subset c(d_k)$ for all k, then $I \subset c(d)$.

Given $d, e \in \mathcal{D}$, let (d, e) denote the ordered pair and define the segment:

$$[d, e] = \{(1-\lambda)d + \lambda e | 0 \leqq \lambda \leqq 1\}.$$

The points $d, e \in \mathcal{D}$ are called compatible if $c(d) = c(e)$ and $c(\delta) \subset c(d)$ for all $\delta \in [d, e]$.

THEOREM 2.1 (Hager [3]). *If γ satisfies*

$$\|z(d) - z(e)\|_{\mathcal{S}} \leqq \gamma \|d - e\|_{\mathcal{D}} \tag{2.2}$$

for all compatible $d, e \in \mathcal{D}$, then γ satisfies (2.2) *for all $d, e \in \mathcal{D}$.*

As a consequence of Theorem 2.1 we obtain the following result.

THEOREM 2.2. *Let $[a, b] \subset R$ be an interval and let $y_i: [a, b] \to R^n$ be a Lipschitz continuous function with Lipschitz constant α_i, $i = 1, \cdots, r$. Let $y: [a, b] \to R^n$ be a continuous function having the property that for all $t \in [a, b]$: $y(t) = y_i(t)$ for some $i \in \{1, \cdots, r\}$.*

Then y is Lipschitz continuous with Lipschitz constant $\alpha := \max_{i=1,\cdots,r} \alpha_i$.

Proof. We reformulate Theorem 2.2 in terms of Hager's theorem. Put $\mathcal{D} = [a, b]$, $\mathcal{S} = R^n$. The map c is defined as follows. Put $m = r$ and $c(t) = \{i \in \{1, \cdots, r\} | y(t) = y_i(t)\}$. Let $\{t_k\} \subset [a, b]$ be a sequence and suppose that for some j, $y_j(t_k) = y(t_k)$ for all k. If $t_k \to \bar{t}$, then by continuity, $y_j(t_k) \to y_j(\bar{t})$ and $y(t_k) \to y(\bar{t})$. This establishes (2.1). The map y plays the role of the map z. Finally, it suffices to show the inequality $\|y(t_1) - y(t_2)\| \leqq \alpha |t_1 - t_2|$ for all $t_1, t_2 \in [a, b]$ with $c(t_1) = c(t_2)$. But the latter fact is obvious since $\alpha = \max_{i=1,\cdots,r} \alpha_i$. □

3. An equivalent formulation of Kojima's strong stability concept. In [9] Kojima introduced strongly stable Kuhn-Tucker points. In fact, Kojima states the concept of strong stability in a more or less topological way (which incorporates continuous dependence on the data) and then he shows the equivalence of it with a condition on corresponding partial derivatives. We take this equivalent formulation as a definition.

DEFINITION 3.1 (cf. [9]). Let $\bar{x}$ be a critical point for $P(\bar{t})$ with Lagrange parameters $\bar{\lambda}_i$, $\bar{\mu}_j$ satisfying (1.5) and suppose that $\bar{x}$ is a Kuhn-Tucker point (i.e. $\bar{\mu}_j \geqq 0$, $j \in J_0(\bar{z})$). Put $J_0^+(\bar{z}) = \{j \in J_0(\bar{z}) | \bar{\mu}_j > 0\}$. Let L be the Lagrange function (cf. (1.6)) and for all $\tilde{J}$ with $J_0^+(\bar{z}) \subset \tilde{J} \subset J_0(\bar{z})$ define

$$T(\tilde{J}) = \bigcap_{i \in I} \operatorname{Ker} D_x h_i(\bar{z}) \cap \bigcap_{j \in \tilde{J}} \operatorname{Ker} D_x g_j(\bar{z}). \tag{3.1}$$

Then $\bar{x}$ is strongly stable if for all $\tilde{J}$ with $J_0^+(\bar{z}) \subset \tilde{J} \subset J_0(\bar{z})$:

K1. $D_x^2 L(\bar{z})_{|T(\tilde{J})}$ is nonsingular

K2. sign det $(D_x^2 L(\bar{z})_{|T(\tilde{J})})$ is constant.

We will present a new condition which is equivalent with K1, K2. In fact, this condition clarifies some ideas behind Kojima's work, such as the stationary index, introduced in [9]. We show that $D_x^2 L(\bar{z})_{|T(J_0^+(\bar{z}))}$ is in a certain sense a "positive extension" of $D_x^2 L(\bar{z})_{|T(J_0(\bar{z}))}$:

THEOREM 3.1. *Let $\bar{x}$ be a Kuhn-Tucker point for $P(\bar{t})$. In the terminology of Definition* 3.1, *let V be a matrix whose columns form a basis for $T(J_0^+(\bar{z}))$. Moreover,*

let $T(J_0^+(\bar z)/J_0(\bar z))$ *denote the orthogonal complement in* $T(J_0^+(\bar z))$ *of* $T(J_0(\bar z))$. *Then,* $\bar x$ *is strongly stable iff* K1*, K2* *hold*:

K1*. $V^T D_x^2 L(\bar z) V$ *is nonsingular.*

K2*. $(V^T D_x^2 L(\bar z) V)^{-1}_{|T(J_0^+(\bar z)/J_0(\bar z))}$ *is positive definite.*

The proof of Theorem 3.1 is a direct consequence of Theorem 3.2 below. But first, we need the following lemma (cf. [6], [7] for a proof).

Let Ind (index) denote "the number of negative eigenvalues" and let $\perp$ denote the orthogonal complement.

LEMMA 3.1. *Let A be a nonsingular symmetric $n\times n$-matrix, L be a linear subspace of R^n. If $A^{-1}_{|L^\perp}$ is nonsingular, then we have*:

a. $A_{|L}$ *is nonsingular.*

b. $\operatorname{Ind}(A) = \operatorname{Ind}(A_{|L}) + \operatorname{Ind}(A^{-1}_{|L^\perp})$.

Let L_a, L_b be linear subspaces of R^n, $L_a \subset L_b$ and $L_a \neq L_b$. A finite sequence of linear subspaces $L_1, L_2, \cdots, L_p$ of R^n is called a simple chain from L_a to L_b if $L_1 = L_a$, $L_p = L_b$, $L_i \subset L_{i+1}$ and $\dim(L_{i+1}) = \dim(L_i) + 1$ for $i = 1, \cdots, p-1$.

THEOREM 3.2. *Let A be a nonsingular symmetric $n \times n$-matrix and $L \subset R^n$ a linear subspace. Then $A^{-1}_{|L^\perp}$ is positive definite iff there is some simple chain $L_1, \cdots, L_r$ from L to R^n such that*:

S1. $A_{|L_i}$ *is nonsingular for* $i = 1, \cdots, r$.

S2. $\operatorname{sign\,det}(A_{|L_i})$ *is constant,* $i = 1, \cdots, r$.

Proof. The theorem in the cases $\dim(L) = 0, n$ being obvious, we assume that $0 < \dim(L) < n$.

The "only if" part. Suppose that $A^{-1}_{|L^\perp}$ is positive definite and let $L_1, \cdots, L_r$ be any simple chain from L to R^n. Since $L_i^\perp \subset L^\perp$ it follows that $A^{-1}_{|L_i^\perp}$ is positive definite as well. From Lemma 3.1 we conclude that $A_{|L_i}$ is nonsingular and

$$\operatorname{Ind}(A) = \operatorname{Ind}(A_{|L_i}) + \operatorname{Ind}(A^{-1}_{|L_i^\perp}) = \operatorname{Ind}(A_{|L_i}). \tag{3.2}$$

But (3.2) and the fact that both A and $A_{|L_i}$ are nonsingular imply that $\operatorname{sign\,det}(A) = \operatorname{sign\,det}(A_{|L_i})$.

The "if part". Let $L_1, \cdots, L_r$ be a simple chain from L to R^n such that S1, S2 hold. From S1 and Lemma 3.1a we conclude that $A^{-1}_{|L_i^\perp}$ is nonsingular. If we can show that $\operatorname{Ind}(A_{|L_i}) = \operatorname{Ind}(A_{|L_{i+1}})$, $i = 1, \cdots, r-1$, then we are done. In fact, in that case, it follows that $\operatorname{Ind}(A_{|L_1}) = \operatorname{Ind}(A_{|L_r})$; hence, $\operatorname{Ind}(A_{|L}) = \operatorname{Ind}(A)$.

Now we can apply Lemma 3.1b and conclude that $\operatorname{Ind}(A^{-1}_{|L^\perp}) = 0$. This implies that A^{-1} is positive semi-definite on $L^\perp$. But since we already know that $A^{-1}_{|L^\perp}$ is nonsingular, it follows that $A^{-1}_{|L^\perp}$ is positive definite.

So it remains to show that $\operatorname{Ind}(A_{|L_i}) = \operatorname{Ind}(A_{|L_{i+1}})$. Let V_i, V_{i+1} be a matrix whose columns form a basis for L_i, L_{i+1} (V_{i+1} extending V_i), and let T_i be the one-dimensional linear subspace of L_{i+1} orthogonal to L_i. From S_1 we see that both $V_i^T A V_i$ and $V_{i+1}^T A V_{i+1}$ are nonsingular. Now we replace in Lemma 3.1 the matrix A by $(V_{i+1}^T A V_{i+1})^{-1}$, the linear space L by T_i and obtain from Lemma 3.1a, in view of the nonsingularity of $V_i^T A V_i$, that $(V_{i+1}^T A V_{i+1})^{-1}_{|T_i}$ is nonsingular. Keeping this result in mind, we replace in Lemma 3.1 the matrix A by $V_{i+1}^T A V_{i+1}$, the linear subspace L by L_i and we obtain:

$$\operatorname{Ind}(V_{i+1}^T A V_{i+1}) = \operatorname{Ind}(V_i^T A V_i) + \operatorname{Ind} \underbrace{((V_{i+1}^T A V_{i+1})^{-1}_{|T_i})}_{\delta_i}. \tag{3.3}$$

Since T_i is one-dimensional, we see that δ_i in (3.3) is a number which is either equal to one or zero. If δ_i equals one, then $\operatorname{Ind}(V_{i+1}^T A V_{i+1})$ is even (resp. odd) whenever

Ind $(V_i^T A V_i)$ is odd (resp. even). Consequently, we obtain:

$$\text{sign det}\,(V_{i+1}^T A V_{i+1}) = -\text{sign det}\,(V_i^T A V_i). \tag{3.4}$$

However, (3.4) is in contradistinction to S2. Hence, we obtain $\delta_i = 0$ and (3.3) yields the desired result. □

Remark 3.1. From the proof of Theorem 3.2 we see that we may replace in the formulation of Theorem 3.2 "iff there is some simple chain $\cdots$" by "iff for every simple chain $L_1, \cdots, L_r$ from L to R^n the conditions S1, S2 hold:".

Remark 3.2. A special case of Lemma 3.1 is obtained as follows: A nonsingular symmetric $n \times n$-matrix A is positive definite iff for some (and hence every) linear subspace $L \subset R^n$ we have: $A_{|L}$ and $A_{|L^\perp}^{-1}$ are positive definite. This particular case has been proved by Fujiwara et al. [2] by means of tools from optimization theory. Finally, we mention the following: at the level of positive (semi-) definiteness there is an interesting extension proved by Han and Mangasarian [4] where linear subspaces are replaced by certain cones.

4. On the local Lipschitz continuity of the Kuhn–Tucker curve. Strong stability of a Kuhn-Tucker point implies continuity of the Kuhn-Tucker curve. This is an immediate consequence of Kojima's theory [9]. We show by means of application of Hager's theorem to certain canonical curves, that local Lipschitz continuity of the KT-curve is then automatically true. We mention that local Lipschitz continuity is also established by means of two different approaches, in fact by Kojima and Hirabayashi [10] and Robinson [13].

THEOREM 4.1. *Let $\bar{x}$ be a Kuhn-Tucker point for $P(\bar{t})$ which is strongly stable. Then, in some neighborhood $\mathcal{O}$ of $\bar{z} = (\bar{x}, \bar{t})$ the Kuhn-Tucker set can be parametrized as a continuous function of the parameter t, say $\{(x(t), t), t \in (a, b)\}$. Moreover, the function $t \to x(t)$ is locally Lipschitz continuous. The same holds w.r.t. the corresponding Lagrange parameters $\lambda_i(t)$, $\mu_j(t)$ under the convention that $\mu_j(t) = 0$ if $g_j(x(t), t) \neq 0$.*

For the proof of Theorem 4.1 we need the following simple lemma (see for example [5]).

LEMMA 4.1. *Let A be a symmetric $n \times n$-matrix, B be an $n \times k$-matrix, and let V be a matrix whose columns form a basis for* Ker B^T. *Then the matrix Q,*

$$Q = \begin{pmatrix} A & B \\ B^T & 0 \end{pmatrix},$$

is nonsingular iff rank $B = k$ *and $V^T A V$ is nonsingular.*

Proof of Theorem 4.1. As we mentioned in the beginning of this section we only show that the continuity of the Kuhn-Tucker curve [9] automatically implies local Lipschitz continuity. Note that for every $\tilde{J}$, $J_0^+(\bar{z}) \subset \tilde{J} \subset J_0(\bar{z})$, the point $\bar{x}$ is a critical point for $\tilde{P}(\bar{t})$:

$$\tilde{P}(\bar{t}): \quad \text{Minimize } f(x, \bar{t}) \text{ subject to } h_i(x, \bar{t}) = 0,\ i \in I,\ g_j(x, \bar{t}) = 0,\ j \in \tilde{J}.$$

Otherwise stated, $(\bar{x}, \bar{t}, \bar{\lambda}_i, i \in I, \bar{\mu}_j, j \in \tilde{J})$ is a zero point of the C^1-map $\tilde{\mathcal{F}}$:

$$\tilde{\mathcal{F}}: \begin{pmatrix} x \\ t \\ \lambda_i,\ i \in I \\ \mu_j,\ j \in \tilde{J} \end{pmatrix} \to \begin{pmatrix} D_x^T f(x, t) - \sum_{i \in I} \lambda_i D_x^T h_i(x, t) - \sum_{j \in \tilde{J}} \mu_j D_x^T g_j(x, t) \\ -h_i(x, t),\ i \in I \\ -g_j(x, t),\ j \in \tilde{J} \end{pmatrix}$$

where $\bar{\lambda}_i$, $i \in I$, $\bar{\mu}_j$, $j \in J_0(\bar{z})$ are the Lagrange parameters corresponding to $\bar{x}$ as a critical point for $P(\bar{t})$.

The map $\tilde{\mathcal{F}}$ is of class C^1. Since $\bar{x}$ is strongly stable, we see that the Jacobian matrix of $\tilde{\mathcal{F}}$ w.r.t. $(x, \lambda_i, i \in I, \mu_j, j \in \tilde{J})$ at $(\bar{x}, \bar{t}, \bar{\lambda}_i, i \in I, \bar{\mu}_j, j \in \tilde{J})$ is nonsingular, thereby using Lemma 4.1. By means of the implicit function theorem we obtain locally a unique C^1-curve $(\tilde{x}(t), t, \tilde{\lambda}_i(t), i \in I, \tilde{\mu}_j(t), j \in \tilde{J})$ along which $\tilde{\mathcal{F}}$ vanishes identically. In particular, $\tilde{x}(t)$, $\tilde{\lambda}_i(t)$, $i \in I$, $\tilde{\mu}_j(t)$, $j \in \tilde{J}$, are locally Lipschitz continuous. For t in some open neighborhood (a, b) of $\bar{t}$ the Kuhn-Tucker curve $\{(x(t), t) \mid t \in (a, b)\}$ obviously has the property that $x(t) = \tilde{x}(t)$, where $\tilde{x}(t)$ corresponds to some $\tilde{J}$, $J_0^+(\bar{z}) \subset \tilde{J} \subset J_0(\bar{z})$. But then, Theorem 2.2 implies that $t \to x(t)$ is locally Lipschitz continuous. With the convention that $\tilde{\mu}_j(t) = 0$ if $j \notin \tilde{J}$, the corresponding result for the Lagrange parameters follows in a similar way. □

5. On the piecewise C^1-differentiability of the Kuhn–Tucker curve. In general, under the assumptions of Theorem 4.1, local Lipschitz continuity of the Kuhn-Tucker curve will not imply piecewise C^1-differentiability (shortly PC^1). So, we need an additional condition in order to guarantee the latter fact. The idea behind such an additional condition comes from the recent investigations of Jongen et al. [7] where generic properties of (generalized) critical point sets are studied. On the other hand and independently from [7], this is also connected with assumptions on regular values of certain maps used by Kojima and Hirabayashi [11]. Let us start with an illustrating simple example.

Example 5.1. Consider $f(x, t) := x^2 - \phi(t)x \to \min$, $x \geqq 0$. Put $\phi(t) = t^6 \sin(1/t)$, $t \neq 0$ and $\phi(0) = 0$. Then, f is of class C^2. Clearly, $\bar{x} = 0$ is the unique minimum as the parameter t equals zero. Moreover, $\bar{x} = 0$ is then strongly stable, but the Kuhn-Tucker curve is merely local Lipschitz continuous and not PC^1. A similar example can be constructed with ϕ (and hence f) being of class C^∞. So, smoothness of the data is not sufficient for guaranteeing the KT-curve to be PC^1. Note, on the other hand, that a similar example cannot be constructed with ϕ analytic. □

Let $\bar{x}$ be a Kuhn-Tucker point for $P(\bar{t})$ which is strongly stable. In view of Theorem 4.1 the Kuhn-Tucker curve is locally Lipschitz continuous in a neighborhood of $\bar{z} = (\bar{x}, \bar{t})$. In order to study the local PC^1-differentiability we obviously may restrict ourselves to the case that $J_0^+(\bar{z}) \neq J_0(\bar{z})$, i.e. the strict complementarity is violated. In fact, if $J_0^+(\bar{z}) = J_0(\bar{z})$, then the KT-curve is locally already of class C^1. Now, we introduce the following conditions B1, B2.

Condition B1. $J_0(\bar{z}) \backslash J_0^+(\bar{z})$ is a singleton (i.e. the strict complementarity is violated for exactly one binding inequality constraint).

Without loss of generality, we put (if Condition B1 holds):

$$J_0(\bar{z}) = \{1, \cdots, p\}, \qquad J_0^+(\bar{z}) = \{1, \cdots, p-1\}, \tag{5.1}$$

$$B(\bar{z}) = -[D_x^T h_1, \cdots, D_x^T g_p]_{|\bar{z}}. \tag{5.2}$$

Condition B2. Adopting the notation of (5.1), (5.2) the matrix H is nonsingular, where

$$H = \begin{pmatrix} D_x^2 L & B & D_t D_x^T L \\ B^T & 0 & D_t \begin{Bmatrix} -h_i, i \in I \\ -g_j, j = 1, \cdots, p \end{Bmatrix} \\ 0 & (0, \cdots, 0, 1) & 0 \end{pmatrix}_{|\bar{z}}.$$

We need the following simple lemma (without proof).

LEMMA 5.1. *Let A be a nonsingular $n \times n$-matrix, $k > n$ and Q a $k \times k$-matrix:*

$$Q = \begin{pmatrix} A & C \\ E & D \end{pmatrix}.$$

Then Q is nonsingular iff $D - EA^{-1}C$ is nonsingular.

THEOREM 5.1. *Let $\bar{x}$ be a Kuhn-Tucker point for $P(\bar{t})$ which is strongly stable. If Conditions* B1, B2 *hold at $\bar{z} = (\bar{x}, \bar{t})$, then the Kuhn-Tucker curve is locally PC^1-differentiable.*

Proof. Let $\bar{\lambda}_i$, $i \in I$, $\bar{\mu}_j$, $j \in J_0(\bar{z})$ be the Lagrange parameters corresponding to $\bar{x}$. From B1 and the convention (5.1) we see that $\bar{\mu}_p = 0$ and $\bar{\mu}_j > 0$ for $j = 1, \cdots, p-1$. Let $\tilde{P}(t)$ be a problem the only difference of which from $P(t)$ is that the constraint g_p is considered as an equality constraint. Then, $\bar{x}$ is a Kuhn-Tucker point for $\tilde{P}(\bar{t})$ which is nondegenerate as a critical point. But then, in some open neighborhood of $\bar{z}$ the KT-curve w.r.t. $\tilde{P}(t)$ is of class C^1 (cf. proof of Theorem 4.1), say $t \to [\tilde{x}(t), t, \tilde{\lambda}_i(t), i \in I, \tilde{\mu}_j(t), j \in J_0(\bar{z})]$. For t in some neighborhood of $\bar{t}$ we observe: $(\tilde{x}(t), t)$ is a critical point for $P(t)$ as well and it is a Kuhn-Tucker point for $P(t)$ iff $\tilde{\mu}_p(t) \geqq 0$. Since $\tilde{\mu}_p(\bar{t}) = \bar{\mu}_p = 0$, a moment of reflection shows that we are done if we can show that $\dot{\tilde{\mu}}_p(\bar{t}) = d\mu_p(\bar{t})/dt \neq 0$.

The derivatives $\dot{\tilde{x}}(\bar{t})$, $\dot{\tilde{\lambda}}_i(\bar{t})$, $\dot{\tilde{\mu}}_j(\bar{t})$ are obtained as the solution of the system:

$$\begin{pmatrix} D_x^2 L & B \\ B^T & 0 \end{pmatrix}_{\bar{z}} \begin{pmatrix} \dot{\tilde{x}} \\ \dot{\tilde{\lambda}}_i,\ i = 1, \cdots, m \\ \dot{\tilde{\mu}}_j,\ j = 1, \cdots, p \end{pmatrix}_{\bar{t}} + \begin{pmatrix} D_t D_x^T L \\ -D_t h_i,\ i = 1, \cdots, m \\ -D_t g_j,\ j = 1, \cdots, p \end{pmatrix}_{\bar{z}} = 0, \tag{5.3}$$

with B as in (5.2). The system (5.3) is determined by means of the implicit function theorem, introducing a corresponding map $\tilde{\mathcal{T}}$ as in the proof of Theorem 4.1. We denote:

$$\begin{pmatrix} D_x^2 L & B \\ B^T & 0 \end{pmatrix}_{\bar{z}} = \begin{pmatrix} A & b \\ b^T & 0 \end{pmatrix}, \quad b \text{ the last column of } B(\bar{z}), \tag{5.4}$$

$$\begin{pmatrix} D_t D_x^T L \\ -D_t h_i,\ i \in I \\ -D_t g_j,\ j \in J_0(\bar{z}) \end{pmatrix}_{\bar{z}} = \begin{pmatrix} v \\ -D_t g_p(\bar{z}) \end{pmatrix}. \tag{5.5}$$

From Lemma 4.1, Condition B1 and the fact that $\bar{x}$ is a strongly stable KT-point it follows that in (5.4) both A and $\left(\begin{smallmatrix} A & b \\ b^T & 0 \end{smallmatrix}\right)$ are nonsingular.

Now we can compute:

$$\dot{\tilde{\mu}}_p(\bar{t}) = (b^T A^{-1} b)^{-1}(-D_t g_p(\bar{z}) - b^T A^{-1} v).$$

However, using Lemma 5.1, it follows that the matrix H (cf. Condition B2) is nonsingular iff $(b^T A^{-1} b)^{-1}(-D_t g_p(\bar{z}) - b^T A v) \neq 0$. This completes the proof of Theorem 5.1. □

Remark 5.1. In the proof of Theorem 5.1 we used the fact that $\dot{\tilde{\mu}}_p(\bar{t})$ is unequal to zero. It is easily seen that the corresponding derivative in Example 5.1 actually vanishes!

Acknowledgments. We would like to thank Bruno Brosowski who invited both of us to Oberwolfach where this work was initiated. Moreover, we are grateful to the anonymous referees. Their precise and positive criticism improved the quality and readability of our manuscript.

REFERENCES

[1] A. V. FIACCO, *Sensitivity analysis for nonlinear programming using penalty methods*, Math. Programming, 10 (1976), pp. 287-311.

[2] O. FUJIWARA, S. P. HAN AND O. L. MANGASARIAN, *Local duality of nonlinear programs*, this Journal, 22 (1984), pp. 162-169.

[3] W. W. HAGER, *Lipschitz continuity for constrained processes*, this Journal, 17 (1979), pp. 321-338.

[4] S. P. HAN AND O. L. MANGASARIAN, *Conjugate cone characterization of positive definite and semidefinite matrices*, Linear Algebra Appl., 56 (1984), pp. 89-103.

[5] S. P. HAN AND O. FUJIWARA, *An inertia theorem for symmetric matrices and its application to nonlinear programming*, Linear Algebra Appl., to appear.

[6] H. TH. JONGEN, P. JONKER AND F. TWILT, *One-parameter families of optimization problems: equality constraints*, Memorandum 420, Twente University of Technology, Enschede, the Netherlands, 1983, J. Optim. Theory Appl., special issue (A. V. Fiacco, ed.), to appear.

[7] ———, *Critical sets in parametric optimization*, Memorandum 433, Twente University of Technology, Enschede, the Netherlands, 1983; submitted.

[8] ———, *Nonlinear Optimization in R^n, I. Morse Theory, Chebyshev Approximation.* Methoden und Verfahren der Math. Physik, 29, Peter Lang Verlag, Frankfurt a.M., 1983.

[9] M. KOJIMA, *Strongly stable stationary solutions in nonlinear programs*, in Analysis and Computation of Fixed Points, S. M. Robinson, ed., Academic Press, New York, 1980.

[10] M. KOJIMA AND R. HIRABAYASHI, *Some results on the strong stability in nonlinear programs.* Technical Report 4, Dept. Management Science and Engineering, Tokyo Institute of Technology, 1980.

[11] ———, *Continuous deformations of nonlinear programs.* Math. Programming Study, 21 (1984), pp. 150-198.

[12] M. MARCUS AND H. MINC, *A Survey of Matrix Theory and Matrix Inequalities*, Allyn and Bacon, Boston, 1964.

[13] S. M. ROBINSON, *Strongly regular generalized equations*, Math. Oper. Res., 5 (1980), pp. 43-62.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 2, March 1986

THE GENERIC LOCAL TIME-OPTIMAL STABILIZING CONTROLS IN DIMENSION 3*

ALBERTO BRESSAN†

Abstract. This paper studies the control system

$$\dot{x}(t) = X(x(t)) + Y(x(t))u(t), \qquad X(p_0) = 0, |u(t)| \leqq 1,$$

where X and Y are $\mathscr{C}^\infty$ vector fields on a 3-dimensional manifold $\mathcal{M}$. Under generic assumptions on X, Y, the structure of the time-optimal stabilizing controls is completely determined in a neighborhood of p_0. The proofs rely on a systematic use of a local asymptotic approximation of X and Y by means of vector fields which generate a nilpotent Lie algebra.

Key words. nonlinear control system, time optimal trajectory, asymptotic nilpotent approximation

AMS(MOS) subject classifications. 49B10, 93C10

1. Introduction. Let $\mathcal{M}$ be a 3-dimensional manifold, $p_0 \in \mathcal{M}$ and let X, Y be smooth vector fields on $\mathcal{M}$ with $X(p_0) = 0$. Consider the control system

$$\begin{aligned} \dot{y}(t) &= X(y(t)) + Y(y(t))u(t), \\ y(0) &= p_0, \end{aligned} \tag{1.1}$$

where the scalar control $u(\cdot)$ is measurable and satisfies $|u(t)| \leqq 1$ almost everywhere. This paper provides a description of all admissible controls that steer the system (1.1) in minimum time from p_0 to any point p in a neighborhood of p_0. We show that the structure of the local time-optimal trajectories is completely determined by the Lie brackets up to order three of X and Y at p_0, under the generic assumptions

(A1) The vectors Y, $[Y, X]$ and $[[Y, X], X]$ are linearly independent at p_0,

(A2) $[Y, [Y, X]](p_0) = \bar{k}_1 Y(p_0) + \bar{k}_2 [Y, X](p_0) + \bar{k}_3 [[Y, X], X](p_0)$ with $|\bar{k}_3| \neq 1$.

For the system (1.1), a numerical algorithm yielding a stabilizing control was studied in [7]. Sussmann [12] provided a complete description of time-optimal trajectories for analytic systems in the plane. The present work is part of a general program of research whose goal is to determine the local properties of control systems of the form (1.1) from the linear relations among the Lie brackets of X and Y at p_0. Our main technique is the local approximation of (1.1) by means of a nilpotent system defined on the same state space [1]. Somewhat different approximations were discussed in [3], [6] and applied in [8], [11] to obtain results on local controllability. From (1.1), a suitable rescaling of time and space coordinates leads us to the system

$$\begin{aligned} (\dot{x}_1, \dot{x}_2, \dot{x}_3) &= (u, x_1, x_2 + kx_1^2/2) + h(x), \\ (x_1, x_2, x_3)(0) &= (0, 0, 0), \qquad t \in [0, 1], \end{aligned} \tag{1.2}$$

where $k = \bar{k}_3$, and the vector field $h(\cdot)$ is as small as we please, together with all of its high-order partial derivatives. In the special case $h \equiv 0$, the trajectories of (1.2) are easily computed as integrals of the control. The time-optimal controllability problem can therefore be explicitly solved applying Pontryagin's maximum principle. We use the directional convexity of the reachable set and a global necessary condition [2] to rule out the optimality of bang-bang controls with more than two switchings. In the

* Received by the editors July 17, 1984, and in revised form December 20, 1984. This research was sponsored by the U.S. Army under contract DAAG29-80-C-0041.

† Istituto di Matematica Applicata, Università di Padova, Padova 35100, Italy.

general case, h can be regarded as a small perturbation. Repeated applications of the implicit function theorem complete the proof. The asymptotic approximation technique used here appears to be quite general and might be effective in the study of higher dimensional systems as well.

2. The main theorem. As a preliminary, note that if (A1) holds, by the implicit function theorem the equation

$$[Y,[Y,X]](y) = k_1(y)\,Y(y) + k_2(y)[Y,X](y) + k(y)[[Y,X],X](y) \tag{2.1}$$

uniquely defines the smooth functions $k_i(y)$ in a neighborhood V of p_0. If (A2) holds with $|\bar{k}_3| > 1$, we can also assume $|k_3(y)| > 1$ for all $y \in V$. Two special families of trajectories will be considered.

DEFINITION. Let $y(\cdot)$ be an absolutely continuous map from $[0, T]$ into $\mathcal{M}$ with $y(0) = p_0$. We say that y is a BBB-trajectory for the system (1.1) if there exist $0 \leqq \tau_1 \leqq \tau_2 \leqq T$ such that

$$\dot{y} = X(y) + Y(y) \quad \text{or} \quad \dot{y} = X(y) - Y(y) \tag{2.2}$$

on each one of the (possibly empty) subintervals $(0, \tau_1)$, (τ_1, τ_2), (τ_2, T). We call $y(\cdot)$ a BSB-trajectory if there exist $0 \leqq \tau_1 < \tau_2 \leqq T$ such that (2.2) holds on $(0, \tau_1)$ and on (τ_2, T), while

$$\dot{y} = X(y) + k_3^{-1}(y)\,Y(y) \tag{2.3}$$

on (τ_1, τ_2).

Our main result states that the bang-bang and the partially singular trajectories just defined are locally the only optimal ones.

THEOREM 1. *Consider the system* (1.1) *and let* (A1), (A2) *hold.*

i) *If* $|\bar{k}_3| < 1$, *then there exists a neighborhood* V *of* p_0 *in* $\mathcal{M}$ *such that every time-optimal trajectory steering* p_0 *to a point* $p \in V$ *is a BBB-trajectory.*

ii) *If* $|\bar{k}_3| > 1$, *then there exists a neighborhood* V *of* p_0 *such that every trajectory steeering* p_0 *to a point* $p \in V$ *in minimum time is either a BBB- or a BSB-trajectory.*

By inverting time and the vector fields X, Y, Theorem 1 thus yields the solution of the generic local time-optimal stabilization problem in dimension three. A noteworthy consequence is that, at least for analytic X and Y, this solution can be written in regular feedback form [13]. When $|\bar{k}| < 1$, (1.1) behaves essentially like a linear system. Part i) in Theorem 1 could already be deduced from [10]. When $|\bar{k}_3| > 1$, the nonlinearities begin to play a major role, and a careful analysis is required. In §§ 3, 4 we prove that Theorem 1 is a consequence of an analogous result (Theorem 2) concerning the system (1.2). The main steps in the proof of Theorem 2 are collected in § 5. Technical details are then worked out in §§ 6–10, which may be skipped in a first reading.

3. An equivalent result. By introducing a suitable set of coordinates, (1.1) will be transformed into a more tractable system on $\mathbb{R}^3$. In the following, the variable in $\mathbb{R}^3$ is $x = (x_1, x_2, x_3)$ and $\{e_1, e_2, e_3\}$ denotes the canonical orthonormal basis. Given a smooth vector field $g = (g_1, g_2, g_3)$ on $\mathbb{R}^3$, its partial derivatives are written

$$g_{i,j} = \frac{\partial g_i}{\partial x_j}, \quad g_{i,jk} = \frac{\partial^2 g_i}{\partial x_j\,\partial x_k}, \quad \cdots;$$

∇g denotes the 3×3 matrix $(g_{i,j})$ of first order partials of g. Consider the map

$$\theta\colon (s_1, s_2, s_3) \to (\exp s_1 Y) \cdot (\exp s_2 [Y,X]) \cdot (\exp s_3 [[Y,X],X])(p_0), \tag{3.1}$$

where $(\exp sZ)(p)$ is the value at time s of the solution of the Cauchy problem

$$\dot{y}(t) = Z(y(t)), \qquad y(0) = p \in \mathcal{M}.$$

Because of (A1), θ defines a local chart of a neighborhood of p_0. In this chart, the system (1.1) becomes

$$\dot{x} = f(x) + e_1 u, \qquad x(0) = 0 \in \mathbb{R}^3. \tag{3.2}$$

The vector field f can be written in the form

$$f(x) = (\bar{k}_1 x_1^2/2,\ x_1 + \bar{k}_2 x_1^2/2,\ x_2 + \bar{k}_3 x_1^2/2) + \tilde{f}(x) \tag{3.3}$$

with $\tilde{f}_{i,j}(0) = \tilde{f}_{i,11}(0) = 0$ for $i = 1, 2, 3$, $j = 1, 2$.

Since the problem is local, we can assume that θ is defined on some open ball $B_r \subseteq \mathbb{R}^3$ centered at the origin with radius r, and that f can be extended outside B_r to a $\mathcal{C}^\infty$ vector field, still called f, with compact support. We now apply to (3.2) the asymptotic rescaling procedure discussed in [1]. Consider the orthogonal decomposition $\mathbb{R}^3 = W_1 \oplus W_2 \oplus W_3$ with $W_i = \{\xi e_i;\ \xi \in \mathbb{R}\}$. Let $\pi_i : \mathbb{R}^3 \to W_i$ be the canonical projections. Given an admissible control $u(\cdot)$, let $t \to x(u, t)$ be the corresponding trajectory of (3.2). If u is defined on the time-interval $[0, \varepsilon]$, construct the rescaled control $u_\varepsilon : [0, 1] \to \mathbb{R}$ by setting $u_\varepsilon(t) = u(\varepsilon t)$. Moreover, set

$$x^\varepsilon(u_\varepsilon, t) = \sum_{i=1}^{3} \varepsilon^{-i} \pi_i(x(u, \varepsilon t)). \tag{3.4}$$

A direct computation shows that x^ε is the response of the system

$$\dot{x}(t) = f^\varepsilon(x(t)) + e_1 u_\varepsilon(t), \qquad x(0) = 0 \in \mathbb{R}^3 \tag{3.5}$$

with $f^\varepsilon = (f_1^\varepsilon, f_2^\varepsilon, f_3^\varepsilon)$,

$$f_i^\varepsilon(x) = \varepsilon^{1-i} f_i\left(\sum_{j=1}^{3} \varepsilon^j \pi_j(x)\right). \tag{3.6}$$

For every $\varepsilon > 0$, (3.5) is merely a linear rescaling of (3.2). Therefore, a control u is time-optimal for (3.2) on $[0, \varepsilon]$ if and only if the corresponding u_ε is time-optimal for (3.5) on $[0, 1]$. Because of (3.3), the main result proved in [1] now implies that, as $\varepsilon \to 0$, f^ε converges to the vector field

$$\bar{f}(x) = (0,\ x_1,\ x_2 + \bar{k}_3 x_1^2/2) \tag{3.7}$$

together with all partial derivatives, uniformly on bounded sets. Theorem 1 thus becomes a consequence of the following result concerning the system (1.2). If $k \geqq 0$, we write Ω_k for the open box $(-2, 2) \times (-1, 1) \times (-1 - k, 1 + k) \subset \mathbb{R}^3$, $\mathcal{C}^3(\Omega_k)$ for the Banach space of three times continuously differentiable vector fields on Ω_k, and we let $\mathcal{F}$ be the family of all neighborhoods of the null vector field in $\mathcal{C}^3(\Omega_k)$.

THEOREM 2. a) *If $0 \leqq k < 1$, then there exists $\mathcal{V} \in \mathcal{F}$ such that for all $h \in \mathcal{V}$, $0 \leqq T \leqq 1$, every time-optimal control $u(\cdot)$ for (1.2) on $[0, T]$ is bang-bang with at most two switchings.*

b) *If $k > 1$, then there exists $\mathcal{V} \in \mathcal{F}$ such that, given any $h \in \mathcal{V}$, every time-optimal control u for (1.2) on $[0, T] \subseteq [0, 1]$ has the following property. Either u is bang-bang with finitely many switchings on $[0, T]$, or there exists $0 \leqq t_1 < t_2 \leqq T$ such that $u(t)$ is constantly equal to $+1$ or -1 on $[0, t_1]$ and on $[t_2, T]$, while $u(t) = k_3^{-1}(x(t))$ on (t_1, t_2). Here $k_3(x)$ is the third coefficient in the linear relation*

$$[e_1, [e_1, g]](x) = k_1(x) e_1 + k_2(x)[e_1, g](x) + k_3(x)[[e_1, g], g](x), \tag{3.8}$$

with $g = \bar{f} + h$.

c) *If $k>1$, then there exists $\mathcal{V}\in\mathcal{F}$ such that, if $h\in\mathcal{V}$ and u is a bang-bang control with initial switchings at times $0<t_1<t_2<t_3=1$, then u is not time-optimal for* (1.2) *after time* 1.

As usual, statements concerning controls in $\mathcal{L}^1$ are always meant "up to $\mathcal{L}^1$-equivalence."

4. Proof of Theorem 1. Let Theorem 2 hold. By possibly replacing Y with $-Y$ in (A2) we can assume $\bar{k}_3\geqq 0$. Consider the case $0\leqq\bar{k}_3<1$ first. Set $k=\bar{k}_3$ and choose the neighborhood $\mathcal{V}\in\mathcal{F}$ according to a) in Theorem 2. Choose $\varepsilon>0$ so small that the reachable set at time ε for the sytem (1.1) is contained within the range of the chart θ, i.e. $R(\varepsilon)\subset\theta(B_r)$, and such that $\varepsilon\Omega\subset B_r$, $h=f^\varepsilon-\bar{f}\in\mathcal{V}$. This is possible because, as $\varepsilon\to 0$, the convergence of f^ε to $\bar{f}$ in (3.6), (3.7) is uniform on the bounded set Ω_k [1]. If the control u steers the system (1.1) from p_0 to some point $p\in R(\varepsilon)$ in minimum time $\eta\leqq\varepsilon$, then the control $t\to u_\varepsilon(t)=u(\varepsilon t)$ is time optimal for the system (1.2) on the interval $[0,\eta\varepsilon^{-1}]\subseteq[0,1]$. By a) in Theorem 2, u_ε is bang-bang with at most two switchings, hence the same holds for u. Taking $V=R(\varepsilon)$, this proves i) in Theorem 1.

The proof of ii) is similar. If $\bar{k}_3>1$, set $k=\bar{k}_3$ and choose $\mathcal{V}\in\mathcal{F}$ according to b) and c) in Theorem 2. Choose $\varepsilon>0$ such that $R(\varepsilon)\subset\theta(B_r)$, $\varepsilon\Omega_k\subset B_r$, $f^\eta-\bar{f}\in\mathcal{V}$ for every $\eta\in[0,\varepsilon]$. If $0<\eta\leqq\varepsilon$ and the control u is time-optimal for (1.1) on $[0,\eta]$, then, setting $h=f^\varepsilon-\bar{f}$, the control $t\to u_\varepsilon(t)=u(\varepsilon t)$ is optimal for (1.2) on $[0,\eta\varepsilon^{-1}]\subseteq[0,1]$. By b) in Theorem 2, either u_ε is partly singular, or u_ε is bang-bang with finitely many switchings, hence the same holds for u. In the first case, comparing (3.8) with (2.1) one concludes that u generates a BSB-trajectory, because the linear relations among the Lie brackets of the vector fields f, e_1 are preserved under the transformation (3.6). In the second case, if u has more than two switchings inside $[0,\eta]$, let $0<t_1<t_2<t_3=\eta'<\eta$ be its first three switching times. The control $t\to u_{\eta'}(t)=u(\eta' t)$ has then its third switch at $t=1$. Since $f^{\eta'}-\bar{f}\in\mathcal{V}$, using c) we see that $u_{\eta'}$ is not optimal after time 1, hence u is not optimal at time $\eta>\eta'$, a contradiction. Taking $V=R(\varepsilon)$, this completes the proof of part ii).

5. Sketch of the proof of Theorem 2. In the following, we denote $\bar{f}(x)$ the vector field with components $(0, x_1, x_2+kx_1^2/2)$; h is the small perturbation and $g=\bar{f}+h$. We write B_ε for the open ball centered at the origin with radius ε. When $h\equiv 0$, the exact solution of (1.2) is

$$
\begin{aligned}
x_1(u,t)&=\int_0^t u(s)\,ds,\\
x_2(u,t)&=\int_0^t (t-s)u(s)\,ds,\\
x_3(u,t)&=\frac{1}{2}\int_0^t (t-s)^2u(s)\,ds+\frac{k}{2}\int_0^t\left(\int_0^s u(r)\,dr\right)^2 ds.
\end{aligned}
\tag{5.1}
$$

If u is an admissible control, i.e. if $|u(t)|\leqq 1$ almost everywhere, then for $t\in[0,1]$ the trajectory $t\to x(u,t)$ is contained inside the closed box $[-1,1]\times[-\frac{1}{2},\frac{1}{2}]\times[-(k+1)/6,(k+1)/6]$. By a classical perturbation theorem [5], there exists a bounded neighborhood $\mathcal{V}_0\in\mathcal{F}$ such that, if $h\in\mathcal{V}_0$, every admissible trajectory for (1.2) remains inside Ω_k during the time interval $[0,1]$. The neighborhood $\mathcal{V}_0$ now chosen will be kept fixed throughout. The first part of our proof will single out all solutions of the

Pontryagin's equations for (1.2) on any interval $[0, T] \subseteq [0, 1]$.

$$(5.2)_1 \quad (\dot{x}_1, \dot{x}_2, \dot{x}_3) = \left(u + h_1(x), x_1 + h_2(x), x_2 + \frac{k}{2}x_1^2 + h_3(x)\right),$$

$$(5.2)_2 \quad (\dot{\lambda}_1, \dot{\lambda}_2, \dot{\lambda}_3) = -\left(\lambda_2 + kx_1\lambda_3 + \sum_{i=1}^{3} h_{i,1}\lambda_i, \lambda_3 + \sum_{i=1}^{3} h_{i,2}\lambda_i, \sum_{i=1}^{3} h_{i,3}\lambda_i\right),$$

$$(5.2)_3 \quad (x_1, x_2, x_3)(0) = (0, 0, 0), \qquad (\lambda_1, \lambda_2, \lambda_3)(T) = (\bar{\lambda}_1, \bar{\lambda}_2, \bar{\lambda}_3),$$

$$(5.2)_4 \quad u(t) \in \operatorname{sgn} \lambda_1(t) \quad \text{a.e. on } [0, T],$$

where $\bar{\lambda} = (\bar{\lambda}_1, \bar{\lambda}_2, \bar{\lambda}_3) \neq (0, 0, 0)$, $0 < T \leqq 1$ and the convention $\operatorname{sgn} 0 = [-1, 1]$ is used. Notice that for every data $\bar{\lambda}$ and T, $(5.2)_{1\text{-}4}$ has at least one solution. Indeed, the compactness of the reachable set $R(T)$ implies the existence of a control $\tilde{u}$ for which $x(\tilde{u}, T) = \max\{\langle \bar{\lambda}, x\rangle; x \in R(T)\}$. Such $\tilde{u}$ clearly yields a solution of (5.2). Different types of extremal controls arise, depending on the direction of $\bar{\lambda}$.

PROPOSITION 1. *There exists $\mathcal{V}_1 \in \mathcal{F}$ such that, if $h \in \mathcal{V}_1$ and $\bar{\lambda}_3^2 \leqq (12k+16)^{-2}(\bar{\lambda}_1^2 + \bar{\lambda}_2^2)$, then the solution (u, x, λ) of (5.2) is unique and the corresponding control u is bang-bang with at most one switching.*

PROPOSITION 2. *For every $\varepsilon > 0$ there exists $\mathcal{V}_2 \in \mathcal{F}$ such that, if $h \in \mathcal{V}_2$ and $\bar{\lambda}_3^2 \geqq (12k+16)^{-2}(\bar{\lambda}_1^2 + \bar{\lambda}_2^2)$, then any solution (u, x, λ) of (5.2) satisfies*

$$\ddot{\lambda}_1(t) \in [(1 - k \operatorname{sgn} \lambda_1(t)) + B_\varepsilon]\bar{\lambda}_3 \quad \text{a.e. on } [0, T]. \tag{5.3}$$

The two above results together imply part a) of Theorem 2. Indeed, let $0 \leqq k < 1$ and choose the neighborhoods $\mathcal{V}_1$, $\mathcal{V}_2$ according to Propositions 1 and 2 with $\varepsilon = (1-k)/2$. If $h \in \mathcal{V}_1 \cap \mathcal{V}_2$ and if (u, x, λ) is a solution of (5.2), then either $\bar{\lambda}_3^2 \leqq (12k+16)^{-2}(\bar{\lambda}_1^2 + \bar{\lambda}_2^2)$ and by Proposition 1 u is bang-bang with at most one switching, or $\bar{\lambda}_3^2 \geqq (12k+16)^{-2}(\bar{\lambda}_1^2 + \bar{\lambda}_2^2)$. In this case, by (5.3) and the choice of ε, $\ddot{\lambda}_1(t)$ has a.e. the same sign of $\lambda_3(T) = \bar{\lambda}_3 \neq 0$. Hence λ_1 is either strictly concave or strictly convex on $[0, T]$ and can vanish at most at two distinct points. The corresponding control u is therefore bang-bang with no more than two switchings. Next, we assume $k > 1$ and study the case where the third component of $\bar{\lambda}$ is large compared with the others.

PROPOSITION 3. *If $k > 1$, there exists $\mathcal{V}_3 \in \mathcal{F}$ such that every solution (u, x, λ) of (5.2) with $h \in \mathcal{V}_3$, $\bar{\lambda}_3^2 \geqq (12k+16)^{-2}(\bar{\lambda}_1^2 + \bar{\lambda}_2^2)$, $\bar{\lambda}_3 < 0$, has the following property. There exist $0 \leqq \tau_1 \leqq \tau_2 \leqq T$ such that u is constantly equal to $+1$ or -1 on $[0, \tau_1]$ and on $[\tau_2, T]$, while $u(t) = k_3^{-1}(x(t))$ on (τ_1, τ_2). Here $k_3(x)$ is the scalar function defined at (3.8).*

PROPOSITION 4. *If $k > 1$, there exists $\mathcal{V}_4 \in \mathcal{F}$ such that, for every solution (u, x, λ) of (5.2) with $h \in \mathcal{V}_4$, $\bar{\lambda}_3^2 \geqq (12k+16)^{-2}(\bar{\lambda}_1^2 + \bar{\lambda}_2^2)$ and $\bar{\lambda}_3 > 0$, either the control u is bang-bang with finitely many switchings on $[0, T]$, or $u(t) = k_3^{-1}(x(t))$ throughout $[0, T]$.*

Propositions 1, 3 and 4 clearly imply part b) of Theorem 2. To prove c), define the set of vectors

$$\Lambda = \{w = (w_1, w_2, w_3) \in \mathbb{R}^3;\ w_3^2 \geqq (12k+16)^2(w_1^2 + w_2^2)\}.$$

Choose $\mathcal{V}_1 \in \mathcal{F}$ according to Proposition 1. An application of Theorem 2 in [2] yields

COROLLARY 1. *If $h \in \mathcal{V}_1$, then the reachable set $R(1)$ for the system (1.2) is Λ-convex, i.e. $R(1)$ contains the point $\xi p + (1-\xi)q$ whenever $p, q \in R(1)$, $\xi \in [0, 1]$ and $p - q \in \Lambda$.*

Let now u be a bang-bang control satisfying Pontryagin's conditions and having a third switch at time $t = 1$. To prove that the value $x(u, 1)$ of the corresponding trajectory at time 1 lies in the interior of $R(1)$, it suffices to exhibit a second admissible control, say u', such that

$$x_i(u', 1) = x_i(u, 1) \quad \text{for } i = 1, 2, \qquad x_3(u', 1) > x_3(u, 1). \tag{5.4}$$

Indeed, if (u, x, λ) is a solution of (5.2) and if the vector field $h(\cdot)$ is sufficiently small, then $\lambda_3(1) > 0$ because of Propositions 1 and 3. The vector $w = x(u', 1) - x(u, 1) = (0, 0, x_3(u', 1) - x_3(u, 1))$ therefore has a positive inner product with $\lambda(1)$ and lies in the interior of Λ. By [2, Thm. 1], $x(u, 1) \in \text{int } R(1)$. To complete the proof, we only need to show that such a control u' always exists. For $a, b, c \geqq 0$ define the control $u^+ = u^+(a, b, c)$ by setting

$$
\begin{aligned}
&u^+(a, b, c)(t) = 1 \qquad \text{for } t \in [0, a) \cup [a+b, a+b+c),\\
&u^+(a, b, c)(t) = -1 \quad \text{for } t \in [a, a+b) \cup [a+b+c, \infty).
\end{aligned} \tag{5.5}
$$

If $\alpha, \beta, \gamma \geqq 0$, define $u^-(\alpha, \beta\, \gamma)(t) = -u^+(\alpha, \beta, \gamma)(t)$. Call $x^+ = x^+(a, b, c)$ the point reached by the system (1.2) at time $T = a+b+c$, subject to the control $u^+(a, b, c)$ and define $x^- = x^-(\alpha, \beta, \gamma)$ similarly. In the special case $h \equiv 0$, the components of x^+, x^- can be explicitly computed from (5.1):

$$
\begin{aligned}
&x_1^+ = a - b + c, \qquad x_2^+ = (a+b+c)^2/2 - (b+c)^2 + c^2,\\
&x_3^+ = \tfrac{1}{3}\{\tfrac{1}{2}(a+b+c)^3 - (b+c)^3 + c^3 + k[a^3 + (b-a)^3 + \tfrac{1}{2}(c-b+a)^3]\},\\
&x_1^- = -\alpha + \beta - \gamma, \qquad x_2^- = -(\alpha+\beta+\gamma)^2/2 + (\beta+\gamma)^2 - \gamma^2,\\
&x_3^- = \tfrac{1}{3}\{-\tfrac{1}{2}(\alpha+\beta+\gamma)^3 + (\beta+\gamma)^3 - \gamma^3 + k[\alpha^3 + (\beta-\alpha)^3 + \tfrac{1}{2}(\gamma-\beta+\alpha)^3]\}.
\end{aligned} \tag{5.6}
$$

The three conditions

$$
x_1^+ = x_1^-, \quad x_2^+ = x_2^-, \quad a+b+c = \alpha+\beta+\gamma = T \tag{5.7}
$$

imply the relations

$$
\alpha = bc/(a+c), \quad \beta = a+c, \quad \gamma = ab/(a+c), \tag{5.8}
$$

$$
a = \beta\gamma/(\alpha+\gamma), \quad b = \alpha+\gamma, \quad c = \alpha\beta/(\alpha+\gamma). \tag{5.9}
$$

When these are satisfied, we have $\Delta x = x^+(a, b, c) - x^-(\alpha, \beta, \gamma) = (0, 0, x_3^+ - x_3^-)$ and a direct calculation (see Appendix) shows that

$$
\begin{aligned}
x_3^+ - x_3^- &= [(a+b+c) - k(a-b+c)]abc/(a+c)\\
&= [(\alpha+\beta+\gamma) + k(\alpha-\beta+\gamma)]\alpha\beta\gamma/(\alpha+\gamma).
\end{aligned} \tag{5.10}
$$

If $a, b, c > 0$ and $u^+(a, b, c)$ satisfies the maximum principle on $[0, T+\varepsilon]$ for some $\varepsilon > 0$, then the corresponding adjoint variable λ in (5.2) satisfies

$$
\begin{aligned}
&\lambda_3(t) = \bar{\lambda}_3 > 0 \quad \forall t \in [0, T],\\
&\lambda_1(a) = \lambda_1(a+b) = \lambda_1(a+b+c) = 0,\\
&\ddot{\lambda}_1(t) = (1+k)\bar{\lambda}_3 \quad \text{for } t \in (a, a+b),\\
&\ddot{\lambda}_1(t) = (1-k)\bar{\lambda}_3 \quad \text{for } t \in (a+b, a+b+c).
\end{aligned}
$$

The above relations imply $(k+1)b = (k-1)c$. Using this equality in (5.10) we obtain

$$
x_3^+ - x_3^- = (1-k)a^2bc/(a+c) < 0. \tag{5.11}
$$

If $u = u^+(a, b, c)$, consider the control $u' = u^-(\alpha, \beta, \gamma)$ with α, β, γ defined at (5.8). When $T = a+b+c = 1$, (5.7) and (5.11) imply (5.4). Therefore u cannot be optimal after time $T = 1$. The case where the bang-bang control u takes initially the value -1 can be treated similarly. Let $u = u^-(\alpha, \beta, \gamma)$ for some $\alpha, \beta, \gamma > 0$. If Pontryagin's equations (5.2) are satisfied, then $(k-1)\beta = (k+1)\gamma$. Consider the control $u' =$

$u^+(a, b, c)$ with a, b, c defined in terms of α, β, γ at (5.9). From (5.10) and the above equality we now obtain

$$x_3^+ - x_3^- = (k+1)\alpha^2\beta\gamma/(\alpha+\gamma) > 0. \tag{5.12}$$

When $T = \alpha+\beta+\gamma = 1$, (5.7) and (5.12) imply (5.4). Therefore $u = u^-$ cannot be optimal after time $T = 1$. This establishes part c) of Theorem 2 in the case $h \equiv 0$. Thanks to the implicit function theorem, the above arguments remain valid when a small perturbation h is added to the vector field $\bar{f}$ in (1.2).

PROPOSITION 5. *There exists $\mathcal{V}_5 \in \mathcal{F}$ such that, if $h \in \mathcal{V}_5$ and if u is a bang-bang control with initial switchings at time t_i: $0 < t_1 < t_2 < t_3 = 1$ which satisfies Pontryagin's equations* (5.2) *on* $[0, 1]$ *with* $\lambda_1(1) = 0$, *then there exists a second admissible control u' such that* (5.4) *holds.*

This will complete the proof of Theorem 2.

6. Proof of Proposition 1.

LEMMA 1. *Let $k \geqq 0$, $\lambda \in \mathbb{R}^3$ with $|\lambda| = (\lambda_1^2+\lambda_2^2+\lambda_3^2)^{1/2} = 1$. Set $\eta = (12k+16)^{-1}$ and assume $\lambda_3^2 \leqq \eta^2(\lambda_1^2+\lambda_2^2)$. Then at least one of the following holds*

i) $|\lambda_1| \geqq |\lambda_2| + (2k+1)|\lambda_3| + (2k+4)\eta$;

ii) $|\lambda_2| \geqq (2k+1)|\lambda_3| + (2k+4)\eta$.

Indeed, if ii) fails, since $|\lambda_3| \leqq \eta$ we have

$$\begin{aligned} |\lambda_1| &\geqq 1 - |\lambda_2| - |\lambda_3| \geqq 1 - [(2k+1)|\lambda_3| + (2k+4)\eta] - \eta \\ &\geqq (8k+10)\eta \geqq |\lambda_2| + (2k+1)|\lambda_3| + (2k+4)\eta. \end{aligned}$$

LEMMA 2. *There exists a constant $M > 0$ such that every solution (u, x, λ) of $(5.2)_{1\text{-}4}$ with $|\bar{\lambda}| = 1$, $h \in \mathcal{V}_0$, satisfies*

$$M^{-1} \leqq |\lambda(t)| \leqq M \quad \forall t \in [0, T], \tag{6.1}$$

$$|\dot{x}_i(t)| \leqq M, \qquad |\dot{\lambda}_i(t)| \leqq M, \qquad i = 1, 2, 3, \quad t \in [0, T]. \tag{6.2}$$

Proof. Since $\mathcal{V}_0$ is bounded in $\mathcal{C}^3(\Omega_k)$, the operator norms of the matrices $\nabla g(x)$ of first order partial derivatives of $g = \bar{f} + h$ satisfy a uniform bound, say $|\nabla g(x)| \leqq N$, for all $h \in \mathcal{V}_0$, $x \in \Omega_k$.

By $(5.2)_2$, (6.1) holds with $M = e^N$. The bounds in (6.2) follows from $(5.2)_{1\text{-}2}$ and (6.1), with a possibly larger constant M.

To prove Proposition 1, it clearly suffices to consider the case $|\bar{\lambda}| = 1$. Set $\eta = (12k+16)^{-1}$ and define $\eta' = \eta/3M$, with M being the constant in (6.1), (6.2). Choose a neighborhood $\mathcal{V}_1 \subseteq \mathcal{V}_0$ in $\mathcal{F}$ such that $|h_{i,j}(x)| < \eta'$ for all $x \in \Omega_k$, $h \in \mathcal{V}_1$, $i, j \in \{1, 2, 3\}$. By Lemma 1, two cases must be considered.

Case 1. Let $|\bar{\lambda}_1| \geqq |\bar{\lambda}_2| + (2k+1)|\bar{\lambda}_3| + (2k+4)\eta$. Then for $t \in [0, T] \subseteq [0, 1]$, using $(5.2)_2$ we obtain

$$|\dot{\lambda}_3(t)| \leqq 3\eta' M = \eta,$$

$$|\lambda_3(t)| \leqq |\bar{\lambda}_3| + \eta, \tag{6.3}$$

$$|\dot{\lambda}_2(t)| \leqq |\bar{\lambda}_3| + \eta + 3\eta' M, \tag{6.4}$$

$$|\lambda_2(t)| \leqq |\bar{\lambda}_2| + |\bar{\lambda}_3| + 2\eta,$$

$$|\dot{\lambda}_1(t)| \leqq |\bar{\lambda}_2| + |\bar{\lambda}_3| + 2\eta + 2k(|\bar{\lambda}_3| + \eta) + 3\eta' M,$$

$$|\lambda_1(t)| \geqq |\bar{\lambda}_1| - (|\bar{\lambda}_2| + |\bar{\lambda}_3| + 2\eta) - 2k(|\bar{\lambda}_3| + \eta) - \eta \geqq \eta > 0.$$

Therefore $\lambda_1(t) \neq 0$ throughout the interval $[0, T]$. From $(5.2)_4$ we deduce $u(t) =$

sgn $\lambda_1(t) = \text{sgn}\,\bar\lambda_1$. The control u is thus uniquely determined and constant throughout $[0, T]$.

Case 2. Let $|\bar\lambda_2| \geqq (2k+1)|\bar\lambda_3| + (2k+4)\eta$. From $(5.2)_{1\text{-}2}$, using (6.3) and (6.4) we now obtain

$$\begin{aligned} &|\lambda_2(t)| \geqq |\bar\lambda_2| - |\bar\lambda_3| - 2\eta, \\ &|\dot\lambda_1(t)| \geqq (|\bar\lambda_2| - |\bar\lambda_3| - 2\eta) - 2k(|\bar\lambda_3| + \eta) - 3\eta' M \geqq \eta > 0. \end{aligned} \tag{6.5}$$

By (6.5), $\lambda_1(\cdot)$ is a strictly monotone function, with at most one zero. By $(5.2)_4$, the corresponding control $u(\cdot)$ is bang-bang with at most one switching inside $[0, T]$. We claim that such a control u is unique, whenever $h \in \mathscr{V}_1$, for a suitably small neighborhood $\mathscr{V}_1 \in \mathscr{F}$. To set the ideas, assume $\bar\lambda_2 > 0$, the case $\bar\lambda_2 < 0$ being entirely analogous. Define the set

$$\Gamma = \{\lambda \in \mathbb{R}^3;\ |\lambda| = 1,\ \lambda_3^2 \leqq \eta^2(\lambda_1^2 + \lambda_2^2),\ \lambda_2 \geqq (2k+1)|\lambda_3| + (2k+4)\eta\}$$

and fix $\bar\lambda \in \Gamma$, $0 < T \leqq 1$. For $\tau \in [0, T]$ define the control $u(\tau, \cdot)$ by setting $u(\tau, t) = 1$ when $t \in [0, \tau]$, $u(\tau, t) = -1$ when $t \in (\tau, T]$, and let $x(\tau, \cdot)$, $\lambda(\tau, \cdot)$ be the solutions of $(5.2)_{1\text{-}3}$ corresponding to the control $u(\tau, \cdot)$. Since $\bar\lambda \in \Gamma$, we already know that any solution of $(5.2)_{1\text{-}4}$ is of the form $(u(\tau, \cdot), x(\tau, \cdot), \lambda(\tau, \cdot))$ for some $\tau \in [0, T]$. Notice that $(5.2)_4$ holds iff either $\tau = 0$ and $\lambda_1(0, 0) \leqq 0$, or $0 < \tau < T$ and $\lambda_1(\tau, \tau) = 0$, or $\tau = T$ and $\lambda_1(T, T) \geqq 0$. Uniqueness will be established by proving that

$$\frac{d}{d\tau}\lambda_1(\tau, \tau) < 0 \quad \forall \tau \in [0, T]. \tag{6.6}$$

When $h \equiv 0$ in (1.2), a direct calculation yields

$$\begin{aligned} &x_1(\tau, s) = 2\tau - s \quad \forall s \in [\tau, T], \\ &\lambda_3(\tau, s) = \bar\lambda_3, \qquad \lambda_2(\tau, s) = \bar\lambda_2 + (T-s)\bar\lambda_3, \\ &\lambda_1(\tau, t) = \bar\lambda_1 + \int_t^T [\bar\lambda_2 + (T-s)\bar\lambda_3 + k(2\tau - s)\bar\lambda_3]\, ds, \\ &\frac{d}{d\tau}\lambda_1(\tau, \tau) = -\bar\lambda_2 - (T-\tau)\bar\lambda_3 - k\tau\bar\lambda_3 + 2k(T-\tau)\bar\lambda_3 \\ &\qquad\qquad \leqq -\bar\lambda_2 + (1+2k)|\bar\lambda_3| + k\eta \leqq -(k+3)\eta < 0. \end{aligned} \tag{6.7}$$

This proves (6.6) when $h \equiv 0$. To cover the general case, notice that (6.7) holds uniformly as $(\tau, T, \bar\lambda)$ range in the compact set $\{\tau, T \in \mathbb{R};\ 0 \leqq \tau \leqq T \leqq 1\} \times \Gamma$. Moreover, by the implicit function theorem, the total derivative of $\lambda_1(\tau, \tau)$ w.r.t. τ depends continuously on τ, T, $\bar\lambda$ and on the partial derivatives of order $\leqq 2$ of the vector field h. Therefore, if the neighborhood $\mathscr{V}_1 \in \mathscr{F}$ is suitably small, (6.6) still holds for any $h \in \mathscr{V}_1$. This completes the uniqueness proof.

7. Proof of Proposition 2. Again it is not restrictive to assume $|\bar\lambda| = 1$. In this case, $\bar\lambda_3^2 \geqq (12k+16)^{-2}(\bar\lambda_2^2 + \bar\lambda_3^2)$ implies $|\bar\lambda_3| \geqq (24k+32)^{-1}$. Let M be the constant in (6.1), (6.2) and choose some $\sigma > 0$ for which

$$(24k+32)(9+9M+10k)M\sigma \leqq \varepsilon. \tag{7.1}$$

Choose $\mathscr{V}_2 \in \mathscr{F}$ contained in $\mathscr{V}_0$ such that

$$|h_{i,j}(x)| \leqq \sigma, \quad |h_{i,jl}(x)| \leqq \sigma, \quad |h_1(x)| \leqq \sigma \tag{7.2}$$

for all $h \in \mathcal{V}_2$, $x \in \Omega_k$, $i, j, l \in \{1, 2, 3\}$. Since the right-hand side of $(4.3)_2$ is absolutely continuous, we can differentiate $(4.3)_2$ once more:

$$\ddot{\lambda}_1 = -\dot{\lambda}_2 - k\dot{x}_1\lambda_3 - kx_1\dot{\lambda}_3 - \sum_{i=1}^{3}\sum_{j=1}^{3} h_{i,1j}(x)\dot{x}_j\lambda_i - \sum_{i=1}^{3} h_{i,1}(x)\dot{\lambda}_i. \tag{7.3}$$

Using the bounds (6.1), (6.2), (7.1), (7.2) and the relations

$$\begin{aligned} &-\dot{\lambda}_2 - k\dot{x}_1\lambda_3 = \lambda_3 + \sum_{i=1}^{3} h_{i,2}(x)\lambda_i - ku\lambda_3 - kh_1(x)\lambda_3, \\ &|\dot{\lambda}_3| = \left|\sum_{i=1}^{3} h_{i,3}\lambda_i\right| \leqq 3\sigma M, \qquad |\lambda_3(t) - \bar{\lambda}| \leqq 3\sigma M, \qquad |x_1| \leqq 2, \end{aligned} \tag{7.4}$$

we obtain

$$|\ddot{\lambda}_1(t) - (1 - ku(t))\bar{\lambda}_3| \leqq (9 + 10k + 9M)M\sigma \leqq \varepsilon(24k + 32)^{-1} \leqq \varepsilon|\bar{\lambda}_3|.$$

8. Proof of Proposition 3. Set $\varepsilon = (k-1)/2$ and choose $\mathcal{V}' \in \mathcal{F}$ according to Proposition 2. Choose $\mathcal{V}'' \in \mathcal{F}$ so small that, whenever $h \in \mathcal{V}''$ and $g \in \bar{f} + h$, the following conditions hold at every point $x \in \Omega_k$.

i) The vectors e_1, $[e_1, g](x)$ and $[[e_1, g], g](x)$ are linearly independent.

ii) In (3.8), $k_3(x) > 1$.

Such a $\mathcal{V}''$ exists. Indeed, when $h \equiv 0$ we have $g \equiv \bar{f}$ and $[e_1, \bar{f}](x) = (0, 1, kx_1)$, $[[e_1, \bar{f}], \bar{f}](x) = (0, 0, 1)$, $[e_1, [e_1, \bar{f}]](x) = (0, 0, k)$. In this case the coefficients of the linear combination (3.8) are $k_1(x) = k_2(x) = 0$, $k_3(x) = k > 1$. By continuity, the conditions i) and ii) remain valid when h ranges within a suitably small neighborhood of the null vector field in $\mathcal{C}^3(\Omega_k)$. Now set $\mathcal{V}_3 = \mathcal{V}' \cap \mathcal{V}''$ and let (u, x, λ) be a solution of (5.2) with $\bar{\lambda}_3^2 \geqq (12k + 16)^{-2}(\bar{\lambda}_1^2 + \bar{\lambda}_2^2)$, $\bar{\lambda}_3 < 0$. We claim that $S = \{t \in [0, T];\ \lambda_1(t) = 0\}$ is a closed interval, possibly empty. If $t_1, t_2 \in S$, let $|\lambda_1(\tau)| = \max\{|\lambda_1(t)|;\ t_1 \leqq t \leqq t_2\}$. If $\lambda_1(\tau) \neq 0$, then $u(t) = \operatorname{sgn} \lambda_1(t)$ is constant on a neighborhood of τ, hence λ_1 is twice differentiable at τ. Since $\bar{\lambda}_3 < 0$, (5.3) and the choice of ε imply that $\operatorname{sgn} \ddot{\lambda}_1(\tau) = \operatorname{sgn} \lambda_1(\tau)$, a contradiction that proves our claim. If S is empty, Proposition 3 trivially holds by setting $\tau_1 = \tau_2 = 0$. If S contains a single point τ, set $\tau_1 = \tau_2 = \tau$. Finally, let S be a nondegenerate interval, say $[\tau_1, \tau_2]$. We need to show that $u(t) = k_3^{-1}(x(t))$ a.e. on S. The relations $\lambda_1(t) = \dot{\lambda}_1(t) = \ddot{\lambda}_1(t) = 0$ imply

$$\begin{aligned} &\langle \lambda(t), e_1 \rangle = 0, \\ &\langle -\dot{\lambda}(t), e_1 \rangle = \langle \lambda(t), \nabla g(x(t)) e_1 \rangle = \langle \lambda(t), [e_1, g](x(t)) \rangle = 0, \\ &\langle \ddot{\lambda}_1(t), e_1 \rangle = -\frac{d}{dt}\langle \lambda(t), [e_1, g](x(t)) \rangle \\ &\qquad = \langle \lambda(t), \nabla g(x(t)) \cdot [e_1, g](x(t)) - \nabla [e_1, g](x(t)) \cdot (g(x(t)) + u(t)e_1) \rangle \\ &\qquad = \langle \lambda(t), [[e_1, g], g](x(t)) - u(t)[e_1[e_1, g]](x(t)) \rangle = 0. \end{aligned}$$

Since $\lambda(t)$ never vanishes, for $t \in (\tau_1, \tau_2)$ the vectors e_1, $[e_1, g](x(t))$ and $[[e_1, g], g](x(t)) - u(t)[e_1, [e_1, g]](x(t))$, being orthogonal to $\lambda(t)$, are linearly dependent. Because of the assumption i), $u(t)$ is uniquely determined and thus coincides with $k_3^{-1}(x(t))$, defined by (3.8).

9. Proof of Proposition 4. Some preliminary technical results are needed.

LEMMA 4. *Let $\tau > 0$ and let ϕ be a twice differentiable concave scalar function, with $\phi(0) = \phi(\tau) = 0$, $\dot{\phi}(0) > 0$, and let σ, m_1, m_2 be positive constants such that*

$$-m_2 \leqq \ddot{\phi}(t) \leqq -m_1 < 0, \quad |\ddot{\phi}(t) - \ddot{\phi}(t')| \leqq \sigma|t - t'| \tag{9.1}$$

for all $t, t' \in [0, \tau]$. *Then*

$$|\dot\phi(\tau)| \geqq \dot\phi(0) - 4\sigma(m_1+2m_2)m_1^{-3}\dot\phi^2(0). \tag{9.2}$$

Proof. The first assumption in (9.1) implies $\tau \in [2\dot\phi(0)/m_2, 2\dot\phi(0)/m_1]$. Let $a = -\ddot\phi(0) > 0$ and define the energy $E(t) = \dot\phi^2(t)/2 + a\phi(t)$. Then

$$\left|\frac{dE(t)}{dt}\right| = |\dot\phi(t)(\ddot\phi(t)+a)| \leqq \dot\phi(0)\left(1+\frac{2m_2}{m_1}\right)\sigma t.$$

Integrating from 0 to τ we obtain

$$|E(\tau) - E(0)| \leqq 2\sigma(m_1+2m_2)m_1^{-3}\dot\phi^3(0). \tag{9.3}$$

This implies (9.2) because

$$\begin{aligned}|\dot\phi(\tau)| - |\dot\phi(0)| &= (\dot\phi^2(\tau) - \dot\phi^2(0))(|\dot\phi(\tau)| + |\dot\phi(0)|)^{-1} \\ &\leqq 2|E(\tau) - E(0)||\dot\phi(0)|^{-1}.\end{aligned}$$

LEMMA 5. *Let* $(d_n)_{n\geqq 1}$ *be a sequence of strictly positive numbers such that* $d_{n+1} \geqq d_n - Cd_n^2$ *for some constant* $C > 1$ *and all* $n \geqq 1$. *Then* $\sum_{n=1}^{\infty} d_n = +\infty$.

Proof. If the series converges, then $d_n \to 0$, hence $d_n \leqq 1/2C$ for all $n \geqq N$, with N suitably large. We claim that $d_{N+n} \geqq n^{-1}d_{N+1}$ for all $n \geqq 1$. Indeed, if this inequality holds for some n, then

$$\begin{aligned}d_{N+n+1} &\geqq \min\{x - Cx^2;\ d_{N+1}/n \leqq x \leqq 1/2C\} \\ &= \frac{1}{n}d_{N+1} - \frac{C}{n^2}d_{N+1}^2 \geqq \left(\frac{1}{n} - \frac{C}{n^2}\cdot\frac{1}{2C}\right)d_{N+1} \geqq d_{N+1}/(n+1).\end{aligned}$$

By induction, our claim holds for every $n \geqq 1$, showing that the series diverges, a contradiction.

LEMMA 6. *Let* $h \in \mathscr{V}_0$ *and let* $t \to (x(t), \lambda(t))$ *be any local solution of the autonomous differential equation on* $\mathbb{R}^6$:

$$\dot x(t) = g(x(t)) + e_1, \qquad \dot\lambda(t) = -\lambda(t)\cdot\nabla g(x(t))$$

obtained by setting $u(t) \equiv 1$ *in* $(5.2)_{1\text{-}2}$. *There exists a constant* σ' *such that*

$$\left|\frac{d^3}{dt^3}\lambda_1(t)\right| \leqq \sigma'|\lambda(t)|, \qquad \left|\frac{d}{dt}\lambda_3(t)\right| \leqq \sigma'|\lambda(t)| \tag{9.4}$$

whenever $x(t) \in \Omega_k$. *The smallest possible constant* σ' *in* (9.4) *approaches zero as the vector field* $h = g - \bar f$ *tends to zero in* $\mathscr{C}^3(\Omega_k)$. *The same holds for the system*

$$\dot x(t) = g(x(t)) - e_1, \qquad \dot\lambda(t) = -\lambda(t)\cdot\nabla g(x(t)).$$

All of the above is clear because the left-hand sides in (9.4) depend continuously on x, λ and on the vector field $h \in \mathscr{C}^3(\Omega_k)$, and vanish identically when $h \equiv 0$.

Proposition 4 can now be proved. Fix $\varepsilon = (k-1)/2$, choose $\mathscr{V}_2, \mathscr{V}_3 \in \mathscr{F}$ according to Propositions 2 and 3 and set $\mathscr{V}_4 = \mathscr{V}_2 \cap \mathscr{V}_3$. Let $h \in \mathscr{V}_4$ and let (u, x, λ) be a solution of (5.2) with $|\bar\lambda| = 1$, $\bar\lambda$ satisfying the assumptions made in Proposition 4. If $\lambda_1(t) = 0$ for all $t \in [0, T]$, then $(u, x, -\lambda)$ is another solution of (5.2); hence by Proposition 3 $u(t) = k_3^{-1}(x(t))$ for all t. Now assume $\lambda_1(\tau) \neq 0$ for some $\tau \in [0, T]$. Then $[\tau, T]$ contains only finitely many zeros of λ_1. To see this, set $m_1 = (k-1-\varepsilon)\bar\lambda_3$, $m_2 = (k+1+\varepsilon)\bar\lambda_3$. Whenever $\lambda_1(t) \neq 0$, u is constantly equal to $\operatorname{sgn}\lambda_1(t)$ on a neighborhood of t; hence

λ_1 is three times differentiable at t. By (5.3)

$$-m_2 \leqq \frac{d^2}{dt^2}|\lambda_1(t)| \leqq -m_1 < 0. \tag{9.5}$$

If λ_1 vanishes infinitely many times inside $[\tau, T]$, let τ_0 be the smallest time. Recursively, set $\tau_{n+1} = \inf\{t \in (\tau_n, T];\ \lambda_1(t) = 0\}$. By (9.5), $\dot\lambda_1(\tau_0) \neq 0$ and τ_0 is an isolated zero of λ_1. By induction, one easily checks that the same holds for every n; hence the sequence $(\tau_n)_{n \geqq 1}$ is strictly increasing. We now apply Lemma 4 to the function $\phi(t) = |\lambda_1(\tau_n + t)|$ for each interval $[\tau_n, \tau_{n+1}]$. The second estimate in (9.1) is obtained from (9.4) and (6.1), setting $\sigma = M\sigma'$. Using (9.2) we deduce

$$|\dot\lambda_1(\tau_{n+1})| \geqq |\dot\lambda_1(\tau_n)| - 4\sigma(m_1 + 2m_2)m_1^{-3}|\dot\lambda_1(\tau_n)|^2.$$

If infinitely many τ_n were defined, by Lemma 5 $\sum_{n=0}^{\infty} |\dot\lambda_1(\tau_n)| = +\infty$. From (9.5) it follows that $\tau_{n+1} - \tau_n \geqq 2|\dot\lambda_1(\tau_n)|m_2^{-1}$, hence $\lim_{n\to\infty} \tau_n = +\infty$, providing a contradiction. An analogous argument shows that λ_1 can have only finitely many zeros inside $[0, \tau]$. Hence the corresponding control u is bang-bang with finitely many switchings.

10. Proof of Proposition 5. We restrict the analysis to the case where $u(t) = +1$ on the initial interval $[0, t_1)$. When $u(t) = -1$ on $[0, t_1)$ an entirely analogous argument applies.

LEMMA 7. *For every h in a suitably small neighborhood $V \in \mathcal{F}$, there exists a unique one-parameter family of bang-bang controls $u(\xi) = u^+(a(\xi), b(\xi), c(\xi))$, $\xi \in [0, 1/2]$, having a first switch at time $t = \xi$ and a third switch at $t = 1$, which satisfy Pontryagin's equations* (5.2) *on the time interval $[\xi, 1]$ with $\lambda_1(\xi) = \lambda_1(1) = 0$.*

Proof. Whenever $h \in \mathcal{V}$ is small enough, the proofs of Propositions 1 to 3 show that the adjoint variable $\lambda(\cdot)$ in (5.2) corresponding to a bang-bang control with at least two switchings inside $[0, 1]$ must satisfy

$$\lambda_3(t) > 0, \quad |\ddot\lambda_1(t)/\lambda_3(1) - (1 - ku(t))| \leqq (k-1)/2 \tag{10.1}$$

a.e. on $[0, 1]$. To construct the one-parameter family $u(\xi)$, for a fixed $h \in \mathcal{C}^3(\Omega_k)$, $g = \bar f + h$ and $\xi \in [0, \frac{1}{2}]$, let $u = u^+(\xi, t_2 - \xi, 1 - t_2)$ be the control whose value is initially $+1$ and has switchings at times ξ, t_2, 1, as in (5.5). Consider the Cauchy problem on $\mathbb{R}^6$, starting at time $t = \xi$:

$$\begin{aligned} &\dot x(t) = g(x(t)) + e_1 u(t), \qquad \dot\lambda(t) = -\lambda(t)\nabla g(x(t)), \\ &x(\xi) = (\exp \xi(g + e_1))(0), \quad \lambda(\xi) = (0, \nu, 1) \end{aligned} \tag{10.2}$$

for some $\nu \in \mathbb{R}$. The above data determine uniquely a trajectory $t \to (x(t), \lambda(t))$. From (10.1) it is clear that the control $u = u^+(\xi, t_2 - \xi, 1 - t_2)$ satisfies the Maximum Principle (5.2) on a neighborhood of the interval $[\xi, 1]$ iff $\lambda_1(t_2) = \lambda_1(1) = 0$. We claim that for $\mathcal{V} \in \mathcal{F}$ suitably small, the conditions

$$\lambda_1(t_2) = \lambda_1(1) = 0, \qquad \xi < t_2 < 1 \tag{10.3}$$

implicitly define t_2, ν uniquely as functions of h, ξ, for all $h \in \mathcal{V}$, $\xi \in [0, 1/2]$. Indeed, when $h \equiv 0$, the equations (10.1), (10.3) can be solved explicitly, first for ν as a function of t_2 and ξ, then for t_2 in terms of ξ:

$$\lambda_1(t_2) = (t_2 - \xi)(-\nu - k\xi) + (t_2 - \xi)^2(1 - k)/2. \tag{10.4}$$

The right-hand side of (10.4) vanishes at the point $t_2 \in (\xi, 1)$ iff $\nu = (t_2 - \xi)(1 - k)/2 - k\xi$.

In this case

$$\lambda_1(1) = (1-t_2)(t_2-\xi)(1+k)/2 + (1-t_2)^2(1-k)/2. \tag{10.5}$$

Hence $\lambda_1(1) = 0$ iff

$$(t_2-\xi)/(1-t_2) = (k-1)/(k+1). \tag{10.6}$$

The exact value of t_2 as a function of ξ is immediately obtained from (10.6). From (10.6) it also follows that

$$(t_2-\xi) > (k-1)/4(k+1), \quad 1+\xi-2t_2 > 0, \quad 1-t_2 \geqq 1/4 \tag{10.7}$$

for all $\xi \in [0, 1/2]$. Differentiating (10.4) and (10.5) w.r.t ν and t_2 respectively and using (10.7) we obtain

$$\frac{\partial \lambda_1(t_2)}{\partial \nu} = \xi - t_2 < -\frac{k-1}{4(k+1)} < 0, \tag{10.8}$$

$$\frac{\partial \lambda_1(1)}{\partial t_2} = (1+\xi-2t_2)(k+1)/2 + (k-1)(1-t_2) > (k-1)/4 > 0. \tag{10.9}$$

By the implicit function theorem, there exists a neighborhood $\mathcal{V} \in \mathcal{F}$ such that (10.2), (10.3) determine (t_2, ν) uniquely as $\mathcal{C}^3$ functions of (h, ξ) in $\mathcal{V} \times [0, 1/2]$. This proves Lemma 7, by setting $a(\xi) = \xi$, $b(\xi) = t_2(\xi) - \xi$, $c(\xi) = 1 - t_2(\xi)$.

Next, it will be shown that Proposition 5 holds if the bang-bang control u belongs to the one-parameter family $u^+(a(\xi), b(\xi), c(\xi))$ just defined. To this purpose we need a technical result, whose proof is straightforward.

LEMMA 8. *Let $\mathcal{V} \in \mathcal{F}$ and let $(h, \xi) \to \phi(h, \xi)$ be a $\mathcal{C}^2$ map from $\mathcal{V} \times [0, 1/2]$ into $\mathbb{R}$ such that $\phi(h, 0) = 0$ for all $h \in \mathcal{V}$ and $\phi(0, \xi) > 0$ for all $\xi \in (0, 1/2]$. Assume that either* i) $(\partial\phi/\partial\xi)(0, 0) > 0$ *or* ii) $(\partial\phi/\partial\xi)(h, 0) = 0$ *for all $h \in \mathcal{V}$ and $(\partial^2\phi/\partial\xi^2)(0, 0) > 0$. Then $\phi(h, \xi) > 0$ for all $\xi \in (0, 1/2]$ and all h in some neighborhood of the null vector field in $\mathcal{C}^3(\Omega_k)$.*

For $h \in \mathcal{V}$ suitably small, we now construct a second one-parameter family of bang-bang controls $u'(\xi) = u^-(\alpha(\xi), \beta(\xi), \gamma(\xi))$, choosing α, β, γ such that $\alpha + \beta + \gamma = 1$ and the equalities in (5.4) hold, i.e.

$$\begin{aligned} &\pi_i(\exp \gamma(\xi)(g-e_1))(\exp \beta(\xi)(g+e_1))(\exp \alpha(\xi)(g-e_1))(0) \\ &\quad = \pi_i(\exp c(\xi)(g+e_1))(\exp b(\xi)(g-e_1))(\exp a(\xi)(g+e_1))(0) \end{aligned} \tag{10.10}$$

for $i = 1, 2$. When $h \equiv 0$, (10.6) implies

$$a(\xi) = \xi, \qquad b(\xi) = (k-1)(1-\xi)/2k, \qquad c(\xi) = (k+1)(1-\xi)/2k \tag{10.11}$$

and $\alpha(\xi)$, $\beta(\xi)$, $\gamma(\xi)$ are obtained substituting the values (10.11) in (5.8). By the implicit function theorem, the condition $\alpha(\xi) + \beta(\xi) + \gamma(\xi) = 1$ together with (10.10) defines a $\mathcal{C}^3$ map $(h, \xi) \to (\alpha, \beta, \gamma)$ on $\mathcal{V} \times [0, 1/2]$, for a suitably small neighborhood $\mathcal{V} \in \mathcal{F}$. Notice that when $h \equiv 0$ and ξ ranges inside $[0, 1/2]$, $\alpha(\xi)$ and $\beta(\xi)$ are strictly positive, while $\gamma(\xi) > 0$ for $\xi > 0$. Moreover, $(d\gamma/d\xi) = (k-1)/(k+1) > 0$ at $\xi = 0$. Setting $\phi = \gamma$ in Lemma 8, we deduce $\gamma(\xi) > 0$ for all $(h, \xi) \in \mathcal{V} \times [0, 1/2]$ with $\mathcal{V}$ small enough. Therefore the bang-bang control $u'(\xi) = u^-(\alpha(\xi), \beta(\xi), \gamma(\xi))$ is well defined. To prove the last inequality in (5.4), set $\phi(h, \xi) = x_3(u'(\xi), 1) - x_3(u(\xi), 1)$. For any fixed h, when $\xi = 0$ (10.10) has the obvious solution $\alpha(0) = b(0)$, $\beta(0) = c(0)$, $\gamma(0) = a(0) = 0$. Call $\tilde{u}$ the control $u^+(a(0), b(0), c(0))$, which coincides with $u^-(\alpha(0), \beta(0), \gamma(0))$ for all $t \in [0, 1]$, and let $t \to (\tilde{x}(t), \tilde{\lambda}(t))$ be the corresponding trajectory and adjoint variable in (10.2). Since $\tilde{\lambda}_1$ vanishes at times 0, $t_2 = b(0)$, 1, as $\xi \to 0$ we have

$$\langle\tilde{\lambda}(1), x(u^+(a(\xi), b(\xi), c(\xi)), 1) - \tilde{x}(1)\rangle\xi^{-1}$$
$$= \left[\int_0^1 \tilde{\lambda}_1(t)[u^+(a(\xi), b(\xi), c(\xi))(t) - u^+(a(0), b(0), c(0))(t)]\, dt + O(\xi^2)\right]\xi^{-1}$$
$$= o(\xi).$$

The same holds for $u^-(\alpha(\xi), \beta(\xi), \gamma(\xi))$, therefore

$$\lim_{\xi\to 0}\langle\tilde{\lambda}(1), x(u'(\xi), 1) - x(u(\xi), 1)\rangle\bar{\xi}^{-1} = \tilde{\lambda}_3(1)(\partial\phi/\partial\xi)(h, 0) = 0. \tag{10.12}$$

From (10.12) we deduce $(\partial\phi/\partial\xi)(h, 0) = 0$. When $h \equiv 0$, (5.10) and (10.11) imply

$$\phi(0, \xi) = \frac{(k-1)^2(k+1)(1-\xi)^2}{2k[2k\xi + (k+1)(1-\xi)]},$$

hence $(\partial^2\phi/\partial\xi^2)(0, 0) = (k-1)^2/k > 0$. By Lemma 8, $x_3(u'(\xi), 1) - x_3(u(\xi), 1) > 0$ for all $\xi \in (0, 1/2]$ and h in a neighborhood of the null vector field.

To conclude the proof of Proposition 5, notice that for every constant $\varepsilon' > 0$, in (5.3) we can choose $\varepsilon > 0$ so small that the conditions

$$|\ddot{\lambda}_1(t)/\bar{\lambda}_3 - (1-k)| \leqq \varepsilon \quad \text{for } t \in (0, t_1) \cup (t_2, 1),$$
$$|\ddot{\lambda}_1(t)/\bar{\lambda}_3 - (1+k)| \leqq \varepsilon \quad \text{for } t \in (t_1, t_2)$$

together with $\lambda_1(t_1) = \lambda_1(t_2) = \lambda_1(1) = 0$, $\lambda_1(t) > 0$ on $(0, t_1)$ imply

$$|(t_2 - t_1)/(1 - t_2) - (k-1)/(k+1)| \leqq \varepsilon', \qquad t_1 \leqq (1 - t_2) + \varepsilon'. \tag{10.13}$$

For $\varepsilon' > 0$ suitably small, (10.13) implies $t_1 \in [0, 1/2]$. Therefore, if $h \in \mathcal{V}$ is small enough, a bang-bang control u, which is initially positive and has switchings at times $0 < t_1 < t_2 < t_3 = 1$, can satisfy Pontryagin's equations (5.2) only if $t_1 \leqq 1/2$. But in this case u is the member of the one-parameter family of control functions $u^+(a(\xi), b(\xi), c(\xi))$ obtained by setting $\xi = t_1$. Hence Proposition 5 holds for u.

Appendix. The equalities (5.10) are obtained from (5.6) to (5.9), using the relations $ab = \beta\gamma$, $\alpha\beta = bc$, as follows.

$$\begin{aligned}
3(x_3^+ - x_3^-) &= (a+b+c)^3 - (b+c)^3 + c^3 - (\beta+\gamma)^3 + \gamma^3 \\
&\quad + k[a^3 + (b-a)^3 - \alpha^3 - (\beta-\alpha)^3 + (a-b+c)^3] \\
&= a^3 + 3a^2(b+c) + 3a(b+c)^2 + (b+c)^3 - (b+c)^3 + c^3 - \beta^3 - 3\beta^2\gamma \\
&\quad - 3\beta\gamma^2 - \gamma^3 + \gamma^3 + k[a^3 + (b-a)^3 - \alpha^3 - \beta^3 + 3\beta^2\alpha - 3\beta\alpha^2 \\
&\qquad\qquad + \alpha^3 + c^3 - 3c^2(b-a) + 3c(b-a)^2 - (b-a)^3] \\
&= a^3 + 3a^2b + 3a^2c + 3ab^2 + 6abc + 3ac^2 + c^3 - (a^3 + 3a^2c + 3ac^2 + c^3) \\
&\quad - 3a^2b^2/(a+c) - 3(a^2b + abc) \\
&\quad + k[a^3 - (a^3 + 3a^2c + 3ac^2 + c^3) + 3(abc + bc^2) - 3b^2c^2/(a+c) \\
&\qquad\qquad + c^3 - 3bc^2 + 3ac^2 + 3b^2c - 6abc + 3a^2c] \\
&= 3ab^2 + 3abc - 3a^2b^2/(a+c) + k[-3abc - 3b^2c^2/(a+c) + 3b^2c],
\end{aligned}$$

$$\begin{aligned}
x_3^+ - x_3^- &= (a^2b^2 + ab^2c + a^2bc + abc^2 - a^2b^2)/(a+c) \\
&\quad - k(a^2bc + abc^2 + b^2c^2 - ab^2c - b^2c^2)/(a+c) \\
&= \frac{abc}{a+c}[a+b+c-k(a-b+c)] = \frac{\alpha\beta\gamma}{\alpha+\gamma}[\alpha+\beta+\gamma+k(\alpha-\beta+\gamma)].
\end{aligned}$$

REFERENCES

[1] A. BRESSAN, *Local asymptotic approximation of nonlinear control systems*, Int. J. Control, to appear.
[2] ———, *Directional convexity and finite optimality conditions*, J. Math. Anal. Appl., submitted.
[3] R. BROCKETT, *Volterra series and geometric control theory*, Automatica, 12 (1976), pp. 162–176.
[4] J. DIÉUDONNE, *Foundations of Modern Analysis*, Academic Press, New York, 1969.
[5] A. F. FILIPPOV, *Classical solutions of differential equations with multivalued right-hand side*, this Journal, 5 (1967), pp. 609–621.
[6] M. FLIESS, *Fonctionnelles causales nonlinéaires et indéterminées noncommutatives*, Bull. Soc. Math. France, 109 (1981), pp. 3–40.
[7] H. HERMES, *On the synthesis of a stabilizing feedback control via Lie algebraic methods*, this Journal, 18 (1980), pp. 352–361.
[8] ———, *Control systems which generate decomposable Lie algebras*, J. Differential Equations, 44 (1982), pp. 166–187.
[9] H. HERMES AND J. P. LASALLE, *Functional Analysis and Time Optimal Control*, Academic Press, New York, 1969.
[10] H. SUSSMANN, *A bang–bang theorem with bounds on the number of switchings*, this Journal, 17 (1979), pp. 629–651.
[11] ———, *Lie brackets and local controllability: a sufficient condition for scalar-input systems*, this Journal, 21 (1983), pp. 686–713.
[12] ———, *Trajectory analysis and regular synthesis for analytic optimal control problems in the plane*, this Journal, submitted.
[13] ———, *Analytic stratifications and control theory*. Proc. International Congress of Mathematicians, Helsinki, 1978, pp. 865–871.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 2, March 1986

003

AN OPTIMIZATION PROBLEM WITH VOLUME CONSTRAINT*

N. AGUILERA†, H. W. ALT‡ AND L. A. CAFFARELLI§

Abstract. We consider the optimization problem of minimizing Dirichlet integral $\int_\Omega |\nabla u|^2$ with volume constraint on the set $\{u>0\}$. We use a penalization method which for small values of the penalization parameter leads to a solution of the original problem.

Key words. partial differential equations, free boundary, optimization, optimum design

AMS(MOS) subject classifications. 35A15, 35J65, 49A22

1. Introduction. A classical variational problem asks for the properties of the following optimal configuration:

Given a perfect conductor $\Gamma_0=\partial\Omega$, Ω a bounded or unbounded domain, find a second one $\Gamma_1=\Omega\cap\partial D$, D a subdomain of Ω containing Γ_0 as boundary, such that their mutual energy is minimized among all bodies D with a prescribed volume ω_0.

Mathematically the problem under consideration is the following:

Let $\Omega\subset\mathbb{R}^n$ be a domain with bounded Lipschitz boundary $\partial\Omega$, Ω being bounded or unbounded, and $u_0\in H^{1,2}(\Omega)$ nonnegative. We assume that u_0 is strictly positive in a neighborhood of at least one C^2-regular point of $\partial\Omega$. Furthermore let $0<\omega_0<\mathcal{L}^n(\Omega)$, where $\mathcal{L}^n$ denotes the n-dimensional Lebesgue measure. We look for a function $u\in H^{1,2}(\Omega)$ with

$$u=u_0 \quad \text{on } \partial\Omega \quad \text{and} \quad \mathcal{L}^n(\{u>0\})=\omega_0$$

minimizing Dirichlet integral among all functions with these properties, that is,

$$\mathcal{J}(u)\leqq\mathcal{J}(v) \quad \text{for all } v\in K_0,$$

where

$$\mathcal{J}(v):=\int_\Omega |\nabla v|^2,$$

$$K_0:=\{v\in L^1_{\text{loc}}(\Omega);\ \nabla v\in L^2(\Omega),\ v=u_0 \text{ on } \partial\Omega,\ v\geqq 0, \text{ and } \mathcal{L}^n(\{v>0\})=\omega_0\}.$$

Although the existence of a weak solution is not hard to prove, the regularity properties of such a solution and its free boundary $\partial\{u>0\}\cap\Omega$ are not easy to establish, mainly because it is hard, at this stage, to make volume preserving perturbations.

Formally, by Hadamard's variational formula (see [3]), the solution u satisfies the free boundary condition

$$-\partial_\nu u=\lambda \quad \text{on } \partial\{u>0\}\cap\Omega$$

for some $\lambda>0$. Solution to such a free boundary problem can be obtained by minimizing the energy functional

$$\tilde{\mathcal{J}}_\lambda(v):=\int_\Omega (|\nabla v|^2+\lambda^2\chi(\{v>0\}))$$

on the convex set

$$K:=\{v\in L^1_{\text{loc}}(\Omega);\ \nabla v\in L^2(\Omega),\ v=u_0 \text{ on } \partial\Omega, \text{ and } v\geqq 0\}.$$

* Received by the editors June 13, 1984, and in revised form October 30, 1984.

† School of Mathematics, University of Minnesota, Minneapolis, Minnesota 55455.

‡ Institut für Angewandte Mathematik, Universität Bonn, Wegelerstrasse 6, 5300 Bonn 1, West Germany.

§ Department of Mathematics, University of Chicago, Chicago, Illinois 60637.

A complete mathematical description has been given by two of the authors ([A−C]), who proved that a minimizer $\tilde{u}_\lambda$ of $\tilde{\mathcal{J}}_\lambda$ is Lipschitz continuous. Furthermore, the free boundary $\partial\{\tilde{u}_\lambda > 0\} \cap \Omega$ is, if $n \geqq 3$ except for a closed set of $n-1$ Hausdorff dimension zero, an analytic hypersurface, on which $-\partial_\nu \tilde{u}_\lambda = \lambda$ is satisfied.

The relation between solutions of the original functional $\mathcal{J}$ and solutions of $\tilde{\mathcal{J}}_\lambda$ is clear. If, for a choice of the parameter λ, we have $\mathcal{L}^n(\{\tilde{u}_\lambda > 0\}) = \omega_0$, then $\tilde{u}_\lambda$ becomes a minimizer of $\mathcal{J}$. Unfortunately, given a value of ω_0, it is not clear that a corresponding value of λ exists, even if Ω is the unit ball (see [A-C, 2.6]). However for the exterior problem, for instance if $\mathbb{R}^n \backslash \Omega$ is starlike, the existence of λ is guaranteed (see [A]).

To solve the original problem in a way that will allow us to perform nonvolume preserving variations, we will use a penalization technique. Therefore for $\varepsilon > 0$ we consider the functional

$$\mathcal{J}_\varepsilon(v) := \int_\Omega |\nabla v|^2 + f_\varepsilon(\mathcal{L}^n(\{v > 0\})) \quad \text{for } v \in K$$

with

$$f_\varepsilon(s) := \begin{cases} \dfrac{1}{\varepsilon}(s - \omega_0) & \text{for } s \geqq \omega_0, \\ \varepsilon(s - \omega_0) & \text{for } s \leqq \omega_0. \end{cases}$$

We will show that minimizers u_ε of $\mathcal{J}_\varepsilon$ are weak solutions of the free boundary problem $-\partial_\nu u_\varepsilon = \lambda_\varepsilon$ for some λ_ε in the sense of [A-C] and therefore the theory established there can be applied.

Finally, we like to fetch from u_ε a solution u to our original variational problem for $\mathcal{J}$. Perhaps the most interesting feature of this note is the fact that it is not necessary to pass to the limit in ε to obtain u, since for ε small enough the volume of $\{u_\varepsilon > 0\}$ automatically adjusts to ω_0.

The reason is that by a simple comparison argument λ_ε stays away from zero and infinity independently of ε, that is,

$$0 < c \leqq \lambda_\varepsilon \leqq C < \infty \quad \text{for all small } \varepsilon.$$

Then using the fact that $f'_\varepsilon(s)$ jumps from a small to a large number at $s = \omega_0$ we conclude that $\mathcal{L}^n(\{u_\varepsilon > 0\}) = \omega_0$. In fact, if we assume that $\mathcal{L}^n(\{u_\varepsilon > 0\}) > \omega_0$, then an inward perturbation of $\{u_\varepsilon > 0\}$ with volume change δV will induce a change $\delta f_\varepsilon = -(1/\varepsilon)\delta V$, while the Dirichlet integral will change by $\lambda_\varepsilon^2 \delta V \leqq C^2 \delta V$, for $(1/\varepsilon) > C^2$ a contradiction to u_ε being a minimizer. If $\mathcal{L}^n(\{u_\varepsilon > 0\}) < \omega_0$ we argue similarly.

2. Solution for the functional $\mathcal{J}_\varepsilon$.

THEOREM. *There exists a minimizer $u_\varepsilon \in \mathcal{K}$ of $\mathcal{J}_\varepsilon$ with the following properties:*

1) *Lipschitz continuity*: $u_\varepsilon \in C^{0,1}_{\text{loc}}(\Omega)$.

2) *Nondegeneracy*: $c\ \text{dist}\,(x, \partial\{u_\varepsilon > 0\}) \leqq u_\varepsilon(x) \leqq C\ \text{dist}\,(x, \partial\{u_\varepsilon > 0\})$ *for* $x \in D$.

3) *Positive density*: $c \leqq \mathcal{L}^n(B_r(x) \cap \{u_\varepsilon > 0\}) / \mathcal{L}^n(B_r(x)) \leqq 1 - c$ *for* $B_r(x) \subset D$.

Here D is any relative compact subdomain of Ω, and the constants c, C depend on D and ε.

Proof. The proofs follow exactly those in [A-C, §§ 1-3]. Here are some details: If we choose u_0 with $\mathcal{L}^n(\{u_0 > 0\}) \leqq \omega_0$ we get $\mathcal{J}_\varepsilon(u_0) \leqq C$ (uniformly in ε). Also $\mathcal{J}(v) \geqq -\omega_0$ for all $v \in \mathcal{K}$. Therefore a minimizing sequence u_k exists, and

$$\int_\Omega |\nabla u_k|^2 \leqq C, \qquad \mathcal{L}^n(\{u_k > 0\}) \leqq C.$$

Hence for some $u_\varepsilon \in \mathcal{K}$ and a subsequence

$$\nabla u_k \to \nabla u_\varepsilon \quad \text{weakly in } L^2(\Omega),$$

$$u_k \to u_\varepsilon \quad \text{almost everywhere in } \Omega.$$

Consequently

$$\mathcal{L}^n(\{u_\varepsilon > 0\}) \leqq \liminf_{k\to\infty} \mathcal{L}^n(\{u_k > 0\}),$$

$$\int_\Omega |\nabla u_\varepsilon|^2 \leqq \liminf_{k\to\infty} \int_\Omega |\nabla u_k|^2,$$

so that u_ε is a minimizer.

Using the fact that

$$\varepsilon(s_2 - s_1) \leqq f_\varepsilon(s_2 - s_1) \leqq \frac{1}{\varepsilon}(s_2 - s_1) \quad \text{for } s_1 \leqq s_2,$$

we obtain the properties 1)-3) as in [A-C, § 3].

This puts our solution u_ε under the hypothesis of [A-C, § 4] and therefore representation Theorem 4.5 and Theorem 4.8 are valid.

THEOREM 2. 1) $\mathcal{H}^{n-1}(D \cap \partial(u_\varepsilon > 0)) < \infty$ *for* $D \Subset \Omega$.

2) *There is a Borel function* q_{u_ε} *such that in the sense of distributions*

$$\Delta u_\varepsilon = q_{u_\varepsilon} \mathcal{H}^{n-1} \llcorner \partial\{u_\varepsilon > 0\},$$

that is, for $\zeta \in C_0^\infty(\Omega)$ *we have*

$$-\int_\Omega \nabla u_\varepsilon \nabla \zeta = \int_{\Omega \cap \partial\{u_\varepsilon > 0\}} \zeta q_{u_\varepsilon} \, d\mathcal{H}^{n-1}.$$

3) *For* $D \Subset \Omega$ *there are constants* $0 < c \leqq C < \infty$ *depending on* u_ε, Ω, D *such that for balls* $B_r(x_0) \subset D$ *with center* $x_0 \in \partial\{u_\varepsilon > 0\}$

$$c \leqq q_{u_\varepsilon}(x_0) \leqq C,$$

$$cr^{n-1} \leqq \mathcal{H}^{n-1}(B_r(x_0) \cap \partial\{u_\varepsilon > 0\}) \leqq Cr^{n-1}.$$

4) *For* $\mathcal{H}^{n-1}$ *almost all* $x_0 \in \partial_{\text{red}}\{u_\varepsilon > 0\}$

$$\text{Tan}\,(\partial\{u_\varepsilon > 0\}, x_0) = \{x;\, x \cdot \nu_{u_\varepsilon}(x_0) = 0\},$$

which determines the outward normal ν_{u_ε} *of* $\{u_\varepsilon > 0\}$ *at* x_0, *and*

$$u_\varepsilon(x_0 + x) = q_{u_\varepsilon}(x_0) \max(-x \cdot \nu_{u_\varepsilon}(x_0), 0) + o(x) \quad \text{as } x \to 0.$$

5) $\mathcal{H}^{n-1}(\partial\{u_\varepsilon > 0\} \setminus \partial_{\text{red}}\{u_\varepsilon > 0\}) = 0$.

The last statement is a consequence of the positive density property (see proof of [A-C, 5.5]).

With this theorem at hand it is not difficult to identify q_{u_ε}.

THEOREM 3. *For some positive constant* $\lambda_\varepsilon > 0$

$$q_{u_\varepsilon} = \lambda_\varepsilon \qquad \mathcal{H}^{n-1} \text{ almost everywhere on } \partial_{\text{red}}\{u_\varepsilon > 0\}.$$

Proof. We write u instead of u_ε. Let x_0, x_1 be two points in $\partial_{\text{red}}\{u > 0\}$ at which u behaves as in Theorem 2.4.

Assume that $q_u(x_0) > q_u(x_1)$. We construct a perturbation outwards to $\{u > 0\}$ at x_0, inwards at x_1 that preserves volume up to higher order terms. That is, we replace

u for small ρ by the function

$$v_\rho(\tau_\rho(x)) := u(x),$$

where

$$\tau_\rho(x) := \begin{cases} x + \kappa\rho\phi\left(\dfrac{|x-x_0|}{\rho}\right)\nu_u(x_0) & \text{for } x \in B_\rho(x_0), \\ x - \kappa\rho\phi\left(\dfrac{|x-x_1|}{\rho}\right)\nu_u(x_1) & \text{for } x \in B_\rho(x_1), \\ x & \text{elsewhere.} \end{cases}$$

κ is a small positive constant and ϕ is a nonnegative C_0^∞ function, $\phi \not\equiv 0$, supported in the unit interval. If κ is small enough (independent of ρ) then τ_ρ is a diffeomorphism with

$$D\tau_\rho(x) = I + \kappa D\eta_i\left(\frac{x-x_i}{\rho}\right) \quad \text{in } B_\rho(x_i),$$

$$\eta_i(y) := (-1)^i\phi(|y|)\nu_u(x_i).$$

Introducing the normalized functions

$$u_{i\rho}(y) := \frac{1}{\rho}\,u(x_i + \rho y),$$

the first part of Theorem 2.4 says that the set $B_1(0) \cap \{u_{i\rho} > 0\}$ approaches

$$\{y \in B_1(0);\ y \cdot \nu_i < 0\}, \qquad \nu_i := \nu_u(x_i),$$

as $\rho \to 0$. Therefore

$$\rho^{-n}\mathcal{L}^n(B_\rho(x_i) \cap \{v_\rho > 0\}) = \rho^{-n}\int_{B_1(0)\cap\{u_{i\rho}>0\}} \det D\tau_\rho(x_i + \rho y)\,dy$$

$$\to \int_{\{y\cdot\nu_i<0\}} \left(1 + (-1)^i\kappa\phi'(|y|)\frac{y}{|y|}\cdot\nu_i\right)dy.$$

Since the last integral is independent of the direction of ν_i, we conclude

$$\rho^{-n}(\mathcal{L}^n(\{v_\rho > 0\}) - \mathcal{L}^n(\{u > 0\})) \to 0;$$

therefore

$$f_\varepsilon(\mathcal{L}^n(\{v_\rho > 0\})) - f_\varepsilon(\mathcal{L}^n(\{u > 0\})) \leqq \frac{1}{\varepsilon}\,o(\rho^n).$$

We now compute the change in the Dirichlet integral. Again we normalize to the unit ball and obtain

$$\rho^{-n}\int_{B_\rho(x_i)} (|\nabla v_\rho|^2 - |\nabla u|^2) = \int_{B_1(0)\cap\{u_{i\rho}>0\}} (|\nabla u_{i\rho}(I + \kappa D\eta_i)^{-1}|^2 \det(I + \kappa D\eta_i) - |\nabla u_{i\rho}|^2)$$

$$= \kappa\int_{B_1(0)\cap\{u_{i\rho}>0\}} (|\nabla u_{i\rho}|^2\nabla\cdot\eta_i - 2\nabla u_{i\rho} D\eta_i \nabla u_{i\rho}) + o(\kappa).$$

From Theorem 2.4 we know that

$$u_{i\rho}(y) \to q_u(x_i)\max(-y\cdot\nu_i, 0)$$

uniformly in any bounded domain. Also for any $\delta > 0$

$$u_{i\rho}(y) > 0 \quad \text{for } y \in B_2(0), \quad y \cdot \nu_i < -\delta/2,$$

$$u_{i\rho}(y) = 0 \quad \text{for } y \in B_2(0), \quad y \cdot \nu_i > \delta/2,$$

provided ρ is small enough. Hence

$$\nabla u_{i\rho} \to -q_u(x_i)\nu_i$$

uniformly in $B_1(0) \cap \{y \cdot \nu_i < -\delta\}$. Since $\nabla u_{i\rho}$ are uniformly bounded, we conclude

$$\nabla u_{i\rho} \to -q_u(x_i)\nu_i \chi(B_1(0) \cap \{y \cdot \nu_i < 0\})$$

in $L^p(B_1(0))$ for any $p < \infty$. Therefore the above expression converges to

$$\kappa \int_{B_1(0)\cap\{y\cdot\nu_i<0\}} q_u(x_i)^2(\nabla \cdot \eta_i - 2\nu_i D\eta_i \nu_i) + o(\kappa)$$

$$= -\kappa q_u(x_i)^2 \int_{B_1(0)\cap\{y\cdot\nu_i<0\}} \partial_{\nu i}\eta_i \cdot \nu_i + o(\kappa)$$

$$= -\kappa(-1)^i q_u(x_i)^2 \int_{B_1(0)\cap\{y\cdot\nu_i=0\}} \phi(|y|)\, d\mathcal{H}^{n-1}(y) + o(\kappa),$$

or

$$\int_\Omega |\nabla v_\rho|^2 - \int_\Omega |\nabla u|^2 = \rho^n \left(\kappa(q_u(x_1)^2 - q_u(x_0)^2) \int_{B_1(0)} \phi(|y|)\, d\mathcal{L}^{n-1}(y) + o(\kappa) \right) + o(\rho^n).$$

If $q_u(x_1) < q_u(x_0)$ then the main term becomes negative if we choose κ small enough. For small ρ this contradicts the minimum property of u. This proves the theorem.

With the last theorem we complete the proof of the fact that u_ε satisfies the properties of a weak solution in [A-C, Def. 5.1]. Therefore [A-C, Thms. 8.3, 8.4] apply to it: The free boundary of u_ε is a locally smooth surface, except possibly for a closed set of $n-1$ Hausdorff measure zero. Along the regular part of it

$$-\partial_\nu u_\varepsilon = \lambda_\varepsilon.$$

3. Behavior of the solution for small ε. To complete our study we will now show that for ε small enough

$$\mathcal{L}^n(\{u_\varepsilon > 0\}) = \omega_0.$$

Remark. $0 < c \leqq \mathcal{L}^n(\{u_\varepsilon > 0\}) \leqq \omega_0 + C\varepsilon$.

Proof. In the proof of Theorem 1 we saw that

$$\mathcal{J}_\varepsilon(u_\varepsilon) \leqq \mathcal{J}_\varepsilon(u_0) \leqq C,$$

where C is independent of ε. Hence

$$f_\varepsilon(\mathcal{L}^n(\{u_\varepsilon > 0\})) \leqq C,$$

which proves the estimate from above. Also

$$\int_\Omega |\nabla u_\varepsilon|^2 \leqq C.$$

Since u_ε takes values u_0 on $\partial\Omega$ we obtain integrating along lines with $D := \Omega \cap B_\delta(\partial\Omega)$

$$\left(\int_{\partial\Omega} u_0\right)^2 \leqq C(\delta)\cdot \mathcal{L}^n(D\cap\{u_\varepsilon>0\})\int_D (|\nabla u_\varepsilon|^2+u_\varepsilon^2),$$

the last integral being bounded uniformly in ε; hence the estimate from below is proved.

As a consequence we obtain

LEMMA 5. *$\lambda_\varepsilon \leqq C$, where C is independent of ε.*

Proof. First we consider the case that Ω is bounded, and let w be the harmonic function in Ω with boundary values u_0. Then $w>0$ in Ω. Next choose $\delta>0$ so that the measure ω_1 of the set

$$D := \Omega\setminus\bar{B}_\delta(\partial\Omega)$$

exceeds ω_0, and $\mathcal{L}^n(\Omega\setminus D)$ is smaller than the constant c in Remark 4. Then for small ε

$$\mathcal{L}^n(D\cap\{u_\varepsilon>0\}) \leqq \omega_0 + C\varepsilon \leqq \frac{\omega_0+\omega_1}{2} < \omega_1 = \mathcal{L}^n(D)$$

and

$$\mathcal{L}^n(D\cap\{u_\varepsilon>0\}) \geqq \mathcal{L}^n(\{u_\varepsilon>0\}) - \mathcal{L}^n(\Omega\setminus D) \geqq c - \mathcal{L}^n(\Omega\setminus D) > 0.$$

Therefore by the isoperimetric inequality

$$\mathcal{H}^{n-1}(D\cap\partial\{u_\varepsilon>0\}) \geqq c(\mathcal{L}^n(D\cap\{u_\varepsilon>0\}))^{(n-1)/n} \geqq c>0.$$

Since u_ε have bounded Dirichlet integrals, we conclude using Theorem 2.2

$$\begin{aligned} C &\geqq \int_\Omega \nabla(u_\varepsilon - w)\nabla u_\varepsilon = \int_{\Omega\cap\partial\{u_\varepsilon>0\}} w\lambda_\varepsilon \, d\mathcal{H}^{n-1} \\ &\geqq \lambda_\varepsilon \inf_D w\cdot \mathcal{H}^{n-1}(D\cap\partial\{u_\varepsilon>0\}) \geqq c\lambda_\varepsilon. \end{aligned}$$

In the unbounded case we choose a fixed ball $B_R(0)$ containing $\partial\Omega$ with $\mathcal{L}^n(\Omega\cap B_R(0)) > \omega_0$. We argue with $\Omega\cap B_R(0)$ instead of Ω, let $w=0$ on $\partial B_R(0)$, and compute

$$\int_\Omega \nabla \min(u_\varepsilon - w, 0)\nabla u_\varepsilon.$$

Furthermore we use the fact that in Remark 4 the measure of $\{u_\varepsilon>0\}$ was estimated from below near $\partial\Omega$.

We remark that if u_0 is strictly positive on $\partial\Omega$ and $\partial\Omega$ is smooth, the proof can be performed as follows:

$$\begin{aligned} C &\geqq \int_\Omega |\nabla u_\varepsilon|^2 = \int_{\partial\Omega} u_0\,\partial_\nu u_\varepsilon \\ &\geqq \inf_{\partial\Omega} u_0\cdot \int_{\partial\Omega} \partial_\nu u_\varepsilon = \inf_{\partial\Omega} u_0 \int_{\Omega\cap\partial\{u_\varepsilon>0\}} \partial_{-\nu} u_\varepsilon \, d\mathcal{H}^{n-1} = \lambda_\varepsilon \inf_{\partial\Omega} u_0\, \mathcal{H}^{n-1}(\Omega\cap\partial\{u_\varepsilon>0\}). \end{aligned}$$

LEMMA 6. *$\lambda_\varepsilon \geqq c>0$, where c is independent of ε.*

Proof. Let $B_{\rho_0}(y_0)$, $y_0\in\partial\Omega$ be a ball such that $B_{\rho_0}(y_0)\cap\partial\Omega$ is C^2 and $u_0 \geqq c_0>0$ in $B_{\rho_0}(y_0)$. First let us assume that $u_\varepsilon>0$ in a neighborhood of y_0. Let y_1 be any free boundary point of u_ε in $\Omega\setminus B_\delta(\partial\Omega)$, δ independent of ε (see proof of Lemma 5; in the exterior case y_1 also should be bounded uniformly in ε). Choose a smooth family of

smooth domains $D_t \subset \Omega \cup B_{\rho_0}(y_0)$ such that $D_0 \subset B_{\rho_0}(y_0)\backslash\Omega$ touching $\partial\Omega$ at y_0 and $y_1 \subset D_1$, the smoothness not depending on ε. Let t be the first value for which D_t touches a free boundary point $x_0 \in \partial D_t \cap \partial\{u_\varepsilon > 0\}$.

Consider the harmonic function w in $D_t\backslash\bar{D}_0$ with boundary values c_0 on ∂D_0 and 0 on ∂D_t. Then

$$\partial_{-\nu} w(x_0) \geqq c c_0,$$

where c depends only on n and the geometry of D_0 and D_t, therefore c is independent of ε. Since $u \geqq w$ on $\partial(D_t \cap \Omega)$ we have $u \geqq w$ in $D_t \cap \Omega$. Hence if x_0 is a regular free boundary point

$$\lambda_\varepsilon = \partial_{-\nu} u_\varepsilon(x_0) \geqq \partial_{-\nu} w(x_0) \geqq c > 0,$$

where c is independent of ε, proving the lemma. However since we do not know it, we have to work more.

For small $r > 0$ we have writing u instead of u_ε

$$⨍_{\partial B_r(x_0)} u \geqq ⨍_{\partial B_r(x_0)} w \geqq cr.$$

Here ⨍ denotes the mean value. Let v_0 be the harmonic extension of u on $\partial B_r(x_0)$ into $B_r(x_0)$. Then (see [A-C, 3.2])

$$\int_{B_r(x_0)} (|\nabla u|^2 - |\nabla v_0|^2) = \int_{B_r(x_0)} |\nabla(u - v_0)|^2 \geqq c\mathcal{L}^n(B_r(x_0) \cap \{u = 0\}) \left(\frac{1}{r} ⨍_{\partial B_r(x_0)} u\right)^2$$
$$\geqq c\mathcal{L}^n(B_r(x_0) \cap \{u = 0\}),$$

where c is independent of ε. Consider now a free boundary point x_1 away from x_0. We can choose x_1 to be regular, say, $\partial\{u > 0\}$ is smooth in $B_{r_0}(x_1)$ for some small r_0. Near x_1 we make a smooth perturbation of the set $\{u > 0\}$ decreasing its volume by δ_r where

$$\delta_r := \mathcal{L}^n(B_r(x_0) \cap \{u = 0\}).$$

Let v_1 be the harmonic function in the perturbed set vanishing on its boundary and equal to u on $\partial B_{r_0}(x_1)$. Then

$$\int_{B_{r_0}(x_1)} (|\nabla v_1|^2 - |\nabla u|^2) = \lambda_\varepsilon^2 \delta_r + o(\delta_r).$$

Since the function

$$v := \begin{cases} v_0 & \text{in } B_r(x_0), \\ v_1 & \text{in } B_{r_0}(x_1), \\ u & \text{elsewhere,} \end{cases}$$

satisfies $\mathcal{L}^n(\{v > 0\}) = \mathcal{L}^n(\{u > 0\})$ we obtain

$$0 \leqq \mathcal{J}_\varepsilon(v) - \mathcal{J}_\varepsilon(u)$$
$$= \int_{B_r(x_0)} (|\nabla v_0|^2 - |\nabla u|)^2 + \int_{B_{r_0}(x_1)} (|\nabla v_1|^2 - |\nabla u|^2)$$
$$\leqq -c\delta_r + \lambda_\varepsilon^2 \delta_r + o(\delta_r),$$

that is, $\lambda_\varepsilon \geqq c$.

To prove that u is positive near y_0 we remark that the above estimate

$$\left(\frac{1}{r}\int_{\partial D_r} u\right)^2 \mathcal{L}^n(D_r \cap \{u=0\}) \leqq C \int_{D_r} |\nabla(u-v_0)|^2$$

is true for any domain D_r with smooth boundary, where v_0 is the harmonic extension of u on ∂D_r into D_r. The constant C depends only on the geometry and C^2 smoothness of the set

$$\frac{1}{r} D_r := \{x \in \mathbb{R}^n;\ rx \in D_r\}.$$

If we choose D_r with $\Omega \cap B_r(y_0) \subset D_r \subset B_{2r}(y_0)$ we obtain using the minimum property of u

$$\left(\frac{c_0}{r}\right)^2 \mathcal{L}^n(D_r \cap \{u=0\}) \leqq C\left(\int_{D_r} |\nabla u|^2 - \int_{D_r} |\nabla v_0|^2\right)$$

$$\leqq \frac{C}{\varepsilon} \mathcal{L}^n(D_r \cap \{u=0\});$$

hence $u>0$ in $\Omega \cap B_r(y_0)$ for small r (depending on ε).

We are now ready to prove

THEOREM 7. *For ε small, $\mathcal{L}^n(\{u_\varepsilon > 0\}) = \omega_0$, that is, u_ε minimizes our original functional $\mathcal{J}$ on K_0.*

Proof. By Lemmas 5 and 6

$$c \leqq \lambda_\varepsilon \leqq C.$$

Choose a regular point of the free boundary $\partial\{u_\varepsilon > 0\}$. Since the solution is smooth in a neighborhood of it, we can make regular perturbations as pointed out in the introduction.

We should remark that our results also apply if we replace the volume $\mathcal{L}^n(\{u>0\})$ by

$$\int_\Omega \rho\chi(\{u>0\})$$

with a smooth positive function ρ.

REFERENCES

[A] A. ACKER, *A free boundary optimization problem*, SIAM J. Math. Anal., 9 (1978), pp. 1179-1191.

[A-C] H. W. ALT AND L. A. CAFFARELLI, *Existence and regularity for a minimum problem with free boundary*, J. Reine Angew. Math., 325 (1981), pp. 105-144.

[F] K. FRIEDRICHS, *Über ein Minimumproblem für Potentialströmungen mit freiem Rand*, Math. Ann., 109 (1934), pp. 60-82.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 2, March 1986

004

THE DIRICHLET–NEUMANN BOUNDARY CONTROL PROBLEM ASSOCIATED WITH MAXWELL'S EQUATIONS IN A CYLINDRICAL REGION*

D. L. RUSSELL†

Abstract. In a cylindrical region we consider electromagnetic fields independent of the axial coordinate: controlling the time evolution of such fields by means of boundary currents, likewise independent of the axial direction, is equivalent to controlling, simultaneously, two wave equations; one with boundary control of Dirichlet type, the other of Neumann type. In this paper we provide a preliminary study of control problems of this type and indicate what is necessary for extensions of our work.

AMS (MOS) subject classifications. 93B05, 93C20, 78A25, 35L15, 35L20

Key words. hyperbolic PDE, control, boundary value control, distributed parameter systems, Maxwell equations, electromagnetic equations

1. Background. In this paper we consider a region $\Omega \subseteq R^3$, not necessarily bounded, having piecewise smooth boundary Γ and almost everywhere uniquely defined unit exterior normal vector $\vec{\nu} = \vec{\nu}(x, y, z)$, $(x, y, z) \in \Gamma$. It is assumed that the region Ω is occupied by a medium having constant electrical permittivity ε and constant magnetic permeability μ. We have then, in Ω, the paired electric and magnetic fields

$$\vec{E} = \vec{E}(x, y, z, t),$$
$$\vec{H} = \vec{H}(x, y, z, t),$$

having finite *energy*

$$\mathbf{E}(t) = \frac{1}{2}\int\int\int_{\Omega} (\varepsilon \|\vec{E}\|^2 + \mu \|\vec{H}\|^2)\, dv, \tag{1.1}$$

where $\|\cdot\|$ denotes the usual Euclidean norm in R^3. As is well known ([4], [9]), $\vec{E}$ and $\vec{H}$ satisfy, in Ω, Maxwell's equations

$$\operatorname{curl} \vec{H} = \varepsilon \frac{\partial \vec{E}}{\partial t}, \tag{1.2}$$

$$\operatorname{curl} \vec{E} = -\mu \frac{\partial \vec{H}}{\partial t}, \tag{1.3}$$

$$\operatorname{div} \vec{E} = \rho, \tag{1.4}$$

$$\operatorname{div} \vec{H} = 0, \tag{1.5}$$

where $\rho = \rho(x, y, z, t)$ is the electrical charge density in Ω—which is zero throughout this paper. (That (1.5) might eventually have to be modified to account for magnetic monopoles will trouble us not at all here!)

Control problems associated with Maxwell's equations have been of interest primarily in connection with nuclear fusion applications—in which case ρ is not

* Received by the editors March 20, 1984, and in revised form December 26, 1984. This research was sponsored by the U.S. Army under contract DAAG29-80-C-0041 and supported in part by the U.S. Air Force Office of Scientific Research under grant AFOSR-79-0018.

† Department of Mathematics, University of Wisconsin, Madison, Wisconsin 53706.

identically equal to zero and the Maxwell equations are coupled with the dynamical equations governing the plasma evolution. In this connection we cite the work of P. K. C. Wang [29], [30], [31]. The point of view which we take here is that we cannot hope to treat these more complicated problems until we have a firmer grasp on the control theory of Maxwell's system in its own right. In this direction some work on controllability with control influence distributed throughout Ω has been carried out by G. Chen [2], [3]. We are primarily concerned here with the possibility of influencing the evolution of the fields $\vec{E}$ and $\vec{H}$ by means of an externally determined current $\vec{J}(x, y, z, t)$ flowing tangentially in Γ so that

$$\vec{J}(x, y, z, t) \cdot \vec{\nu}(x, y, z) = 0, \tag{1.6}$$

for $(x, y, z) \in \Gamma$ where $\vec{\nu}(x, y, z)$ is defined. Then we have the boundary condition (see e.g. [4], [28])

$$\mu \vec{H}_\tau(x, y, z, t) = \vec{\nu}(x, y, z) \times \vec{J}(x, y, z, t) \tag{1.7}$$

for $(x, y, z) \in \Gamma$ such that $\vec{\nu}(x, y, z)$ is well defined. Here, and subsequently, the subscript τ refers to the component of the vector in question which is tangential to Γ. Similarly, the subscript ν will denote the normal component. Writing

$$\vec{H} = \vec{H}_\nu + \vec{H}_\tau,$$
$$\vec{J} = \vec{J}_\nu + \vec{J}_\tau = \vec{J}_\tau \quad \text{on } \Gamma,$$

we see that (1.7) becomes $\mu \vec{H}_\tau = \vec{\nu} \times \vec{J}_\tau$, so that $\vec{H}_\tau$ is a vector tangential to Γ and perpendicular to $\vec{J} = \vec{J}_\tau$.

The state space in which we study solutions of the above system will be denoted by $H_{\mathbf{E},d}(\Omega)$; it is a closed subspace of the space $H_{\mathbf{E}}(\Omega)$ of square integrable six-dimensional fields $(\vec{E}(x, y, z, t), \vec{H}(x, y, z, t))$ with the inner product and norm

$$\langle (\vec{E}_1, \vec{H}_1); (\vec{E}_2, \vec{H}_2) \rangle \equiv \iiint_\Omega (\varepsilon \vec{E}_1 \cdot \vec{E}_2 + \mu \vec{H}_1 \cdot \vec{H}_2)\, dv, \tag{1.8}$$

$$\| (\vec{E}, \vec{H}) \|^2 = \langle (\vec{E}, \vec{H}); (\vec{E}; \vec{H}) \rangle. \tag{1.9}$$

Clearly $H_{\mathbf{E}}(\Omega)$ is a real Hilbert space with this inner product. Where a complex space is required, we employ conjugation as usual. The state space $H_{\mathbf{E},d}(\Omega)$ is the closed span in $H_{\mathbf{E}}(\Omega)$ of those continuously differentiable fields $(\vec{E}(x, y, z, t), \vec{H}(x, y, z, t))$ for which

$$\operatorname{div} \vec{E} = \frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z} = 0,$$

$$\operatorname{div} \vec{H} = \frac{\partial H_x}{\partial x} + \frac{\partial H_y}{\partial y} + \frac{\partial H_z}{\partial z} = 0.$$

If $\vec{E}_0, \vec{H}_0$ and $\vec{E}_1, \vec{H}_1$ are two smooth solution pairs for (1.2)-(1.5), (1.7), the first corresponding to $\vec{J} \equiv 0$ on Γ, we see easily that

$$\frac{d}{dt} \langle (\vec{E}_0, \vec{H}_0); (\vec{E}_1, \vec{H}_1) \rangle$$

$$= \iiint_\Omega \left(\varepsilon \left[\vec{E}_0 \cdot \frac{\partial \vec{E}_1}{\partial t} + \frac{\partial \vec{E}_0}{\partial t} \cdot \vec{E}_1 \right] + \mu \left[\vec{H}_0 \cdot \frac{\partial \vec{H}_1}{\partial t} + \frac{\partial \vec{H}_0}{\partial t} \cdot \vec{H}_1 \right] \right) dv$$

$$= (\text{using (1.2), (1.3)})$$

$$= \iiint_{\Omega} (\vec{E}_0 \cdot \operatorname{curl} \vec{H}_1 - \operatorname{curl} \vec{E}_0 \cdot \vec{H}_1 + \operatorname{curl} \vec{H}_0 \cdot \vec{E}_1 - \vec{H}_0 \cdot \operatorname{curl} \vec{E}_1)\, dv$$

$$= (\text{using } \operatorname{div}(E \times H) = \operatorname{curl} \vec{E} \cdot \vec{H} - \vec{E} \cdot \operatorname{curl} \vec{H})$$

$$= -\iiint_{\Omega} [\operatorname{div}(\vec{E}_0 \times \vec{H}_1) + \operatorname{div}(\vec{E}_1 \times \vec{H}_0)]\, dv \tag{1.10}$$

$$= -\iint_{\Gamma} (\vec{E}_0 \times \vec{H}_1 + \vec{E}_1 \times \vec{H}_0) \cdot \nu\, ds$$

$$= -\iint_{\Gamma} (\vec{E}_{0\tau} \times \vec{H}_{1\tau} + \vec{E}_{0\tau} \times \vec{H}_{1\nu} + \vec{E}_{1\tau} \times \vec{H}_{0\tau} + \vec{E}_{1\tau} \times \vec{H}_{0\nu}) \cdot \nu\, ds$$

$$= -\iint_{\Gamma} (\vec{E}_{0\tau} \times \vec{H}_{1\tau} + \vec{E}_{1\tau} \times \vec{H}_{0\tau}) \cdot \nu\, ds$$

$$= (\text{using (1.7) and noting that } \vec{J} \equiv 0 \text{ for } \vec{E}_0, \vec{H}_0)$$

$$= -\iint_{\Gamma} (\vec{E}_{0\tau} \cdot \vec{J})\, ds.$$

If we go through the same computation with $\vec{E}_0$, $\vec{H}_0$, $\vec{E}_1$, $\vec{H}_1$ both replaced by the same $\vec{E}$, $\vec{H}$ satisfying (1.2)-(1.5), (1.7), (1.8), we find that

$$\frac{d\mathbf{E}}{dt} = -\iint_{\Gamma} (\vec{E} \times \vec{H}) \cdot \vec{v}\, ds = -\iint_{\Gamma} \vec{E}_{\tau} \cdot \vec{J}\, ds. \tag{1.11}$$

For $\vec{J} \equiv 0$ generalized solutions of (1.2)-(1.5), (1.7), (1.8) can be discussed in the general context of partial differential equations and strongly continuous semigroups. The generator

$$A(\vec{E}, \vec{H}) = \left(\frac{1}{\varepsilon} \operatorname{curl} \vec{H}, -\frac{1}{\mu} \operatorname{curl} \vec{E}\right) \tag{1.12}$$

with domain consisting of $\vec{E}$, $\vec{H}$ in the Sobolev space $H^1_{\mathbf{E},d}(\Omega)(= H_{\mathbf{E},d}(\Omega) \cap H^1(\Omega))$ having zero divergence and satisfying (cf. (1.7))

$$\vec{H}_{\tau}|_{\Gamma} = 0, \tag{1.13}$$

is antisymmetric and generates a group of isometries in $H_{\mathbf{E},d}(\Omega)$. (See [32], [33], [34] for related work.) Sufficient conditions on $\vec{J}$ so that solutions of the inhomogeneous system (1.2)-(1.5), (1.7), (1.8) lie in $H_{\mathbf{E},d}(\Omega)$ and are strongly continuous there may be obtained much as in [18], [19] but it is not easy to specify necessary and sufficient conditions. Indeed, this is already difficult for the much simpler, but related, wave

equation

$$\mu\varepsilon\frac{\partial^2 w}{\partial t^2}=\frac{\partial^2 w}{\partial x^2}+\frac{\partial^2 w}{\partial y^2}+\frac{\partial^2 w}{\partial z^2}$$

with boundary forcing terms. We will make some comments related to this in § 6.

2. Control problems in a cylindrical region. The main point in this paper is to study the question of controllability of the electromagnetic field $\vec{E}$, $\vec{H}$ by means of the boundary current $\vec{J}=\vec{J}_\tau$. By controllability we mean the possibility of transferring an initial field $\vec{E}(x, y, z, 0)$, $\vec{H}(x, y, z, 0)\in H_{\mathbf{E},d}(\Omega)$, given at time $t=0$, to a prescribed terminal field $\vec{E}(x, y, z, T)$, $\vec{H}(x, y, z, T)\in H_{\mathbf{E},d}(\Omega)$, specified at $t=T>0$, by means of a suitable control current $\vec{J}(x, y, z, t)$ defined for $(x, y, z)\in\Gamma$, $t\in[0, T]$. Because the homogeneous Maxwell equations correspond to a group of isometries in $H_{\mathbf{E},d}(\Omega)$, it is enough to consider the special case wherein

$$\vec{E}(x, y, z, 0)\equiv 0, \tag{2.1}$$

$$\vec{H}(x, y, z, 0)\equiv 0. \tag{2.2}$$

For a given space, $\mathbf{J}$, of admissible control currents $\vec{J}(x, y, z, t)=\vec{J}_\tau(x, y, z, t)$ defined on $\Gamma\times[0, T]$ we define the reachable set $R(T, \mathbf{J})$ to be the subspace of $H_{\mathbf{E},d}(\Omega)$ consisting of states reachable from the zero initial state using controls $\vec{J}\in\mathbf{J}$. Following earlier definitions [26] our system is *approximately controllable in time* T if $R(T, \mathbf{J})$ is dense in $H_{\mathbf{E},d}(\Omega)$ and *exactly controllable in time* T if $R(T, \mathbf{J})=H_{\mathbf{E},d}(\Omega)$ (or some precisely designated subspace of $H_{\mathbf{E},d}(\Omega)$).

At this writing we are not able to discuss the general three-dimensional problem wherein the vector fields $\vec{E}$ and $\vec{H}$ are unrestricted and Ω has a general geometry. We hope in later work to consider at least some three-dimensional cases which arise for special domains Ω. For the present we must content ourselves with an analysis of certain cases in which Ω is a cylinder:

$$\Omega=\mathbf{R}\times(-\infty, \infty)=\{(x, y, z)\,|\,(x, y)\in R\quad R^2,\ z \text{ real}\}$$

$\mathbf{R}$ being an open connected region in R^2 with piecewise smooth boundary $\mathbf{B}$. Thus

$$\partial\Omega=\partial\mathbf{R}\times(-\infty, \infty)=\mathbf{B}\times(-\infty, \infty).$$

Even here we can give results only for special two-dimensional regions $\mathbf{R}$.

The two-dimensional problem in the cylinder $\Omega=\mathbf{R}\times(-\infty, \infty)$ occurs when we confine attention to fields

$$\vec{E}=\vec{E}(x, y, t),\qquad \vec{H}=\vec{H}(x, y, t)$$

which do not depend on the coordinate z corresponding to the axial, or longitudinal, direction of the cylinder. (Note that this is not at all the same thing as requiring that E_z, H_z, the field components in the z direction, should be zero.) We correspondingly consider only control currents

$$\vec{J}=\vec{J}(x, y, t)$$

which do not depend upon z.

We will see in § 3 that the only "interesting" control problems occur in the case where $\vec{J}$ is not permitted to be an arbitrary vector tangent to $\partial\Omega$ but, rather, has a preassigned direction in the tangent space (but arbitrary sign and magnitude). In this section and § 3 we will see that this case reduces to control of two separate wave equations with a single control function entering into both of them.

Then, in §§ 5 and 6 we will investigate the case where Ω is a rectangle and the control boundary is one side and the case where Ω is a disc and control is applied on the whole boundary, respectively. In the first instance we obtain an approximate controllability result and in the second, an exact controllability result relative to transfer between finite energy states. In each case the minimum time for control turns out to be twice the time required for control of a single wave equation under comparable circumstances. It is our expectation that this will turn out to be a general rule but, for the present, this can only be offered as a conjecture.

It is relatively easy to replace the requirement that the fields should be constant relative to the z-direction by the requirement that they should be periodic in the z-direction. This work will appear separately.

Of course the energy $\mathbf{E}$ in the cylinder Ω is infinite if $\vec{E}, \vec{H}$ are not identically zero. We redefine $\mathbf{E}$ to be the energy per unit length of cylinder:

$$\mathbf{E}(t) = \frac{1}{2}\int\int_{\mathbf{R}} (\varepsilon \|\vec{E}(x, y, t)\|^2 + \mu \|\vec{H}(x, y, t)\|^2)\, dx\, dy. \tag{2.3}$$

The space $H_{\mathbf{E},d}(\Omega)$ is now replaced by $H_{\mathbf{E},d}(\mathbf{R})$. Because

$$\frac{\partial E_z(x, y, t)}{\partial z} \equiv 0, \qquad \frac{\partial H_z(x, y, t)}{\partial z} \equiv 0,$$

we have

$$\operatorname{div} \vec{E} = \frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y}, \qquad \operatorname{div} \vec{H} = \frac{\partial H_x}{\partial x} + \frac{\partial H_y}{\partial y}. \tag{2.4}$$

The curl expressions simplify to

$$\operatorname{curl} \vec{E} = \left(\frac{\partial E_z}{\partial y}, -\frac{\partial E_z}{\partial x}, \frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y}\right),$$

$$\operatorname{curl} \vec{H} = \left(\frac{\partial H_z}{\partial y}, -\frac{\partial H_z}{\partial x}, \frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y}\right),$$

so that the equations (1.2), (1.3) become

$$\begin{aligned}
&\text{(i)} \quad \varepsilon \frac{\partial E_x}{\partial t} = \frac{\partial H_z}{\partial y}, && \text{(iv)} \quad \mu \frac{\partial H_x}{\partial t} = -\frac{\partial E_z}{\partial y}, \\
&\text{(ii)} \quad \varepsilon \frac{\partial E_y}{\partial t} = -\frac{\partial H_z}{\partial x}, && \text{(v)} \quad \mu \frac{\partial H_y}{\partial t} = \frac{\partial E_z}{\partial x}, \\
&\text{(iii)} \quad \varepsilon \frac{\partial E_z}{\partial t} = \frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y}, && \text{(vi)} \quad \mu \frac{\partial H_z}{\partial t} = -\frac{\partial E_y}{\partial x} + \frac{\partial E_x}{\partial y}.
\end{aligned} \tag{2.5}$$

It is clear from (2.5) (i)-(vi), that if $\vec{E}(x, y, 0)$, $\vec{H}(x, y, 0)$ are given, then the subsequent evolution of $E_z(x, y, t)$, $H_z(x, y, t)$ determine all of the other components. As for these components themselves, differentiating (2.5)(iii) and (2.5)(vi) with respect to t and then substituting (2.5)(iv), (v) and (2.5)(i), (ii) into the respectively resulting expressions, we obtain the familiar wave equations

$$\mu\varepsilon \frac{\partial^2 E_z}{\partial t^2} = \frac{\partial^2 E_z}{\partial x^2} + \frac{\partial^2 E_z}{\partial y^2}, \tag{2.6}$$

$$\mu\varepsilon\frac{\partial^2 H_z}{\partial t^2}=\frac{\partial^2 H_z}{\partial x^2}+\frac{\partial^2 H_z}{\partial y^2}, \tag{2.7}$$

valid for $(x, y)\in\mathbf{R}$, $t\in[0,\infty)$, provided E_z, H_z have enough derivatives, or provided the equations are interpreted in the distributional sense. Assuming the initial states $\vec{E}(x, y, 0)$, $\vec{H}(x, y, 0)$ are divergence-free, we compute (cf. (2.4))

$$\varepsilon\frac{\partial}{\partial t}\left(\frac{\partial E_x}{\partial x}+\frac{\partial E_y}{\partial y}\right)=(\text{using (2.5)(i), (ii)})=\varepsilon\left(\frac{\partial^2 H_z}{\partial x\,\partial y}-\frac{\partial^2 H_z}{\partial y\,\partial x}\right)=0$$

and similarly

$$\mu\frac{\partial}{\partial t}\left(\frac{\partial H_x}{\partial x}+\frac{\partial H_y}{\partial y}\right)=0$$

and we conclude that the fields remain divergence-free for all time.

Suppose, then, that divergence-free initial states $\vec{E}(x, y, 0)$, $\vec{H}(x, y, 0)$ are given. Then $E_z(x, y, 0)$, $H_z(x, y, 0)$ are known and (2.5)(iii), (vi) determine $(\partial E_z/\partial t)(x, y, 0)$ and $(\partial H_z/\partial t)(x, y, 0)$. If (2.6), (2.7) are then solved with these initial conditions, and appropriate boundary conditions, the complete solution of Maxwell's equations (2.5)(i)-(vi), can be obtained by integrating (2.5)(i), (ii), (iv), (v). Thus it is enough to work with (2.6), (2.7), and it should be noted that the divergence condition does not have any bearing on E_z, H_z; it can be ignored henceforth.

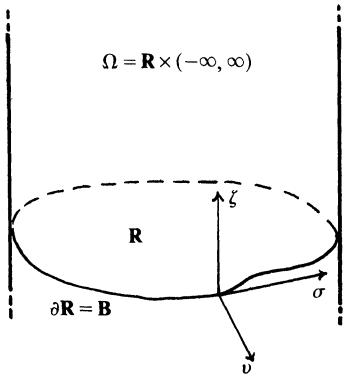


FIG. 1. *The region* **R**.

It is important to recast the boundary condition (1.7) so that it provides boundary conditions for (2.6), (2.7). We ask the reader to consult Fig. 1, where the region **R** with boundary $\partial\mathbf{R}=\mathbf{B}$ is shown. At a point $(x, y)\in\mathbf{B}$ we let $\vec{\nu}=\vec{\nu}(x, y)$ denote the unit exterior normal to **B** and we let $\vec{\sigma}=\vec{\sigma}(x, y)$ denote the positively oriented unit tangent vector to **B** there. With $\vec{\zeta}$, the unit vector in the positive z direction, $\vec{\nu}$, $\vec{\sigma}$, $\vec{\zeta}$ form a positively oriented orthogonal triple of unit vectors. Given an arbitrary vector w, we can decompose it as

$$\vec{w}=(w_\nu, w_\sigma, w_\zeta(=w_z)),$$
$$\|\vec{w}\|^2=w_\nu^2+w_\sigma^2+w_z^2.$$

The tangential part of $\vec{H}$, which we have designated as $\vec{H}_\tau$, may now be represented as

$$\vec{H}_\tau=H_z\vec{\zeta}+H_\sigma\vec{\sigma} \tag{2.8}$$

and the current $\vec{J}=\vec{J}_\tau$ may likewise be represented as

$$\vec{J}_\tau=J_z\vec{\zeta}+J_\sigma\vec{\sigma}.$$

Then

$$\vec{\nu} \times \vec{J} = \vec{\nu} \times \vec{J}_\tau = \vec{\nu} \times (J_z \vec{\zeta} + J_\sigma \vec{\sigma}) = -J_z \vec{\sigma} + J_\sigma \vec{\zeta}. \tag{2.9}$$

Combining (1.8), (2.8), (2.9), we see that on $\mathbf{B}$

$$H_z(x, y, t) = J_\sigma(x, y, t), \tag{2.10}$$

$$H_\sigma(x, y, t) = -J_z(x, y, t). \tag{2.11}$$

Represent $\vec{\nu}$, $\vec{\sigma}$ as

$$\vec{\nu} = \nu_x \vec{\xi} + \nu_y \vec{\eta}, \tag{2.12}$$

$$\vec{\sigma} = \sigma_x \vec{\xi} + \sigma_y \vec{\eta} = -\nu_y \vec{\xi} + \nu_x \vec{\eta}. \tag{2.13}$$

Then compute

$$\begin{aligned} \frac{\partial E_z}{\partial \nu} &= \frac{\partial E_z}{\partial x} \nu_x + \frac{\partial E_z}{\partial y} \nu_y \\ &= (\text{using (1.3)}) \\ &= \mu \frac{\partial H_y}{\partial t} \sigma_y + \mu \frac{\partial H_x}{\partial t} \sigma_x = \mu \frac{\partial H_\sigma}{\partial t} \\ &= (\text{using (2.11)}) = -\frac{\partial J_z}{\partial t}. \end{aligned} \tag{2.14}$$

The equations (2.10), (2.14) provide the needed boundary conditions for (2.6), (2.7) respectively. For H_z we have the Dirichlet-type boundary condition (2.10) while for E_z we have the Neumann-type boundary condition (2.14). If we let

$$\vec{U}(x, y, t) = \frac{\partial \vec{J}}{\partial t}(x, y, t),$$

$$\vec{U} = \vec{U}_\tau = U_\sigma \vec{\sigma} + U_z \vec{\zeta},$$

and differentiate (2.10), we have the more symmetric form

$$\frac{\partial H_z}{\partial t}(x, y, t) = U_\sigma(x, y, t), \quad \frac{\partial E_z}{\partial \nu} = -U_z(x, y, t), \quad (x, y) \in \mathbf{B}. \tag{2.15}$$

We complete this section by discussing the question of expression of the energy per unit cylinder length, (2.3), solely in terms of H_z and E_z.

We consider the equations (2.6), (2.7) with homogeneous boundary conditions

$$\frac{\partial H_z}{\partial t}(x, y, t) = 0, \quad \frac{\partial E_z}{\partial \nu}(x, y, t) = 0, \quad (x, y) \in \mathbf{B}.$$

We use the symbol Δ for the Laplacian:

$$\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}.$$

Initially we take H_z, E_z to lie in the Sobolev space $H^2(\mathbf{R})$. This space must be decomposed in order to attach a meaning to Δ^{-1}.

The boundary condition for H_z may be rewritten as

$$H_z(x, y, t) = h(x, y), \qquad (x, y) \in \mathbf{B},$$

where, by the trace theorem, $h \in H^{3/2}(\mathbf{B})$. Then we can write

$$H_z(x, y, t) = \hat{H}_z(x, y, t) + \tilde{H}_z(x, y)$$

where $\tilde{H}_z(x, y)$ is the solution of

$$\Delta \tilde{H}_z(x, y) = 0, \quad \tilde{H}_z(x, y) = h(x, y), \quad (x, y) \in \mathbf{B}$$

and

$$\hat{H}_z(x, y, t) = 0, \qquad (x, y) \in \mathbf{B}.$$

The inverse Laplacian Δ^{-1} is well defined on the functions $\hat{H}_z$. For E_z we may write

$$E_z(x, y, t) = \hat{E}_z(x, y, t) + \tilde{E}_z(t)$$

where $\tilde{E}_z$, as indicated, is constant with respect to $(x, y) \in \mathbf{R}$ and

$$\int_{\mathbf{B}} \hat{E}_z(x, y, t)\, ds = 0.$$

It is well known that Δ^{-1} is well defined on the functions $\hat{E}_z$.

We proceed first on the assumption that

$$H_z(x, y, t) = \hat{H}_z(x, y, t), \qquad E_z(x, y, t) = \hat{E}_z(x, y, t).$$

We form new solutions of (2.6), (2.7) by setting

$$\mu \frac{\partial G_z}{\partial t} = -E_z, \qquad \varepsilon \frac{\partial F_z}{\partial t} = H_z,$$

$$G_z = \varepsilon \Delta^{-1} \frac{\partial E_z}{\partial t}, \qquad F_z = \mu \Delta^{-1} \frac{\partial H_z}{\partial t}.$$

We then determine G_x, G_y, F_x, F_y, using the equations (2.5) with $\vec{G}$ replacing $\vec{H}$, $\vec{F}$ replacing $\vec{E}$, so that $\vec{F}$ and $\vec{G}$ satisfy Maxwell's equations:

$$\mu \frac{\partial \vec{G}}{\partial t} = -\operatorname{curl} \vec{F}, \qquad \varepsilon \frac{\partial \vec{F}}{\partial t} = \operatorname{curl} \vec{G}.$$

It will then be found that

$$\vec{E} = \operatorname{curl} \vec{F}, \qquad \vec{H} = \operatorname{curl} \vec{G}.$$

Following this, (2.3) can be written as

$$\mathbf{E}(t) = \frac{1}{2} \iint_{\mathbf{R}} (\varepsilon \|\operatorname{curl} \vec{F}\|^2 + \mu \|\operatorname{curl} \vec{G}\|^2)\, dx\, dy$$

$$= \frac{1}{2} \iint_{\mathbf{R}} \varepsilon \left[\left(\frac{\partial F_z}{\partial x} \right)^2 + \left(\frac{\partial F_z}{\partial y} \right)^2 + \left(\frac{\partial F_y}{\partial x} - \frac{\partial F_x}{\partial y} \right)^2 \right] \tag{2.16}$$

$$+ \mu \left[\left(\frac{\partial G_z}{\partial x} \right)^2 + \left(\frac{\partial G_z}{\partial y} \right)^2 + \left(\frac{\partial G_y}{\partial x} - \frac{\partial G_x}{\partial y} \right)^2 \right] dx\, dy.$$

Then from (2.16) we have

$$\mathbf{E}(t)=\frac{1}{2}\iint_{\mathbf{R}}\left\{\varepsilon\left[\left(\frac{\partial F_z}{\partial x}\right)^2+\left(\frac{\partial F_z}{\partial y}\right)^2+\left(\mu\frac{\partial G_z}{\partial t}\right)^2\right]\right.$$
$$\left.+\mu\left[\left(\frac{\partial G_z}{\partial x}\right)^2+\left(\frac{\partial G_z}{\partial y}\right)^2+\left(\varepsilon\frac{\partial F_z}{\partial t}\right)^2\right]\right\}dx\,dy$$
$$=\frac{1}{2}\iint_{\mathbf{R}}\varepsilon\left[\left(\frac{\partial F_z}{\partial x}\right)^2+\left(\frac{\partial F_z}{\partial y}\right)^2+(E_z)^2\right]+\mu\left[\left(\frac{\partial G_z}{\partial x}\right)^2+\left(\frac{\partial G_z}{\partial y}\right)^2+(H_z)^2\right]dx\,dy.$$

Now consider the quadratic form, for $E_z=\hat{E}_z$ and $(\ ,\ )$ the inner product in $L^2(\mathbf{R})$,

$$\left(\frac{\partial E_z}{\partial t},-\Delta^{-1}\frac{\partial E_z}{\partial t}\right)=\left(\mu\frac{\partial^2 G_z}{\partial t^2},-\Delta^{-1}\frac{\partial^2 G_z}{\partial t^2}\right)$$
$$=(\text{since } G_z \text{ satisfies the wave equation } \mu\varepsilon\,\partial^2 G_z/\partial t^2=\Delta G_z$$
$$\text{and the boundary conditions } G_z(x,y,t)=0,\ (x,y)\in\mathbf{B})$$
$$=\frac{1}{\mu\varepsilon^2}(-\Delta G_z,G_z)=\frac{1}{\mu\varepsilon^2}\iint_{\mathbf{R}}\left[\left(\frac{\partial G_z}{\partial x}\right)^2+\left(\frac{\partial G_z}{\partial y}\right)^2\right]dx\,dy.$$

Similarly

$$\left(\frac{\partial H_z}{\partial t},-\Delta^{-1}\frac{\partial H_z}{\partial t}\right)=\frac{1}{\varepsilon\mu^2}\iint_{\mathbf{R}}\left[\left(\frac{\partial F_z}{\partial x}\right)^2+\left(\frac{\partial F_z}{\partial y}\right)^2\right]dx\,dy$$

from which it follows that

$$\mathbf{E}(t)=\frac{(\mu\varepsilon)^2}{2}\left[\left(\frac{\partial E_z}{\partial t},-\Delta^{-1}\frac{\partial E_z}{\partial t}\right)+\left(\frac{\partial H_z}{\partial t},-\Delta^{-1}\frac{\partial H_z}{\partial t}\right)\right]+\frac{1}{2}\iint_{\mathbf{R}}\left[\varepsilon(E_z)^2+\mu(H_z)^2\right]dx\,dy.$$

It is necessary to modify this expression for general E_z, H_z. We begin with

$$E_z(x,y,t)=\tilde{E}_z(t).$$

The only possible solutions of the wave equation (2.6) satisfying $\partial E_z/\partial\nu|_B=0$ and having this form are

$$E_z(x,y,t)=e_0+e_1t$$

where e_0 and e_1 are constants. (Such solutions are consistent with a constant boundary current J for which $J_\sigma\equiv 0$.) The corresponding E_x, E_y, H_z are zero but

$$\varepsilon e_1=\varepsilon\frac{\partial E_z}{\partial t}=\frac{\partial H_y}{\partial x}-\frac{\partial H_x}{\partial y}.$$

It is not possible to express this quantity in terms of E_z itself or H_z. It is better to leave it in the form $\varepsilon\,\partial E_z/\partial t$. Solutions of Maxwell's equations with E_z having this form have energy expressible as a quadratic form in E_z and $\partial E_z/\partial t$.

Next we consider $H_z=\tilde{H}_z$ as described earlier. Such a solution is consistent with a boundary current for which $J_\tau=0$, constant with respect to time but possibly varying

with $(x, y) \in B$. We may take H_x, H_y, E_z all zero. However,

$$\varepsilon \frac{\partial E_x}{\partial t} = \frac{\partial H_z}{\partial y}, \qquad \varepsilon \frac{\partial E_y}{\partial t} = -\frac{\partial H_z}{\partial x}$$

so we may not assume that E_x and E_y are equal to zero. The energy associated with solutions of this type is expressible in terms of

$$\iint_{\mathbf{R}} \left[\left(\frac{\partial H_z}{\partial x} \right)^2 + \left(\frac{\partial H_z}{\partial y} \right)^2 \right] dx\, dy$$

if integration with respect to t is permitted. In the sequel we will not explicitly consider the timewise linear electric fields satisfying the above equations.

We see then that a norm involving only E_z and H_z and compatible with the energy (2.3) may be expressed as

$$|(E_z, H_z)|^2 = (\mu\varepsilon)^2 \left[\left(\frac{\partial \hat{E}_z}{\partial t}, -\Delta^{-1} \frac{\partial \hat{E}_z}{\partial t} \right) + \left(\frac{\partial \hat{H}_z}{\partial t}, -\Delta^{-1} \frac{\partial \hat{H}_z}{\partial t} \right) \right]$$

$$+ \iint_{\mathbf{R}} \left[\varepsilon (\hat{E}_z)^2 + \mu (\hat{H}_z)^2 + \rho_0 (\tilde{E}_z)^2 \right. \tag{2.17}$$

$$\left. + \rho_1 \left(\frac{\partial \tilde{E}_z}{\partial t} \right)^2 + \sigma_0 \left(\frac{\partial \tilde{H}_z}{\partial x} \right)^2 + \sigma_1 \left(\frac{\partial \tilde{H}_z}{\partial y} \right)^2 \right] dx\, dy$$

where ρ_0, ρ_1, σ_0, σ, are positive numbers. It will be seen that this is a weaker norm than the one associated with a pair of wave equations, viz.:

$$|(E_z, H_z)|^2 = \iint_{\mathbf{R}} \left\{ \mu\varepsilon \left[\left(\frac{\partial E_z}{\partial t} \right)^2 + \left(\frac{\partial H_z}{\partial t} \right)^2 \right] + |\nabla E_z|^2 + |\nabla H_z|^2 \right\} dx\, dy. \tag{2.18}$$

We will denote the Hilbert space of states E_z, H_z, $\partial E_z/\partial t$, $\partial H_z/\partial t$ lying in $H^1(\mathbf{R})$, $H^1(\mathbf{R})$, $L^2(\mathbf{R})$, $L^2(\mathbf{R})$, respectively, by $\hat{H}$. This space will be very convenient for use in the remainder of this paper. In some cases we will add boundary conditions to the specification of H, the space with norm $|\cdot|$, without changing the symbol, to correspond to an agreed specification of the states in $\hat{H}$ by similar boundary conditions.

3. Some control configurations. We describe here two possible realizations of the control problem which we have posed and indicate why we have chosen the mathematically more interesting (i.e., more difficult) one to work with in this paper.

Let us assume that $\Gamma = \partial\Omega = B \times (-\infty, \infty)$ is covered by one or more layers of conducting bars, arranged in rows as shown in Fig. 3.1. In the case of a single layer of conducting bars shown in Fig. 2(b), the bars are arranged so that they make an angle θ, $0 < |\theta| < \pi/2$, with the vector $\vec{\sigma}$ (cf. Fig. 1), while in the double layer case (Fig. 2(a)) they are arranged so that the bars in the second layer make an angle ψ, $0 < |\psi| < \pi/2$, $\psi \neq \theta$, with the vector $\vec{\sigma}$. The current in any row of bars parallel to the z-axis is independent of z; i.e., constant for all bars in that row. As we consider successively smaller bars we obtain, as an idealization, the boundary current vector

$$\vec{J}(x, y, t) = J(x, y, t)(\cos\theta \vec{\sigma} + \sin\theta \vec{\zeta}) \tag{3.1}$$

in the single layer case, $J(x, y, t)$ denoting the current strength with the sign determined

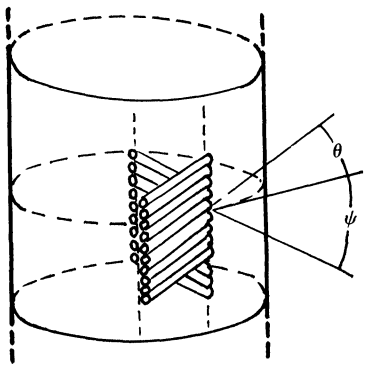


FIG. 2(a). *Double layer control.*

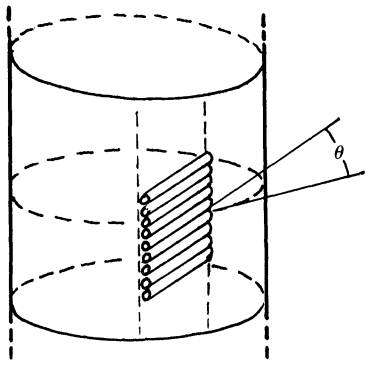


FIG. 2(b). *Single layer control.*

so that J positive yields a positive current component in the $\vec{\sigma}$ direction. The corresponding formula in the double layer case is

$$\vec{J}(x, y, t) = J_1(x, y, t)(\cos\theta\vec{\sigma} + \sin\theta\vec{\zeta}) + J_2(x, y, t)(\cos\psi\vec{\sigma} + \sin\psi\vec{\zeta}). \tag{3.2}$$

The current components are, in the single layer case

$$J_\sigma(x, y, t) = J(x, y, t)\cos\theta,$$

$$J_z(x, y, t) = J(x, y, t)\sin\theta,$$

and in the double layer case,

$$\begin{pmatrix} J_\sigma(x, y, t) \\ J_z(x, y, t) \end{pmatrix} = \begin{pmatrix} \cos\theta & \cos\psi \\ \sin\theta & \sin\psi \end{pmatrix} \begin{pmatrix} J_1(x, y, t) \\ J_2(x, y, t) \end{pmatrix}. \tag{3.3}$$

The determinant of the matrix in (3.3) is $\sin(\psi - \theta) \neq 0$ if $\psi \neq \theta$ in the range $0 < |\theta| < \pi/2$, $0 < |\psi| < \pi/2$. Thus in the double layer case J_σ and J_z are independent if J_1 and J_2 are independent while in the single layer case J_σ and J_z are fixed nonzero multiples of each other.

The double layer case is easily disposed of in the light of earlier work on boundary control of the wave equation. Referring back to (2.10), (2.11), we now have, for $(x, y) \in \mathbf{B} = \partial\mathbf{R}$, $t \in [0, \infty)$,

$$\frac{\partial H_z}{\partial t}(x,y,t) = U_\sigma(x, y, t) = \cos\theta u_1(x, y, t) + \cos\psi u_2(x, y, t),$$

$$\frac{\partial E_z}{\partial \nu}(x, y, t) = -U_z(x, y, t) = -\sin\theta u_1(x, y, t) + \cos\psi u_2(x, y, t),$$

$$u_1(x, y, t) = \frac{\partial J_1}{\partial t}(x, y, t), \qquad u_2(x, y, t) = \frac{\partial J_2}{\partial t}(x, y, t).$$

Since U_σ and U_z are independent if u_1 and u_2 are, the control problem splits into two uncoupled wave-equation problems, one for E_z and one for H_z. These have been discussed thoroughly in [2], [3], [15], [16], [22], [23], [25] with affirmative controllability results for various control configurations and will not concern us further here.

In the remainder of this paper we study the single layer case. If we let

$$u(x, y, t) = \frac{\partial J}{\partial t}(x, y, t) \tag{3.4}$$

we now have the wave equations (2.6), (2.7) for E_z, H_z and the boundary conditions

$$\frac{\partial H_z}{\partial t}(x, y, t) = \cos\theta \frac{\partial J}{\partial t}(x, y, t) \equiv \alpha u(x, y, t), \tag{3.5}$$

$$\frac{\partial E_z}{\partial \nu}(x, y, t) = -\sin\theta \frac{\partial J}{\partial t}(x, y, t) \equiv \beta u(x, y, t). \tag{3.6}$$

The control problems for E_z and H_z are now coupled because the single control function, $u(x, y, t)$, appears in the boundary conditions for both E_z and H_z; we have to control both systems simultaneously using the same control function.

If we rely on experience in a single space dimension, which has proved generally quite helpful in the control theory of a single wave equation, we are led to believe that systems like (2.6), (2.7), (3.5), (3.6) may, in fact, be controllable. Replacing $u(x, y, t)$ by $u_0(t)$, $u_1(t)$ and taking $0 \leqq x \leqq 1$, the one-dimensional equations are, using variables v, w,

$$\rho \frac{\partial^2 v}{\partial t^2} - \frac{\partial^2 v}{\partial x^2} = 0, \tag{3.7}$$

$$\frac{\partial v}{\partial t}(0, t) = \alpha u_0(t), \qquad \frac{\partial v}{\partial t}(1, t) = \alpha u_1(t), \tag{3.8}$$

$$\rho \frac{\partial^2 w}{\partial t^2} - \frac{\partial^2 w}{\partial x^2} = 0, \tag{3.9}$$

$$\frac{\partial w}{\partial x}(0, t) = -\beta u_0(t), \qquad \frac{\partial w}{\partial x}(1, t) = \beta u_1(t), \tag{3.10}$$

(note that $-\partial w/\partial x$ corresponds to the exterior normal derivative at 0). Letting

$$\tilde{v} = \frac{\partial v}{\partial x}, \tag{3.11}$$

$$\tilde{w} = \frac{\partial w}{\partial t}, \tag{3.12}$$

we find that

$$\rho \frac{\partial^2 \tilde{v}}{\partial t^2} - \frac{\partial^2 \tilde{v}}{\partial x^2} = 0, \tag{3.13}$$

and

$$\rho \frac{\partial^2 \tilde{w}}{\partial t^2} - \frac{\partial^2 \tilde{w}}{\partial x^2} = 0. \tag{3.14}$$

Differentiating (3.11) with respect to t and using (3.8), we have

$$\frac{1}{\rho}\frac{\partial^2 v}{\partial x^2}(0, t) = \frac{1}{\rho}\frac{\partial \tilde{v}}{\partial x}(0, t) = \frac{\alpha}{\rho} u_0'(t), \tag{3.15}$$

$$\frac{1}{\rho}\frac{\partial^2 v}{\partial x^2}(1, t) = \frac{1}{\rho}\frac{\partial \tilde{v}}{\partial x}(1, t) = \frac{\alpha}{\rho} u_1'(t), \tag{3.16}$$

while differentiation of (3.12) along with (3.10) yields

$$\frac{\partial^2 w}{\partial t\, \partial x}(0, t) = \frac{\partial \tilde{w}}{\partial x}(0, t) = -\beta u_0'(t), \tag{3.17}$$

$$\frac{\partial^2 w}{\partial t\,\partial x}(1,t)=\frac{\partial \tilde{w}}{\partial x}(1,t)=\beta u_1'(t). \tag{3.18}$$

Combining (3.13) with (3.14), (3.15), (3.16), (3.17), (3.18), we see that $\beta\tilde{v}+\alpha\tilde{w}/\rho$, $\beta\tilde{v}-\alpha\tilde{w}/\rho$ both satisfy the wave equation and

$$\frac{\partial}{\partial x}\left(\beta\tilde{v}+\frac{\alpha}{\rho}\tilde{w}\right)(0,t)=0, \qquad \frac{\partial}{\partial x}\left(\beta\tilde{v}+\frac{\alpha}{\rho}\tilde{w}\right)(1,t)=\frac{2\alpha\beta}{\rho}u_1'(1),$$

$$\frac{\partial}{\partial x}\left(\beta\tilde{v}-\frac{\alpha}{\rho}\tilde{w}\right)(0,t)=\frac{2\alpha\beta}{\rho}u_0'(t), \qquad \frac{\partial}{\partial x}\left(\beta\tilde{v}-\frac{\alpha}{\rho}\tilde{w}\right)(1,t)=0.$$

Thus the control problems for $\beta\tilde{v}+\alpha\tilde{w}/\rho$ and $\beta\tilde{v}-\alpha\tilde{w}/\rho$ are both of Neumann type and are *uncoupled.* Affirmative controllability results are then available from [20], [21], [24].

If we replace $u_0(t)$ (or $u_1(t)$ by 0 in the above, then $\beta\tilde{v}-\alpha\tilde{w}/\rho$ (or $\beta\tilde{v}+\alpha\tilde{w}/\rho$) will become completely uncontrollable and our original system must therefore be uncontrollable. This result at first seems to predict failure for the enterprise which we now undertake for the two-dimensional case.

4. Approximate boundary controllability. By a simple change of scale in the t variable, and renaming of the independent variables, we may assume that the system of interest is

$$\left.\begin{aligned} \frac{\partial^2 v}{\partial t^2}&=\frac{\partial^2 v}{\partial x^2}+\frac{\partial^2 v}{\partial y^2}, \qquad (4.1)\\ \frac{\partial^2 w}{\partial t^2}&=\frac{\partial^2 w}{\partial x^2}+\frac{\partial^2 w}{\partial y^2}, \qquad (4.2)\end{aligned}\right\} \quad t\geqq 0, \quad (x,y)\in\mathbf{R},$$

with boundary conditions

$$\left.\begin{aligned} \frac{\partial v}{\partial t}(x,y,t)&=\alpha u(x,y,t), \qquad (4.3)\\ \frac{\partial w}{\partial \nu}(x,y,t)&=\beta u(x,y,t), \qquad (4.4)\end{aligned}\right\} \quad t\geqq 0, \quad (x,y)\in\mathbf{B}=\partial\Omega.$$

We will not, in general, assume that $u(x,y,t)$ can be selected at will for all values of (x,y,t) shown. More on this later.

Because the system is time reversible, it is sufficient to analyze controllability in terms of control from the zero initial state

$$\left.\begin{aligned} v(x,y,0)=\frac{\partial v}{\partial t}(x,y,0)&=0, \qquad (4.5)\\ w(x,y,0)=\frac{\partial w}{\partial t}(x,y,0)&=0, \qquad (4.6)\end{aligned}\right\} \quad (x,y)\in\mathbf{R},$$

to a final state

$$\left.\begin{aligned} v(x,y,T)=v_0(x,y), \quad \frac{\partial v}{\partial t}(x,y,T)&=v_1(x,y), \qquad (4.7)\\ w(x,y,T)=w_0(x,y), \quad \frac{\partial w}{\partial t}(x,y,T)&=w_1(x,y), \qquad (4.8)\end{aligned}\right\} \quad (x,y)\in\mathbf{R}.$$

We have noted in § 2 that the $\|\cdot\|$-finite states are dense in the $|\cdot|$-finite states. In the present context this means that we can work with the Hilbert space of states v, $\partial v/\partial t$, w, $\partial w/\partial t$ with the inner product

$$\left(\left(v, \frac{\partial v}{\partial t}, w, \frac{\partial w}{\partial t}\right); \left(\tilde{v}, \frac{\partial \tilde{v}}{\partial t}, \tilde{w}, \frac{\partial \tilde{w}}{\partial t}\right)\right)$$

$$= \int\int_{\mathbf{R}} \left[\frac{\partial v}{\partial t}\frac{\partial \tilde{v}}{\partial t} + \frac{\partial w}{\partial t}\frac{\partial \tilde{w}}{\partial t} + \frac{\partial v}{\partial x}\frac{\partial \tilde{v}}{\partial x} + \frac{\partial w}{\partial x}\frac{\partial \tilde{w}}{\partial x} + \frac{\partial v}{\partial y}\frac{\partial \tilde{v}}{\partial y} + \frac{\partial w}{\partial y}\frac{\partial \tilde{w}}{\partial y}\right] dx\, dy, \tag{4.9}$$

the space which we refer to as $\hat{H}$. The norm is $\|\cdot\|$ (cf. (2.18)) with $\mu\varepsilon = 1$. As we have indicated, this is a dense subspace of H, the Hilbert space obtained by use of the norm $|\cdot|$ (cf. (2.17)).

The final states (4.7), (4.8) are not quite arbitrary in $\hat{H}$ if the control u is restricted so that its support is contained in a proper relatively closed subset $B_1 \subset B$. Since the condition

$$\frac{\partial v}{\partial t}(x, y, t) = \alpha u(x, y, t), \qquad (x, y) \in \mathbf{B}$$

applies, we may as well adjoin the additional condition

$$v_0(x, y) = 0, \qquad (x, y) \in \mathbf{B} - \mathbf{B}_1 \equiv \mathbf{B}_0. \tag{4.10}$$

The trace theorem [1], [19] assures us that this describes a closed subspace of $\hat{H}$, which we will call $\hat{H}_1$. The only restriction on $\hat{H}_1$ is (4.10); v_0 is permitted to have arbitrary values in $H^{1/2}(\mathbf{B}_1)$ and w_0, w_1 are unrestricted in $H^1(\mathbf{B})$, $H^0(\mathbf{R}) = L^2(\mathbf{R})$, respectively.

Let U be a given space of admissible control functions, about which we will shortly have more to say. For each control $u \in U$ we assume the existence of a unique solution v_u, w_u of (4.1)–(4.6) for $t \geqq 0$, $(x, y) \in \mathbf{R}$. Very general sufficient conditions for this to be the case are given in [19]. We define the reachable set at time T, $R(U, T)$, to be the set of all final states $v_u(x, y, T)$, $(\partial v_u/\partial t)(x, y, T)$, $w_u(x, y, T)$, $(\partial w_u/\partial t)(x, y, T)$ which may be realized in this way. The set $R(U, T)$ is a subspace of $\hat{H}_1$ if U is a linear space, which we will assume, and our system is approximately controllable in time T if $R(U, T)$ is dense in $\hat{H}_1$ (then $R(U, T)$ is also dense in H because $|\cdot|$ is a weaker norm than $\|\cdot\|$ and $\hat{H}_1$ is dense in H). Evidently $R(U, T)$ is dense in $\hat{H}_1$ just in case, given an arbitrary state $(\tilde{v}_0, \tilde{v}_1, \tilde{w}_0, \tilde{w}_1)$ in $\hat{H}_1$,

$$\left\{\left(\left(v_u(x, y, T), \frac{\partial v_u}{\partial t}(x, y, T), w_u(x, y, T), \frac{\partial w_u}{\partial t}(x, y, T)\right); (\tilde{v}_0, \tilde{v}_1, \tilde{w}_0, \tilde{w}_1)\right) = 0, u \in U\right\}$$

$$\Rightarrow (\tilde{v}_0, \tilde{v}_1, \tilde{w}_0, \tilde{w}_1) = 0. \tag{4.11}$$

Let $\tilde{v}(x, y, t)$, $\tilde{w}(x, y, t)$ be the unique solution of (4.1), (4.2) satisfying the terminal conditions at time T:

$$\tilde{v}(x, y, T) = \tilde{v}_0, \quad \frac{\partial \tilde{v}}{\partial t}(x, y, T) = \tilde{v}_1, \quad \tilde{w}(x, y, T) = \tilde{w}_0, \quad \frac{\partial \tilde{w}}{\partial t}(x, y, T) = \tilde{w}_1, \tag{4.12}$$

and the homogeneous boundary conditions

$$\left.\begin{aligned} &\frac{\partial \tilde{v}}{\partial t}(x, y, t) = 0, \qquad (4.13)\\ &\frac{\partial \tilde{w}}{\partial \nu}(x, y, t) = 0, \qquad (4.14)\end{aligned}\right\} \quad (x, y) \in \mathbf{B}, \quad t \geqq 0.$$

Computing the quantity

$$\frac{d}{dt}\left(\left(v_u(x, y, t), \frac{\partial v_u}{\partial t}(x, y, t), w_u(x, y, t), \frac{\partial w_u}{\partial t}(x, y, t)\right);\right.$$
$$\left.\left(\tilde{v}(x, y, t), \frac{\partial \tilde{v}}{\partial t}(x, y, t), \tilde{w}(x, y, t), \frac{\partial \tilde{w}}{\partial t}(x, y, t)\right)\right),$$

using familiar duality theorems involving the Laplacian and integrating from 0 to T (see [22], [23], [26] for details in the case of a single wave equation) we see that

$$\left(\left(v_u(x, y, T), \frac{\partial v_u}{\partial t}(x, y, T), w_u(x, y, T), \frac{\partial w_u}{\partial t}(x, y, T)\right); (\tilde{v}_0, \tilde{v}_1, \tilde{w}_0, \tilde{w}_1)\right)$$

$$= \int_0^T \int_{\mathbf{B}} \left[\frac{\partial \tilde{v}}{\partial t}(x, y, t)\frac{\partial v_u}{\partial \nu}(x, y, t) + \frac{\partial \tilde{v}}{\partial \nu}(x, y, t)\frac{\partial v_u}{\partial t}(x, y, t)\right. \tag{4.15}$$
$$\left. + \frac{\partial \tilde{w}}{\partial t}(x, y, t)\frac{\partial w_u}{\partial \nu}(x, y, t) + \frac{\partial \tilde{w}}{\partial \nu}(x, y, t)\frac{\partial w_u}{\partial t}(x, y, t)\right] ds\, dt.$$

Then using the boundary conditions (4.3), (4.4), (4.13), (4.14), we see that the above reduces to

$$\int_0^T \int_{\mathbf{B}} \left[\alpha \frac{\partial \tilde{v}}{\partial \nu}(x, y, t) + \beta \frac{\partial \tilde{w}}{\partial t}(x, y, t)\right] u(x, y, t)\, ds\, dt. \tag{4.16}$$

If, as discussed above, we suppose that $\mathbf{B}$ has the disjoint decomposition

$$\mathbf{B} = \mathbf{B}_0 \cup \mathbf{B}_1,$$

with $\mathbf{B}_1$ relatively open in $\mathbf{B}$, and that $u(x, y, t) \equiv 0$, $(x, y) \in \mathbf{B}_0$ while on $\mathbf{B}_1$ u is unrestricted save for the specification of the admissible space (e.g., we might take

$$U = C(\mathbf{B}_1 \times [0, T]), \qquad U = L^2(\mathbf{B}_1 \times [0, T]), \tag{4.17}$$

or any of many other possibilities), and if we suppose the first equation in (4.11) to hold, we conclude that (4.16) vanishes for all $u \in U$. We know from the trace theorem ([1], [19]) that the partial derivatives

$$\frac{\partial \tilde{v}}{\partial t}, \frac{\partial \tilde{v}}{\partial \nu}, \frac{\partial \tilde{w}}{\partial t}, \frac{\partial \tilde{w}}{\partial \nu},$$

restricted to $\mathbf{B}$, all lie in $H^{1/2}(\mathbf{B})$ for $t \in [0, T]$ and vary, with respect to the norm in that space, continuously with respect to t, i.e. they lie in $C(H^{1/2}(\mathbf{B}); [0, T])$. We suppose, as is the case for (4.17), e.g., that U includes a total subspace of the dual space of $C(H^{1/2}(\mathbf{B}_1); [0, T])$. Then the fact that (4.17) is zero for all $u \in U$ implies

$$\alpha \frac{\partial \tilde{v}}{\partial \nu}(x, y, t) + \beta \frac{\partial \tilde{w}}{\partial t}(x, y, t) = 0, \qquad (x, y) \in \mathbf{B}_1, \quad t \in [0, T]. \tag{4.18}$$

We also have (cf. (4.13), (4.14))

$$\frac{\partial \tilde{v}}{\partial t}(x, y, t) = 0, \quad \frac{\partial \tilde{w}}{\partial \nu}(x, y, t) = 0, \quad (x, y) \in \mathbf{B}_1, \quad t \in [0, T]. \tag{4.19}$$

The boundary values of $\tilde{v}$ and $\tilde{w}$ are therefore *overspecified* on $\mathbf{B}_1 \times [0, T]$. The proof of approximate controllability, where it can be carried through, depends upon being able to use this overspecification to show that

$$\tilde{v}(x, y, t) \equiv 0, \quad \tilde{w}(x, y, t) \equiv 0, \quad (x, y) \in \mathbf{R}, \quad t \in [0, T],$$

and therefore to conclude that the implication (4.11) is indeed valid so that $R(U, T)$ is dense in $\hat{H}_1$ and hence in H. We carry this argument out for the case in which $\mathbf{R}$ is a rectangle and $\mathbf{B}_1$ is one of its sides in § 5.

Following the development in [6], it may be seen that our system is exactly controllable in $\hat{H}_1$, using the control space $U = L^2(\mathbf{B}_1 \times [0, T])$, just in case

$$\left\| \alpha \frac{\partial \tilde{v}}{\partial \nu} + \beta \frac{\partial \tilde{w}}{\partial t} \right\|_{L^2(\mathbf{B}_1 \times [0,T])} \geqq K \| (\tilde{v}_0, \tilde{v}_1, \tilde{w}_0, \tilde{w}_1) \|_{\hat{H}} \tag{4.20}$$

for some $K > 0$. In general this is a very difficult result to obtain but we are able to obtain exact controllability, by other means, for the case where $\mathbf{R}$ is a disc in R^2 and $\mathbf{B}_1 = \mathbf{B}$ is its boundary, a circle. This result is developed in § 6 where it will be seen that it is heavily dependent on certain properties of the Bessel functions.

5. The case $\mathbf{R}$ = a rectangle, $\mathbf{B}_1$ = one side. The work here can be carried out for a rectangle with arbitrary dimensions, but all essential ideas are contained in the notationally simpler case

$$\mathbf{R} = \{(x, y) \mid 0 \leqq x \leqq \pi, 0 \leqq y \leqq \pi\}$$

to which attention is restricted henceforth. We will assume that $\mathbf{B}_1$, the portion of the boundary on which control is exercised, is one side of $\mathbf{R}$; without loss of generality it is the set

$$\mathbf{B}_1 = \{(\pi, y) \mid 0 \leqq y \leqq \pi\}. \tag{5.1}$$

We consider then $\tilde{v}$, $\tilde{w}$ satisfying (4.1), (4.2) in $\mathbf{R} \times [0, T]$ for some $T > 0$, and also satisfying boundary conditions

$$\frac{\partial \tilde{v}}{\partial t}(x, y, t) = 0, \quad \frac{\partial \tilde{w}}{\partial \nu}(x, y, t) = 0, \quad (x; y) \in \mathbf{B} = \partial \mathbf{R}, \quad t \in [0, T], \tag{5.2}$$

$$\begin{aligned} &\alpha \frac{\partial \tilde{v}}{\partial \nu}(\pi, y, t) + \beta \frac{\partial \tilde{w}}{\partial t}(\pi, y, t) \\ &\quad = -\alpha \frac{\partial \tilde{v}}{\partial x}(\pi, y, t) + \beta \frac{\partial \tilde{w}}{\partial t}(\pi, y, t) = 0, \qquad 0 \leqq y \leqq \pi, \quad t \in [0, T]. \end{aligned} \tag{5.3}$$

We may assume without loss of generality, since the wave equation is time reversible with either Dirichlet or Neumann boundary conditions, that $\tilde{v}$ and $\tilde{w}$ are extended to satisfy (4.1), (4.2) on $-\infty < t < \infty$ and that the boundary conditions (5.2) hold for $(x, y) \in \mathbf{B}$, $t \in (-\infty, \infty)$. We may not assume that the boundary condition (5.3) is applicable beyond $[0, T]$, however, if controls are restricted to have support in $\mathbf{B}_1 \times [0, T]$. Let $\delta > 0$ and let $s(t)$ be an arbitrary function in $C^\infty(-\infty, \infty)$ with support in

$(-\delta, \delta)$. Define

$$\hat{v}(x, y, t) = \int_{-\infty}^{\infty} s(t-\tau)\tilde{v}(x, y, \tau)\, d\tau, \tag{5.4}$$

$$\hat{w}(x, y, t) = \int_{-\infty}^{\infty} s(t-\tau)\tilde{w}(x, y, \tau)\, d\tau. \tag{5.5}$$

Then $\hat{v}$, $\hat{w}$ are solutions of the wave equations (4.1), (4.2) satisfying boundary conditions

$$\frac{\partial \hat{v}}{\partial t}(x, y, t) = 0, \quad \frac{\partial \hat{w}}{\partial \nu}(x, y, t) = 0, \quad (x, y) \in \mathbf{B} = \partial\mathbf{R}, \quad -\infty < t < \infty, \tag{5.6}$$

while

$$\alpha\frac{\partial \hat{v}}{\partial x}(\pi, y, t) + \beta\frac{\partial \hat{w}}{\partial t}(\pi, y, t) = 0, \quad 0 \leqq y \leqq \pi, \quad t \in [\delta, T-\delta]. \tag{5.7}$$

Moreover, it can be shown that $\hat{v}$, $\hat{w}$ are of class C^∞ for $(x, y) \in \mathbf{R}$, $-\infty < t < \infty$. If we can show $\hat{v} \equiv 0$, $\hat{w} \equiv 0$ for any such choice of s, then $\tilde{v} \equiv 0$, $\tilde{w} \equiv 0$.

Let us define, for $(x, y) \in \mathbf{R}$, $-\infty < t < \infty$,

$$\phi(x, y, t) = \alpha\frac{\partial \hat{v}}{\partial x}(x, y, t) + \beta\frac{\partial \hat{w}}{\partial t}(x, y, t). \tag{5.8}$$

From (5.7) we have

$$\phi(\pi, y, t) = 0, \quad 0 \leqq y \leqq \pi, \quad t \in [\delta, T-\delta]. \tag{5.9}$$

Since α and β are constants, we have

$$\frac{\partial^2 \phi}{\partial t^2} = \frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2}, \qquad (x, y) \in \mathbf{R}, \quad -\infty < t < \infty. \tag{5.10}$$

Let us note that, since $\hat{v}$ satisfies the wave equation in $\mathbf{R} \cup \mathbf{B}$,

$$\begin{aligned} &\alpha\frac{\partial^2 \hat{v}}{\partial t^2}(x, y, t) + \beta\frac{\partial^2 \hat{w}}{\partial t\, \partial x}(x, y, t) \\ &\quad = \alpha\left[\frac{\partial^2 \hat{v}}{\partial x^2}(x, y, t) + \frac{\partial^2 \hat{v}}{\partial y^2}(x, y, t)\right] + \beta\frac{\partial^2 \hat{w}}{\partial t\, \partial x}(x, y, t). \end{aligned} \tag{5.11}$$

Setting $x = \pi$ in (5.11) and differentiating the identities in (5.6) with respect to t, we see that the left-hand side vanishes. Then, comparing (5.11) with (5.8)

$$\frac{\partial \phi}{\partial x}(\pi, y, t) = -\alpha\frac{\partial^2 \hat{v}}{\partial y^2}(\pi, y, t) \equiv \alpha(y), \qquad 0 \leqq y \leqq \pi, \quad \delta \leqq t \leqq T-\delta, \tag{5.12}$$

the last identity being valid as a consequence of the first condition in (5.6).

The two conditions, (5.8) and (5.12), satisfied by ϕ at the boundary $x = \pi$ enable us to use Holmgren's uniqueness theorem (see [5] or [13], e.g.) in much the same way as it was used in the proof of the approximate controllability of the wave equation in [22], [23] to see that if

$$T > 2 + 2\delta \tag{5.13}$$

then ϕ must be independent of t for $1 + \delta \leqq t \leqq T - 1 - \delta$, i.e.

$$\phi(x, y, t) = \phi(x, y), \qquad (x, y) \in \mathbf{R}, \quad 1 + \delta \leqq t \leqq T - 1 - \delta. \tag{5.14}$$

Because $\hat{v}$ and $\hat{w}$ satisfy the wave equation in $\mathbf{R}$ with the homogeneous boundary conditions (5.6), and are of class C^∞ in $\mathbf{R}\cup\mathbf{B}$, we have C^∞-convergent expansions

$$\hat{v}(x, y, t) = \hat{v}_0(x, y) + \sum_{k=1}^{\infty}\sum_{j=1}^{\infty} (v_{kj}\, e^{i\omega_{kj}t} + \bar{v}_{kj}\, e^{-\omega_{kj}t}) \sin kx \sin jy, \tag{5.15}$$

$$\hat{w}(x, y, t) = \hat{w}_0 + \sum_{k=1}^{\infty}\sum_{j=1} (w_{kj}\, e^{i\omega_{kj}t} + \bar{w}_{kj}\, e^{-i\omega_{kj}t}) \cos kx \cos jy, \tag{5.16}$$

where

$$w_{kj} = \sqrt{k^2+j^2}, \tag{5.17}$$

$\hat{v}_0(x, y)$ is a C^∞ function in $\mathbf{R}\cup\mathbf{B}$ such that (cf. (4.10))

$$\hat{v}_0(x, y) = 0, \qquad (x, y)\in \mathbf{B}-\{(\pi, y)\,|\,0\leqq y\leqq \pi\} \tag{5.18}$$

and $\hat{w}_0$ is a constant. Then, from (5.8),

$$\begin{aligned}\phi(x, y, t) - \alpha\frac{\partial \hat{v}_0(x, y)}{\partial x} &= \sum_{k=1}^{\infty} \cos kx \Bigg[\sum_{j=1}^{\infty} (\alpha k v_{kj} \sin jy + i\beta\omega_{kj} w_{kj} \cos jy)\, e^{i\omega_{kj}t} \\ &\qquad + \sum_{j=1}^{\infty} (\alpha k \bar{v}_{kj} \sin jy - i\beta\omega_{kj}\bar{w}_{kj} \cos jy)\, e^{-i\omega_{kj}t}\Bigg],\end{aligned} \tag{5.19}$$

still C^∞-convergent for $(x, y)\in\mathbf{R}\cup\mathbf{B}$, $-\infty<t<\infty$. Noting (5.14), we see that the left-hand side takes the form

$$\phi(x, y, t) - \alpha\frac{\partial \hat{v}_0(x, y)}{\partial x} = \phi(x, y) - \alpha\frac{\partial \hat{v}_0(x, y)}{\partial x} \equiv \hat{\phi}(x, y), \qquad 1+\delta\leqq t\leqq T-1-\delta. \tag{5.20}$$

We now strengthen (5.13) to

$$T > 4+2\delta \tag{5.21}$$

and we see that the time interval in (5.14), (5.20) has length >2, i.e.

$$T-1-\delta-(1+\delta) = T-(2+2\delta) > 2.$$

Since the functions $\sqrt{2/\pi}\cos kx$ are orthonormal on $0\leqq x\leqq\pi$, we conclude from (5.19), (5.20) that for $k = 1, 2, 3, \cdots$

$$\begin{aligned}&\sum_{j=1}^{\infty} (\alpha k v_{kj} \sin jy + i\beta\omega_{kj} w_{kj} \cos jy)\, e^{i\omega_{kj}t} \\ &\qquad + \sum_{j=1}^{\infty} (\alpha k \bar{v}_{kj} \sin jy - i\beta\omega_{kj}\bar{w}_{kj} \cos jy)\, e^{-i\omega_{kj}t} \\ &= \frac{2}{\pi}\int_0^{\pi} \hat{\phi}(x, y) \cos kx\, dx \equiv \Phi_k(y), \qquad 1+\delta\leqq t\leqq T-1-\delta.\end{aligned} \tag{5.22}$$

Classical results of Levinson and Schwartz ([17], [27]), which have frequently been used in control studies of this type (see, e.g., [12], [21]), can now be used to show that for each fixed k, the exponential functions

$$\exp(\pm i\omega_{kj}t) = \exp(\pm\sqrt{k^2+j^2}\,t), \qquad j = 1, 2, 3, \cdots,$$

together with the constant function 1 are strongly independent in $L^2(I)$ for any t-interval I of length >2. This clearly contradicts (5.22) unless we have

$$\Phi_k(y) \equiv 0, \qquad 0 \leqq y \leqq \pi \tag{5.23}$$

and

$$\alpha k v_{kj} \sin jy + i\beta\omega_{kj} w_{kj} \cos jy = 0, \qquad 0 \leqq y \leqq \pi, \quad j = 1, 2, 3, \cdots.$$

Since, for each j, $\sin jy$ and $\cos jy$ are independent on $0 \leqq y \leqq \pi$ and since none of α, k, β, ω_{kj} are zero, we conclude that

$$v_{kj} = 0, \quad w_{kj} = 0, \quad k = 1, 2, 3, \cdots, \quad j = 1, 2, 3, \cdots. \tag{5.24}$$

Since (5.22), (5.23) show that

$$\hat{\phi}(x, y) = \sum_{k=1}^{\infty} \Phi_k(y) \cos kx = 0,$$

(5.19) gives

$$\phi(x, y, t) = \phi(x, y) = \alpha \frac{\partial \hat{v}_0(x, y)}{\partial x}, \qquad (x, y) \in \mathbf{R}, \quad 1 + \delta \leqq t \leqq T - 1 - \delta. \tag{5.25}$$

Noting (5.15) and (5.16) and the fact that $\hat{v}(0, y, t) \equiv 0$, we conclude from (5.23) that

$$\left.\begin{aligned} \hat{v}(x, y, t) &\equiv \hat{v}_0(x, y) \\ \hat{w}(x, y, t) &\equiv \hat{w}_0 \end{aligned}\right\} \qquad 1 + \delta \leqq t \leqq T - 1 - \delta. \tag{5.26}$$

Since $v(x, y, t) \equiv v_0(x, y)$ is a solution of the wave equation with (cf. (5.18))

$$\hat{v}_0(x, y) = 0, \qquad (x, y) \in \mathbf{B} - \{(\pi, y) \mid 0 \leqq y \leqq \pi\}$$

it must in fact be a solution of Laplace's equation there. Then we compute

$$\begin{aligned} &\int_{\mathbf{R}} \left[\left(\frac{\partial \hat{v}_0}{\partial x}(x, y) \right)^2 + \left(\frac{\partial \hat{v}_0}{\partial y}(x, y) \right)^2 + \hat{v}_0(x, y) \left(\frac{\partial^2 \hat{v}_0}{\partial x^2}(x, y) + \frac{\partial^2 \hat{v}_0}{\partial y^2}(x, y) \right) \right] dx\, dy \\ &= \int_{\mathbf{R}} \operatorname{div} (\hat{v}_0(x, y) \operatorname{grad} \hat{v}_0(x, y))\, dx\, dy \\ &= \int_{\mathbf{B}} \hat{v}_0(x, y) \operatorname{grad} \hat{v}_0(x, y) \cdot \nu(x, y)\, ds = \int_0^{\pi} \hat{v}_0(\pi, y) \frac{\partial \hat{v}_0}{\partial x}(\pi, y)\, dy. \end{aligned} \tag{5.27}$$

Combining (5.9) and (5.25) with the fact that $\hat{v}_0$ satisfies Laplace's equation, we conclude from (5.27) that

$$\int_{\mathbf{R}} \left[\left(\frac{\partial \hat{v}_0}{\partial x}(x, y) \right)^2 + \left(\frac{\partial \hat{v}_0}{\partial y}(x, y) \right)^2 \right] dx\, dy = 0$$

and this, together with (5.18), implies

$$\hat{v}_0(x, y) \equiv 0. \tag{5.28}$$

Combining (5.26) and (5.28), we conclude that

$$\left.\begin{aligned} \hat{v}(x, t, t) &\equiv 0 \\ \hat{w}(x, y, t) &\equiv \hat{w}_0 \end{aligned}\right\} \qquad (x, y) \in \mathbf{R}, \quad -\infty < t < \infty, \tag{5.29}$$

the result for $-\infty < t < \infty$ being an immediate consequence of the result for $1 + \delta \leqq t \leqq T - 1 - \delta$. Since this is true for every $\delta > 0$ and every $s(t)$ in (5.4), (5.5), we conclude

that a comparable result obtains for $\hat{v}$, $\hat{w}$ in (4.11), (5.2), (5.3). It follows (since $w=$ constant is a zero state in $\hat{H}$ and in H) that (cf. (4.9) ff.)

$$\|(\tilde{v}_0, \tilde{v}_1, \tilde{w}_0, \tilde{w}_1)\|_{\hat{H}} = \|(\tilde{v}_0, \tilde{v}_1, \tilde{w}_0, \tilde{w}_1)\|_H = 0$$

and, from the discussion in § 4, the approximate controllability result follows for $u \in L^2(\mathbf{B}_1 \times [0, T])$, $T > 4$. The comparable result for a single wave equation appears in [22], [23] where a control time $T > 2$ (in the same geometrical context) is seen to be required. Thus the "critical control time" for the combined Dirichlet-Neumann problem, which we have seen to be equivalent to the Maxwell system under the restrictions which we have imposed, is exactly twice the critical control time for the (single) wave equation.

The approximate controllability just established cannot be strengthened to exact controllability when the controlled portion of the boundary, $\mathbf{B}_1$, is just one side of the rectangle. This follows from the result [8] of Fattorini who has shown that only a very weak form of approximate controllability obtains for the (single) wave equation under these circumstances. It is known from [2], [3], [25] that exact controllability of finite energy states for the (single) wave equation is obtained with control on at least two sides of the rectangle. The question as to whether this remains the case for Maxwell's equations, with the doubled time interval again, is not completely clear at this time but seems likely to be answered in the affirmative.

The difficulty in such work, which would be approached in much the same manner as Graham and the present author treated the wave equation in [12], lies in the imperfect state of the literature as regards the eigenfunctions of the Maxwell system in three-dimensional regions. No definitive description of the complete set of eigenfunctions appears to exist for such a basic region as the three-dimensional ball. The nature of the eigenfunctions in the case of a rectangle can be sorted out with some difficulty and work on the control problem is under way in this direction. Results for general regions, comparable to those in [2], [3], [15], [16], [25] are hampered by the lack of a general energy decay result similar to what is available for the (single) wave equation. We see in the next section, however, that we can obtain exact controllability results for the two-dimensional problem corresponding to the circular cylinder.

6. Some exact controllability results in the case of a circular cylinder. We consider now the case $\Omega = \mathbf{R} \times (-\infty, \infty)$ with

$$\mathbf{R} = \{(x, y) \mid x^2 + y^2 < 1\},$$
$$\mathbf{B} = \partial\mathbf{R} = \{(x, y) \mid x^2 + y^2 = 1\}.$$

With introduction of the usual polar coordinates r, θ, the equations (4.1), (4.2) now become

$$\frac{\partial^2 v}{\partial t^2} = \frac{\partial^2 v}{\partial r^2} + \frac{1}{r}\frac{\partial v}{\partial r} + \frac{1}{r^2}\frac{\partial^2 v}{\partial \theta^2}, \tag{6.1}$$

$$\frac{\partial^2 w}{\partial t^2} = \frac{\partial^2 w}{\partial r^2} + \frac{1}{r}\frac{\partial w}{\partial r} + \frac{1}{r^2}\frac{\partial^2 w}{\partial \theta^2} \tag{6.2}$$

and the boundary conditions (4.3), (4.4) are transformed to

$$\frac{\partial v}{\partial t}(1, \theta, t) = \alpha u(\theta, t), \tag{6.3}$$

$$\frac{\partial w}{\partial r}(1, \theta, t) = \beta u(\theta, t). \tag{6.4}$$

Writing

$$v(r,\theta,t)=\sum_{k=-\infty}^{\infty} v_k(r,t)\,e^{ik\theta}, \qquad v_{-k}=\bar{v}_k, \tag{6.5}$$

$$w(r,\theta,t)=\sum_{k=-\infty}^{\infty} w_k(r,t)\,e^{ik\theta}, \qquad w_{-k}=\bar{w}_k, \tag{6.6}$$

$$u(\theta,t)=\sum_{k=-\infty}^{\infty} u_k(t)\,e^{ik\theta}, \tag{6.7}$$

we arrive at an infinite collection of control problems in the single space dimension, r:

$$\frac{\partial^2 v_k}{\partial t^2}=\frac{\partial^2 v_k}{\partial r^2}+\frac{1}{r}\frac{\partial v_k}{\partial r}-\frac{k^2}{r^2}v_k=0, \qquad -\infty<k<\infty, \tag{6.8}$$

$$\frac{\partial^2 w_k}{\partial t^2}=\frac{\partial^2 w_k}{\partial r^2}+\frac{1}{r}\frac{\partial w_k}{\partial r}-\frac{k^2}{r^2}w_k=0, \qquad -\infty<k<\infty, \tag{6.9}$$

$$\frac{\partial v_k}{\partial t}(1,t)=\alpha u_k(t), \qquad -\infty<k<\infty, \tag{6.10}$$

$$\frac{\partial w_k}{\partial r}(1,t)=\beta u_k(t), \qquad -\infty<k<\infty. \tag{6.11}$$

We will first treat the equation (4.1) with the boundary condition (4.3) which, as we have seen, reduces to the set of problems (6.8), (6.10), $-\infty<k<\infty$. With

$$z(r,\theta,t)=\sum_{k=-\infty}^{\infty} z_k(r,t)\,e^{ik\theta}=\sum_{k=-\infty}^{\infty}\frac{\partial v_k(r,t)}{\partial t}\,e^{ik\theta}=\frac{\partial v}{\partial t}(r,\theta,t)$$

we have the equivalent first order systems

$$\frac{\partial}{\partial t}\begin{pmatrix} v_k(r,t)\\ z_k(r,t)\end{pmatrix}=\begin{pmatrix} 0 & I\\ L_{|k|} & 0\end{pmatrix}\begin{pmatrix} v_k(r,t)\\ z_k(r,t)\end{pmatrix}=L_{|k|}\begin{pmatrix} v_k(r,t)\\ z_k(r,t)\end{pmatrix} \tag{6.12}$$

where $L_{|k|}$ is the differential operator on the right-hand side of (6.8). The boundary conditions (6.10) become

$$z_k(1,t)=\alpha u_k(t), \qquad -\infty<k<\infty. \tag{6.13}$$

The eigenvalues of the operator $L_{|k|}$ with the corresponding homogeneous boundary condition

$$z_k(1,t)=0 \tag{6.14}$$

are

$$0,\ \pm i\omega_{|k|,l}, \qquad l=1,2,3,\cdots,$$

where $\omega_{|k|,l}$ is the lth positive zero of the Bessel function $J_{|k|}(r)$ of order $|k|$. The corresponding vector eigenfunctions are

$$\begin{pmatrix}\phi_{|k|,0}(r)\\ 0\end{pmatrix},\begin{pmatrix}\phi_{|k|,l}(r)\\ \pm i\omega_{|k|,l}\phi_{|k|,l}(r)\end{pmatrix}, \qquad -\infty<k<\infty, \quad l=1,2,3,\cdots,$$

where

$$\begin{aligned}&\phi_{|k|,0}(r)=A_{|k|,0}r^{|k|}, \qquad -\infty<k<\infty,\\ &\phi_{|k|,l}(r)=A_{|k|,l}J_{|k|}(\omega_{|k|,l}r), \qquad -\infty<k<\infty, \quad l=1,2,3,\cdots.\end{aligned} \tag{6.15}$$

The normalization coefficients $A_{|k|,0}$, $A_{|k|,l}$ are chosen so that

$$\int_0^1 r|\phi_{|k|,0}(r)|^2\,dr=\frac{1}{2\pi},\qquad \int_0^1 r|\phi_{|k|,l}(r)|^2\,dr=\frac{1}{2\pi},\qquad l=1,2,3,\cdots. \tag{6.16}$$

Thus

$$A_{|k|,0}=\sqrt{\frac{|k|+1}{\pi}},\qquad -\infty<k<\infty, \tag{6.17}$$

while, as may be seen from [5], e.g.

$$A_{|k|,l}=\frac{\omega_{|k|,l}}{\sqrt{\pi}J_{|k|,l}(\omega_{|k|,l})}. \tag{6.18}$$

The state space in which we wish to work, for the present at least, is (cf. (2.18))

$$\tilde{H}=\left\{\begin{pmatrix}v\\z\end{pmatrix}\,\middle|\, v\in H^1(\mathbf{R}),\ z\in L^2(\mathbf{R})\right\}$$

with the inner product

$$\left(\begin{pmatrix}v_1\\z_1\end{pmatrix},\begin{pmatrix}v_2\\z_2\end{pmatrix}\right)=\int_{\mathbf{R}}(\nabla v_1\cdot\overline{\nabla v_2}+z_1\overline{z_2})\,dx\,dy$$

and associated norm. Since the $\phi_{|k|,l}$ satisfy the homogeneous boundary condition (6.14), one easily sees that

$$\begin{aligned}\left\|\begin{pmatrix}\phi_{|k|,0}\,e^{ik\theta}\\0\end{pmatrix}\right\|_{\tilde{H}}^2&=-\int_{\mathbf{R}}\phi_{|k|,0}\,e^{ik\theta}\overline{\Delta(\phi_{|k|,0}\,e^{ik\theta})}\,dx\,dy\\&\quad+\int_{\partial\mathbf{R}}\phi_{|k|,0}\,e^{ik\theta}\overline{\frac{\partial\phi_{|k|,0}}{\partial r}}\,e^{-ik\theta}\,d\theta+\frac{|k|+1}{\pi}\int_0^{2\pi}|k|\,d\theta\\&=2|k|(|k|+1),\qquad -\infty<k<\infty,\end{aligned} \tag{6.19}$$

while

$$\begin{aligned}\left\|\begin{pmatrix}\phi_{|k|,l}\\\pm i\omega_{|k|,l}\phi_{|k|,l}\end{pmatrix}\right\|_{\tilde{H}}^2&=\lambda_{|k|,l}\int_{\mathbf{R}}|\varepsilon_{|k|,l}|^2\,dx\,dy\\&\quad+\int_{\mathbf{R}}\nabla\phi_{|k|,l}\cdot\overline{\nabla\phi_{|k|,l}}\,dx\,dy=2\lambda_{|k|,l}\int_{\mathbf{R}}|\phi_{|k|,l}|^2\,dx\,dy=2\lambda_{|k|,l},\end{aligned} \tag{6.20}$$

where

$$\lambda_{|k|,l}=(\omega_{|k|,l})^2,\qquad -\infty<k<\infty,\quad l=1,2,3,\cdots.$$

The state $\binom{\phi_{0,0}}{0}$ has zero norm in $\tilde{H}$. Nevertheless we will not neglect this component.

If v, $\tilde{v}$ both satisfy the wave equation and (6.3), (4.13) on ∂R with initial state (4.5) for v we have (cf. (4.16))

$$\left(\begin{pmatrix}v(\cdot,\cdot,T)\\z(\cdot,\cdot,T)\end{pmatrix},\begin{pmatrix}\tilde{v}(\cdot,\cdot,T)\\\tilde{z}(\cdot,\cdot,T)\end{pmatrix}\right)_{\tilde{H}}=\alpha\int_0^T\int_{\mathbf{B}=\partial\mathbf{R}}u(x,y,t)\overline{\frac{\partial\tilde{v}}{\partial\nu}}(x,y,t)\,ds\,dt. \tag{6.21}$$

It may be shown that this result is valid for all u for which the solution (in the generalized sense) v lies in $\tilde{H}$ and varies continuously with respect to t. This class of controls u is discussed in [19] and is known to include, e.g., $u\in C([0,T];H^{1/2}(\mathbf{B}))$.

If we assume $\binom{v}{z}$ given by the $\tilde{H}$-convergent series

$$\begin{pmatrix} v(\cdot,\cdot,t) \\ z(\cdot,\cdot,t) \end{pmatrix} = \sum_{k=-\infty}^{\infty} v_{k,0}(t)(\phi_{|k|,0}\, e^{ik\theta}) + \sum_{k=-\infty}^{\infty}\sum_{l=1}^{\infty}\left[v_{k,l}^{+}(t)\begin{pmatrix} \phi_{|k|,l}\, e^{ik\theta} \\ i\omega_{|k|,l}\phi_{|k|,l}\, e^{ik\theta} \end{pmatrix} + v_{k,l}^{-}(t)\begin{pmatrix} \phi_{|k|,l}\, e^{ik\theta} \\ -i\omega_{|k|,l}\phi_{|k|,l}\, e^{ik\theta} \end{pmatrix}\right]$$

and successively let

$$\begin{pmatrix} \tilde{v}(\cdot,\cdot,t) \\ \tilde{z}(\cdot,\cdot,t) \end{pmatrix} = \begin{pmatrix} \phi_{|k|,0}\, e^{ik\theta} \\ 0 \end{pmatrix}, \exp(i\omega_{|k|,l}(t-T))\begin{pmatrix} \phi_{|k|,l}\, e^{ik\theta} \\ i\omega_{|k|,l}\phi_{|k|,l}\, e^{ik\theta} \end{pmatrix}, \exp(-i\omega_{|k|,l}(t-T))\begin{pmatrix} \phi_{|k|,l}\, e^{ik\theta} \\ -i\omega_{|k|,l}\phi_{|k|,l}\, e^{ik\theta} \end{pmatrix}, \quad -\infty < k < \infty, \quad l = 1, 2, 3, \cdots \tag{6.22}$$

for $T > 0$ we arrive at the equations

$$\begin{aligned} 2|k|(|k|+1)v_{k,0}(T) &= \alpha \int_0^T \int_0^{2\pi} u(\theta, t)\overline{\frac{\partial \phi_{|k|,0}}{\partial r}(1)}\, e^{-ik\theta}\, d\theta\, dt \\ &= 2\pi\alpha \overline{\frac{\partial \phi_{|k|,0}}{\partial r}(1)} \int_0^T u_k(t)\, dt, \end{aligned} \tag{6.23}$$

$$\begin{aligned} 2\lambda_{|k|,l}v_{k,l}^{+}(T) &= \alpha \int_0^T \int_0^{2\pi} u(\theta, t)\exp(i\omega_{|k|,l}(T-t))\overline{\frac{\partial \phi_{|k|,l}}{\partial r}(1)}\, e^{-ik\theta}\, d\theta\, dt \\ &= 2\pi\alpha \overline{\frac{\partial \phi_{|k|,l}}{\partial r}(1)} \int_0^T \exp(i\omega_{|k|,l}(T-t))u_k(t)\, dt, \end{aligned} \tag{6.24}$$

$$\begin{aligned} 2\lambda_{|k|,l}v_{k,l}^{-}(T) &= \alpha \int_0^T \int_0^{2\pi} u(\theta, t)\exp(-i\omega_{|k|,l}(T-t))\overline{\frac{\partial \phi_{|k|,l}}{\partial r}(1)}\, e^{-ik\theta}\, d\theta\, dt \\ &= 2\pi\alpha \overline{\frac{\partial \phi_{|k|,l}}{\partial r}(1)} \int_0^T \exp(-i\omega_{|k|,l}(T-t))u_k(t)\, dt. \end{aligned} \tag{6.25}$$

Thus the Dirichlet boundary control problem for (6.8), (6.10) is reduced to a moment problem (6.23), (6.24), (6.25) for which $u_k(t)$ must be a solution. We proceed in much the same way with the Neumann boundary control problem for (6.9), (6.11). We let

$$\zeta(r, \theta, t) = \sum_{k=-\infty}^{\infty} \zeta_k(r, t)\, e^{ik\theta} = \sum_{k=-\infty}^{\infty} \frac{\partial w_k(r, t)}{\partial t}\, e^{ik\theta} = \frac{\partial w}{\partial t}(r, \theta, t)$$

and obtain, in place of (6.12),

$$\frac{\partial}{\partial t}\begin{pmatrix} w_k(r, t) \\ \zeta_k(r, t) \end{pmatrix} = \begin{pmatrix} 0 & I \\ M_{|k|} & 0 \end{pmatrix}\begin{pmatrix} w_k(r, t) \\ \zeta_k(r, t) \end{pmatrix} = M_{|k|}\begin{pmatrix} v_k(r, t) \\ z_k(r, t) \end{pmatrix}. \tag{6.26}$$

The boundary conditions are now

$$\frac{\partial w_k}{\partial r}(1, t) = \beta u_k(t), \qquad -\infty < k < \infty.$$

The eigenvalues of $M_{|k|}$ with the corresponding homogeneous boundary condition

$$\frac{\partial w_k}{\partial r}(1, t) = 0$$

are, for $k=0$,

$$0, \quad \pm i\nu_{0,l} \quad l=1,2,3,\cdots,$$

where $\nu_{0,l}$ is the lth zero of the differentiated Bessel function, $j_0'(r)$, of order 0, and, for $k\neq 0$,

$$\pm i\nu_{|k|,l}, \qquad l=1,2,3,\cdots,$$

where $\nu_{k,l}$ is the lth zero of $J_k'(r)$. In the case $k=0$ the eigenvalue 0 has double multiplicity. The special solutions taking the place of (6.22) in this case are

$$\begin{pmatrix}\tilde{w}(\cdot,\cdot,t)\\ \tilde{\zeta}(\cdot,\cdot,t)\end{pmatrix}=\begin{pmatrix}\psi_{00}\\ 0\end{pmatrix},\begin{pmatrix}(t-T)\psi_{00}\\ \psi_{00}\end{pmatrix} \tag{6.27}$$

where ψ_{00} is such that (cf. (6.16))

$$\int_0^1 r\psi_{00}^2\,dr=\frac{1}{2\pi}, \quad \text{i.e.,} \quad \psi_{00}=\frac{1}{\sqrt{\pi}}.$$

In all of the other cases the vector eigenfunctions take the form

$$\begin{pmatrix}\psi_{|k|,l}(r)\\ \pm i\nu_{|k|,l}\psi_{|k|,l}(r)\end{pmatrix}, \qquad -\infty<k<\infty, \quad l=1,2,3,\cdots$$

where

$$\psi_{|k|,l}(r)=B_{|k|,l}J_{|k|}(\nu_{|k|,l}r), \qquad -\infty<k<\infty, \quad l=1,2,3,\cdots,$$

the normalization coefficients

$$B_{|k|,l}=\frac{\nu_{|k|,l}}{\sqrt{\pi}\,(\mu_{|k|,l}-k^2)^{1/2}J_{|k|}(\nu_{|k|,l})} \tag{6.28}$$

selected so that

$$\int_0^{\mathbf{R}} r|\psi_{|k|,l}(r)|^2\,dr=\frac{1}{2\pi}.$$

The corresponding special solutions of the homogeneous equation are

$$\begin{aligned}\begin{pmatrix}\tilde{w}(\cdot,\cdot,t)\\ \tilde{\zeta}(\cdot,\cdot,t)\end{pmatrix}&=\exp\left(i\nu_{|k|,l}(t-T)\right)\begin{pmatrix}\psi_{|k|,l}\,e^{ik\theta}\\ i\nu_{|k|,l}\psi_{|k|,l}\,e^{ik\theta}\end{pmatrix},\\ &\exp\left(-i\nu_{|k|,l}(t-T)\right)\begin{pmatrix}\psi_{|k|,l}\,e^{ik\theta}\\ -i\nu_{|k|,l}\psi_{|k|,l}\,e^{ik\theta}\end{pmatrix}.\end{aligned} \tag{6.29}$$

As in (6.20) it may be seen that

$$\left\|\begin{pmatrix}\psi_{|k|,l}\\ \pm i\nu_{|k|,l}\psi_{|k|,l}\end{pmatrix}\right\|_{\tilde{H}}^2=2\mu_{|k|,l}, \qquad \mu_{|k|,l}=(\nu_{|k|,l})^2.$$

Let w satisfy the wave equation and (6.4) with $w(x,y,0)\equiv 0$, $\zeta(x,y,0)=(\partial w/\partial t)(x,y,0)\equiv 0$ in $\mathbf{R}$. We expand $\binom{w}{\zeta}$ in the form

$$\begin{aligned}\begin{pmatrix}w(\cdot,\cdot,t)\\ \zeta(\cdot,\cdot,t)\end{pmatrix}&=w_{00}(t)\begin{pmatrix}\psi_{00}\\ 0\end{pmatrix}+\zeta_{00}(t)\begin{pmatrix}0\\ \psi_{00}\end{pmatrix}\\ &+\sum_{k=-\infty}^{\infty}\sum_{l=1}^{\infty}\left[w_{k,l}^{+}(t)\begin{pmatrix}\psi_{|k|,l}\,e^{ik\theta}\\ i\nu_{|k|,l}\psi_{|k|,l}\,e^{ik\theta}\end{pmatrix}+w_{k,l}^{-}(t)\begin{pmatrix}\psi_{|k|,l}\,e^{ik\theta}\\ -i\nu_{|k|,l}\psi_{|k|,l}\,e^{ik\theta}\end{pmatrix}\right].\end{aligned}$$

If $\tilde{w}$ satisfies the wave equation and the homogeneous boundary condition (cf. (4.14))

$$\frac{\partial \tilde{w}}{\partial \nu}(x, y, t) = 0, \qquad (x, y) \in \mathbf{B}, \quad t \geqq 0,$$

we find (cf. (4.16), (6.21)) that

$$\left(\begin{pmatrix} w(\cdot, \cdot, T) \\ \zeta(\cdot, \cdot, T) \end{pmatrix}, \begin{pmatrix} \tilde{w}(\cdot, \cdot, T) \\ \tilde{\zeta}(\cdot, \cdot, T) \end{pmatrix} \right)_{\tilde{H}} = \beta \int_0^T \int_{\mathbf{B}=\partial\mathbf{R}} u(x, y, t) \overline{\frac{\partial \tilde{w}}{\partial t}(x, y, t)}\, ds\, dt. \tag{6.30}$$

Employing (6.29), (6.3) successively for $\binom{w}{\zeta}$, we arrive at the equations, for $-\infty < k < \infty$, $l = 1, 2, 3, \cdots$,

$$\begin{aligned} 2\mu_{|k|,l} w^+_{k,l}(T) &= \beta \int_0^T \int_0^{2\pi} u(\theta, t) i\nu_{|k|,l} \exp\left(i\nu_{|k|,l}(T-t)\right) \psi_{|k|,l}(1)\, e^{-ik\theta}\, d\theta\, dt \\ &= 2\pi\beta i \nu_{|k|,l} \overline{\psi_{|k|,l}(1)} \int_0^T \exp\left(i\nu_{|k|,l}(T-t)\right) u_k(t)\, dt, \end{aligned} \tag{6.31}$$

$$\begin{aligned} 2\mu_{|k|,l} w^-_{k,l}(T) &= -\beta \int_0^T \int_0^{2\pi} u(\theta, t) i\nu_{|k|,l} \exp\left(-i\nu_{|k|,l}(T-t)\right) \psi_{|k|,l}(1)\, e^{-ik\theta}\, d\theta\, dt \\ &= -2\pi\beta i \nu_{|k|,l} \overline{\psi_{|k|,l}(1)} \int_0^T \exp\left(-i\nu_{|k|,l}(T-t)\right) u_k(t)\, dt. \end{aligned} \tag{6.32}$$

We find also, taking $\begin{pmatrix} \tilde{w} \\ \tilde{\zeta} \end{pmatrix}$ in the second form given in (6.27), that

$$\zeta_{00}(T) = \beta \int_0^T \int_0^{2\pi} u(\theta, t) \overline{\psi_{00}}\, d\theta\, dt = 2\pi\beta \overline{\psi_{00}} \int_0^T u_0(t)\, dt. \tag{6.33}$$

Since this must be true for all T and $(d/dt) w_{00}(t) = \zeta_{00}(t)$, we have also

$$w_{00}(T) = 2\pi\beta \overline{\psi_{00}} \int_0^T (T-t) u_0(t)\, dt. \tag{6.34}$$

Since $\mu_{|k|,l} = (\nu_{|k|,l})^2$, (6.31), (6.32) become

$$\begin{aligned} \frac{\nu_{|k|,l}}{\pi\beta i} w^+_{k,l}(T) &= \overline{\psi_{|k|,l}(1)} \int_0^T \exp\left(i\nu_{|k|,l}(T-t)\right) u_k(t)\, dt \\ &= B_{|k|,l} J_{|k|,l}(\nu_{|k|,l}) \int_0^T \exp\left(i\nu_{|k|,l}(T-t)\right) u_k(t)\, dt, \end{aligned} \tag{6.35}$$

$$\frac{\nu_{|k|,l}}{\pi\beta i} w^-_{k,l}(T) = -B_{|k|,l} J_{|k|,l}(\nu_{|k|,l}) \int_0^T \exp\left(-i\nu_{|k|,l}(T-t)\right) u_k(t)\, dt. \tag{6.36}$$

Taking account of the fact that

$$\overline{\frac{\partial \phi_{|k|,l}}{\partial r}(1)} = \omega_{|k|,l} A_{|k|,l} \frac{\partial J_{|k|,l}}{\partial r}(\omega_{|k|,l}),$$

(6.24) and (6.25) yield

$$\frac{\omega_{|k|,l}}{\pi\alpha} v^+_{k,l}(T) = A_{|k|,l} \frac{\partial J_{|k|,l}}{\partial r}(\omega_{|k|,l}) \int_0^T \exp\left(i\omega_{|k|,l}(T-t)\right) u_k(t)\, dt, \tag{6.37}$$

$$\frac{\omega_{|k|,l}}{\pi\alpha} v^-_{k,l}(T) = A_{|k|,l} \frac{\partial J_{|k|,l}}{\partial r}(\omega_{|k|,l}) \int_0^T \exp\left(-i\omega_{|k|,l}(T-t)\right) u_k(t)\, dt. \tag{6.38}$$

On the other hand

$$\frac{\partial \phi_{|k|,0}}{\partial r}(1) = A_{|k|,0}|k|$$

so (6.23) gives

$$\frac{|k|+1}{\pi\alpha} v_{k,0}(T) = A_{|k|,0} \int_0^T u_k(t)\, dt. \tag{6.39}$$

Using the formula (6.18) and (6.28) for $A_{|k|,l}$ and $B_{|k|,l}$ we have

$$\frac{\nu_{|k|,l}}{\pi\beta i} w^+_{k,l}(T) = \frac{\nu_{|k|,l}}{\sqrt{\pi}(\mu_{|k|,l}-k^2)^{1/2}} \int_0^T \exp\left(i\nu_{|k|,l}(T-t)\right) u_k(t)\, dt, \tag{6.40}$$

$$\frac{\nu_{|k|,l}}{\pi\beta i} w^-_{k,l}(T) = \frac{-\nu_{|k|,l}}{\sqrt{\pi}(\mu_{|k|,l}-k^2)^{1/2}} \int_0^T \exp\left(-i\nu_{|k|,l}(T-t)\right) u_k(t)\, dt, \tag{6.41}$$

$$\frac{\omega_{|k|,l}}{\pi\alpha} v^+_{k,l}(T) = \frac{1}{\sqrt{\pi}} \int_0^T \exp\left(i\omega_{|k|,l}(T-t)\right) u_k(t)\, dt, \tag{6.42}$$

$$\frac{\omega_{|k|,l}}{\pi\alpha} v^-_{k,l}(T) = \frac{1}{\sqrt{\pi}} \int_0^T \exp\left(-i\omega_{|k|,l}(T-t)\right) u_k(t)\, dt. \tag{6.43}$$

The equations (6.39) become, in view of (6.17),

$$\frac{\sqrt{2}\sqrt{|k|(|k|+1)}}{\pi\alpha} v_{k,0}(T) = \frac{\sqrt{2|k|}}{\sqrt{\pi}} \int_0^T u_k(t)\, dt. \tag{6.44}$$

This is valid, but meaningless, for $k=0$. It is easy to see that in the case $k=0$ we should use

$$\frac{1}{\sqrt{\pi}\alpha} v_{00}(T) = \int_0^T u_k(t)\, dt. \tag{6.45}$$

The equations (6.33) and (6.34) are left as they appear. We note that all of the coefficients

$$\frac{\nu_{|k|,l}}{\sqrt{\pi}(\mu_{|k|,l}-k^2)^{1/2}}, \quad \frac{1}{\sqrt{\pi}}, \quad \frac{\sqrt{2|k|}}{\sqrt{\pi}}, \quad k \neq 0, \quad 2\pi\beta$$

are bounded away from zero, uniformly with respect to k.

It is also possible to show, using the work [10], [11] of K. D. Graham, that the numbers

$$0, \nu_{|k|,1}, \omega_{|k|,1}, \nu_{|k|,2}, \omega_{|k|,2}, \cdots, \nu_{|k|,j}, \omega_{|k|,j}, \cdots$$

are separated by a gap at least equal to $\pi/2$, again uniformly with respect to k. Applying the result [14] of A. E. Ingham along with the work of Duffin and Schaeffer [7], much as in [12], [2], [3], we conclude the existence of functions $u_k(t)$ in $L^2[0, T]$, for any fixed $T>4$, solving the above moment problems, $-\infty<k<\infty$. Moreover, the result of Ingham implies as explained in [12], [26], that for each k

$$c^{-2}N_k^2 \leqq \int_0^T |u_k(t)|^2\, dt \leqq C^2 N_k^2$$

where

$$N_k^2 = 2|k|(|k|+1)|v_{k,0}(T)|^2 + \sum_{l=1}^{\infty} \lambda_{|k|,l}|v_{k,l}^+(T)|^2 + \sum_{l=1}^{\infty} \lambda_{|k|,l}|v_{k,l}^-(T)|^2$$
$$+ \sum_{l=1}^{\infty} \mu_{|k|,l}|w_{k,l}^+(T)|^2 + \sum_{l=1}^{\infty} \mu_{|k|,l}|w_{k,l}^-(T)|^2$$

$k = \pm 1, \pm 2, \cdots$. For $k = 0$ we must add $|\zeta_{00}(T)|^2 + |w_{00}(T)|^2$. Since

$$\int_0^T \int_0^{2\pi} |u(\theta, t)|^2 \, d\theta \, dt = \sum_{k=-\infty}^{\infty} \int_0^T |u_k(t)|^2 \, dt \tag{6.46}$$

we see that the above moment problems, equivalent to the control problem, can be solved with (6.46) finite, provided that

$$\sum_{k=-\infty}^{\infty} N_k^2 < \infty,$$

which is the same as saying that the norm of the final state in $\hat{H}$ should be finite. We have, then, the exact controllability result that any $\hat{H}$ state may be controlled to any other $\hat{H}$ state during a time interval of length $T > 4$ with the control configuration we have described here. As discussed in connection with the wave equation in [2], [3], and [12], one cannot be sure that the state of the system remains in $\hat{H}$ for all $t \in [0, T]$. However, in the present case of the Maxwell equations one can show that these states do lie in $H = H_{\mathbf{E},d}(\mathbf{R})$. Again this is exactly twice the critical control time for the (single) wave equation under the same circumstances, as established in [10], [12].

7. Concluding remarks. The approximate controllability results of § 5 would appear to be extendable to domains other than rectangular ones but the precise method of extension remains to be worked out. We will indicate some aspects of this problem which are clear from our current work.

First of all, the result of § 5 is almost trivially extended to the case where control is exercised only on a subset $\{(\pi, y) | 0 \leqq a \leqq y \leqq b \leqq \pi\}$, $b > a$, of $\{(\pi, y) | 0 \leqq y \leqq \pi\}$. The only change is that the interval $1 + \delta \leqq t \leqq T - 1 - \delta$ appearing in (5.14) and subsequently must be modified to $d + \delta \leqq t \leqq T - d - \delta$ where

$$d = \inf_{a \leqq y \leqq b} \left\{ \sup_{\substack{0 \leqq \xi \leqq \pi \\ 0 \leqq \eta \leqq \pi}} \{[(\pi - \xi)^2 + (\eta - y)^2]^{1/2}\} \right\}.$$

If $\phi(\pi, y, t) \equiv (\partial\phi/\partial x)(\pi, y, t) \equiv 0$ for $\delta \leqq t \leqq T - \delta$, $a \leqq y \leqq b$, the Holmgren theorem will still apply to show that $\phi(x, y, t) \equiv 0$, $(x, y) \in \mathbf{R}$, $d + \delta \leqq t \leqq T - d - \delta$. After that the remainder of the proof is the same: the same eigenfunctions and frequencies must be dealt with, the functions $\sin jy$, $\cos jy$ are still independent on $a \leqq y \geqq b$ and $b > a$ and the conditions

$$\hat{v}_0(x, y) = 0, \qquad (x, y) \in \mathbf{B} - \{(\pi, y) | a \leqq y \leqq b\}$$
$$\frac{\partial \hat{v}_0}{\partial x}(\pi, y) = 0, \qquad a \leqq y \leqq b,$$

still show $\hat{v}_0(x, y) \equiv 0$ in $\mathbf{R}$.

The first limitation of the method which we have used in § 5 lies in its dependence on the construction of $\phi(x, y, t)$ as a linear combination of partial derivatives of $\hat{v}$ and $\hat{w}$. It is necessary to have *a solution of the wave equation to which Holmgren's theorem*

may be applied. This part of the proof can still be used for nonrectangular domains as long as a portion of the boundary on which control is applied is a straight line segment. Assuming the segment parallel to the y-axis, one can construct ϕ by the formula (5.8) again and show that ϕ and $\partial\phi/\partial x$ both vanish on the straight line segment in question, allowing subsequent application of the Holmgren theorem to show $\phi(x, y, t) \equiv 0$ for $(x, y) \in \mathbf{R}$ and t in some interval $d + \delta \leqq t \leqq T - d - \delta$, with d depending on the geometry of $\mathbf{R}$. But then we are faced with a second limitation.

The second limitation of the method which we have used lies in its reliance on the specific form of the eigenfunctions and frequencies to pass from $\phi(x, y, t) \equiv 0$ to the conclusion that both $\hat{v}(x, y, t)$ and $\hat{w}(x, y, t)$ are likewise identically zero. It needs to be emphasized that no local analysis will suffice here. In the one-dimensional case (see our remarks at the end of § 3) if the control problem is stated for boundary conditions

$$v(0, t) = 0, \qquad \frac{\partial v}{\partial t}(1, t) = \alpha u(t), \tag{7.1}$$

$$\frac{\partial w}{\partial x}(0, t) = 0, \qquad \frac{\partial w}{\partial x}(1, t) = \beta u(t) \tag{7.2}$$

the $\tilde{v}$, $\tilde{w}$ constructed as in § 4 will satisfy the wave equation and

$$\tilde{v}(0, t) = 0, \qquad \frac{\partial \tilde{v}}{\partial t}(1, t) = 0, \tag{7.3}$$

$$\frac{\partial \tilde{w}}{\partial x}(0, t) = 0, \qquad \frac{\partial \tilde{w}}{\partial x}(1, t) = 0, \tag{7.4}$$

$$\alpha \frac{\partial \tilde{v}}{\partial x}(1, t) + \beta \frac{\partial \tilde{w}}{\partial t}(1, t) \equiv \phi(1, t) = 0. \tag{7.5}$$

Here if we take $\tilde{w}$ to be a nonzero solution of the wave equation satisfying (7.4) and take

$$\tilde{v}(x, t) = -\frac{\beta}{\alpha} \int_0^x \frac{\partial \tilde{w}}{\partial t}(\xi, t)\, d\xi,$$

we clearly have $\tilde{v}(0, t) = 0$,

$$\begin{aligned}
\frac{\partial \tilde{v}}{\partial t}(1, t) &= \frac{\beta}{\alpha} \int_0^1 \frac{\partial^2 \tilde{w}}{\partial t^2}(\xi, t)\, d\xi \\
&= -\frac{\beta}{\alpha} \int_0^1 \frac{\partial^2 \tilde{w}}{\partial \xi^2}(\xi, t)\, d\xi = \frac{\beta}{\alpha}\left(\frac{\partial \tilde{w}}{\partial x}(0, t) - \frac{\partial \tilde{w}}{\partial x}(1, t)\right) = 0, \\
\frac{\partial^2 \tilde{v}}{\partial t^2}(x, t) &= -\frac{\beta}{\alpha} \int_0^x \frac{\partial^3 \tilde{w}}{\partial t^3}(\xi, t)\, d\xi \\
&= -\frac{\beta}{\alpha} \int_0^x \frac{\partial^3 \tilde{w}}{\partial t\, \partial \xi^2}(\xi, t)\, d\xi = -\frac{\beta}{\alpha} \frac{\partial^2 \tilde{w}}{\partial t\, \partial x}(x, t) = \frac{\partial^2 \tilde{v}}{\partial x^2}(x, t)
\end{aligned}$$

so that $\tilde{v}$ satisfies the wave equation and, clearly, (7.5) is also satisfied. Thus the wave equation with (7.1), (7.2) is not approximately controllable; $\phi(x, t) \equiv \alpha(\partial\tilde{v}/\partial x)(x, t) + \beta(\partial\tilde{w}/\partial t)(x, t) \equiv 0$ but this does not imply that $\tilde{v}$ or $\tilde{w}$ are identically equal to zero. The additional condition which makes this work in (3.7)ff. is the fact

that one can show there that

$$-\alpha\frac{\partial\tilde{v}}{\partial x}(0,t)+\beta\frac{\partial\tilde{w}}{\partial t}(0,t)=0.$$

It seems likely that the question of whether or not $\phi=0$ implies that both $\hat{v}$ and $\hat{w}$, equivalently $\tilde{v}$ and $\tilde{w}$, are both zero must eventually reduce to a boundary value problem of an as yet unidentified type.

At the present writing there is only one, rather curious, result which we can offer which yields approximate controllability for a domain $\mathbf{R}$ of rather general shape. We suppose that the "control boundary" $\mathbf{B}_1\subset\mathbf{B}=\partial\mathbf{R}$ includes two nonparallel line segments, l_1 and l_2, with unit exterior normals ν_1 and ν_2. Proceeding as before we can show, applying the Holmgren theorem together with

$$\frac{\partial\hat{v}}{\partial t}=0 \quad \text{on } l_1, l_2,$$

$$\frac{\partial\hat{w}}{\partial\nu_i}=0, \quad i=1,2 \quad \text{on } l_1, l_2, \text{ respectively,}$$

$$\alpha\frac{\partial\hat{v}}{\partial\nu_i}+\beta\frac{\partial\hat{w}}{\partial t}=0, \quad i=1,2 \quad \text{on } l_1, l_2, \text{ respectively,}$$

that both

$$\phi_1=\alpha\frac{\partial\hat{v}}{\partial\nu_1}+\beta\frac{\partial\hat{w}}{\partial t}, \tag{7.6}$$

$$\phi_2=\alpha\frac{\partial\hat{v}}{\partial\nu_2}+\beta\frac{\partial\hat{w}}{\partial t}, \tag{7.7}$$

must vanish identically in $\mathbf{R}$ for $d+\delta\leqq t\leqq T-d-\delta$, $\delta>0$ arbitrary, $d>0$ depending on the geometry of $\mathbf{R}$ and $\mathbf{B}$, the location of l_1 and l_2 within $\mathbf{B}$, etc. But then both ϕ_1 and ϕ_2 must vanish on l_1 (say) for these values of t. Subtracting (7.6) from (7.7), we see that

$$\alpha\left(\frac{\partial\hat{v}}{\partial\nu_1}-\frac{\partial\hat{v}}{\partial\nu_2}\right)=0 \quad \text{on } l_1\times[d+\delta, T-d-\delta].$$

This shows, since l_1 and l_2 are not parallel, that a nontangential derivative of $\hat{v}$ vanishes on $l_1\times[d+\delta, T-d-\delta]$. Combining this with $\partial\hat{v}/\partial t=0$ on l_1 and applying the Holmgren theorem to $\hat{v}$ alone, much as in [5], [13], we are able to conclude $\hat{v}\equiv 0$, provided T is appropriately large. Then one easily has the same result for $\hat{w}$ and approximate controllability follows.

This result gives approximate controllability for $\mathbf{R}$ equal to the interior of any closed polyhedron in R^2 with control on at least two sides.

Further inspection of this argument shows that only l_2 needs to be assumed to be a line segment. That is needed in order to identify ϕ_2 as a solution of the wave equation. We may then take l_1 to be any smooth portion of $\mathbf{B}_1$ which is never parallel to l_1 and achieve the same result.

Finally, let us indicate that we are very much aware of the limitations, from the point of view of actual implementation, of the control configuration discussed in this paper. In principle, at least, the boundary conditions (1.7), (1.8), along with the further "single layer" condition discussed in connection with (3.1), could be achieved with conducting bars attached to terminals as shown in Fig. 3.

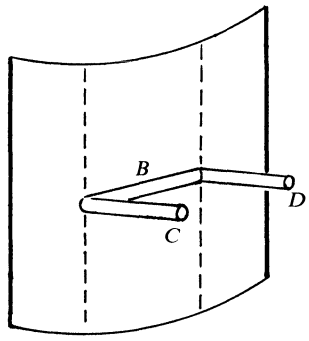


FIG. 3. *Conducting bar and busses.*

We have not considered any effects of propagation delays in the controlling circuits—i.e., we have not assumed that these are distributed parameter systems. This assumption, and evident limitations on the speed with which prescribed currents can be computed and established in the controlling circuits together with sensing limitations, place admittedly severe limitations on what can be done "open loop." It is likely that the eventual significance of our results will be most evident in connection with closed loop behavior wherein time varying magnetic fields $\vec{H}$ near the boundary of Ω induce currents in the bars B which, being resistive, will then act as energy dissipators. We hope to discuss this topic in later work.

Another control configuration may be obtained by supposing the boundary of Ω to be a sheet of material to which electromagnets are attached in a dense array as shown in Fig. 4. If J denotes the current through, the windings of the electromagnets, then we shall have

$$H_\nu = \alpha J$$

where α is dependent on the electromagnet's configuration. The theory in this case will take much the same form as the one discussed in this paper.

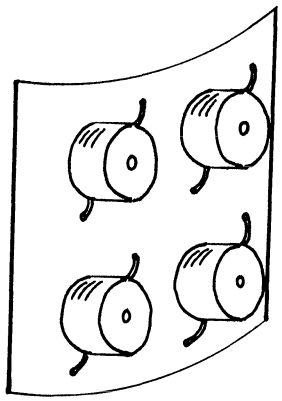


FIG. 4. *Electromagnet array.*

Acknowledgment. The author would like to thank Ms. K. Kime for careful reading of the manuscript and some very important corrections.

REFERENCES

[1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] G. CHEN, *Control and stabilization for the wave equation in a bounded domain*, this Journal, 17 (1979), pp. 66-81.

[3] G. CHEN, *Part* II, this Journal, 19 (1981), pp. 114-122.
[4] P. CLEMMOW, *An Introduction to Electromagnetic Theory*, Cambridge Univ. Press, Cambridge, 1973.
[5] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics*, II: *Partial Differential Equations*, Interscience, New York, 1962.
[6] S. DOLECKI AND D. L. RUSSELL, *A general theory of observation and control*, this Journal, 15 (1977), pp. 185-220.
[7] R. J. DUFFIN AND A. C. SCHAEFFER, *A class of nonharmonic Fourier series*, Trans. Amer. Math. Soc., 72 (1952), pp. 341-366.
[8] H. O. FATTORINI, *Estimates for sequences biorthogonal to certain complex exponentials and boundary control of the wave equation*, Seminaires IRIA, 1976.
[9] K. O. FRIEDRICHS, *Mathematical Methods of Electromagnetic Theory*, Lectures, 1972-73, Courant Institute of Mathematical Sciences, New York Univ., New York, 1974.
[10] K. D. GRAHAM, *On boundary value control of distributed hyperbolic systems*, Dept. Mathematics, Univ. Minnesota, Minneapolis, March 1973.
[11] ———, *Separation of eigenvalues of the wave equation for the unit ball in R^N*, Studies in Applied Mathematics, Vol. LII (1973), pp. 329-343.
[12] K. D. GRAHAM AND D. L. RUSSELL, *Boundary value control of the wave equation in a spherical region*, this Journal, 13 (1975), pp. 174-196.
[13] L. HORMANDER, *Linear Partial Differential Operators*, Springer-Verlag, New York, Heidelberg, Berlin, 1963.
[14] A. E. INGHAM, *Some trigonometrical inequalities with applications to the theory of series*, Math. Z., 41 (1936), pp. 367-379.
[15] J. LAGNESE, *Boundary value control of a class of hyperbolic equations in a general region*, this Journal, 15 (1977), pp. 973-983.
[16] ———, *Exact boundary value controllability of a class of hyperbolic equations*, this Journal, 16 (1978), pp. 1000-1017.
[17] N. LEVINSON, *Gap and density theorems*, AMS Colloquium Publications 26, American Mathematical Society, Providence, RI, 1940.
[18] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, Heidelberg, Berlin, 1971, trans. by S. K. Mitter.
[19] J. L. LIONS AND E. MAGENES, *Problèmes aux limites non-homogènes*, Vol. I and II, Dunod, Paris, 1968.
[20] D. L. RUSSELL, *On boundary-value controllability of linear symmetric hyperbolic systems*, in Mathematical Theory of Control, A. Balakrishnan and L. Neustadt, eds., Academic Press, New York, 1967.
[21] ———, *Nonharmonic Fourier series on the control theory of distributed parameter systems*, J. Math. Anal. Appl., 18 (1967), pp. 542-559.
[22] ———, *Boundary value control theory of the higher dimensional wave equation*, this Journal, 9 (1971), pp. 29-42.
[23] ———, *Part* II, this Journal, 9 (1971), pp. 401-419.
[24] ———, *Control theory of hyperbolic equations related to certain questions in harmonic analysis and spectral theory*, J. Math. Anal. Appl., 40 (1972), pp. 336-368.
[25] ———, *Exact boundary value controllability theorems for wave and heat processes in star-complemented regions*, in Differential Games and Control Theory, Roxin, Liu and Sternberg, eds., Marcel Dekker, New York, 1974.
[26] ———, *Controllability and stabilizability theory for linear partial differential equations: recent progress and open questions*, SIAM Rev., 20 (1978), pp. 639-739.
[27] L. SCHWARTZ, *Etude des sommes d'exponentielles*, deuxieme edition, Hermann, Paris, 1959.
[28] A. N. TIKHONOV AND A. A. SAMARSKII, *Equations of Mathematical Physics*, trans. by A. R. M. Robson and P. Basa, D. M. Brink, ed., Macmillan, New York, 1963.
[29] P. K. C. WANG, *Optimal control of a class of linear symmetric hyperbolic systems with applications to plasma confinement*, J. Math. Anal. Appl., 28 (1969), pp. 594-608.
[30] P. K. C. WANG AND W. A. JANOS, *A control-theoretic approach to the plasma confinement problem*, J. Optim. Theory Appl., 5 (1970), pp. 313-329.
[31] P. K. C. WANG, *Feedback stabilization of highly conducting plasmas*, Phys. Rev. Letters, 24 (1970), pp. 362-364.
[32] C. H. WILCOX, *Wave operators and asymptotic solutions of wave propagation problems of classical physics*, Arch. Rat. Mech. Anal., 22 (1966), pp. 37-78.
[33] ———, *Steady-state wave propagation in homogeneous anisotropic media*, Arch. Rat. Mech. Anal., 25 (1967), pp. 201-242.
[34] ———, *Transient wave propagation in homogeneous anisotropic media*, Arch. Rat. Mech. Anal., 37 (1970), pp. 323-343.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 2, March 1986

005

LOCAL REALIZATIONS OF NONLINEAR CAUSAL OPERATORS*

BRONISŁAW JAKUBCZYK†

Abstract. We give necessary and sufficient conditions for a causal operator to have a local realization of class $C^k(k=1,\cdots,\infty,\omega)$ of the form

$$\dot{x}=f(x,u),\qquad y=h(x).$$

We show that two minimal local realizations of the same response map are locally diffeomorphic.

Key words. causal operators, nonlinear systems, nonlinear realization theory, existence and uniqueness of realizations

1. Introduction. Let Ω, Y be sets and let $\mathcal{U}$, $\mathcal{Y}$ be sets of functions $[0,T)\to\Omega$ and $[0,T]\to Y$, respectively. We shall assume that $\mathcal{U}$ contains all piecewise constant functions and $Y=R^r$. An operator $F:\mathcal{U}\to\mathcal{Y}$ is called (strictly) causal if $u|_{[0,t)}=v|_{[0,t)}$ implies $(F(u))(t)=(F(v))(t)$ for any $u,v\in\mathcal{U}$ and $0\leqq t\leqq T$ (here t has interpretation of time).

In this paper, for general classes of functions $\mathcal{U}$, $\mathcal{Y}$ we give necessary and sufficient conditions for the operator $u\to y=F(u)$ to be locally represented by a system of the form

$$\dot{x}=f(x,u),\quad x(0)=x_0,\quad y=h(x), \tag{1}$$

where $x(t)\in V$ and V is an open subset of R^n for some $n>0$. "Locally" means "for small times" here.

The problem arises in the system theory, where Ω, Y are input and output spaces, respectively; u, y are input and output signals; and F is an operator describing input-output behavior of a system (black box). Then (1) is a desired internal description of the system (called realization), which is to be found.

A solution of a global version of the problem was given in [14] (uniqueness) and [8] (existence). Related results may be found in [1]-[7] and [9]-[13]. In the global version of [8] the time horizon is infinite, $T=\infty$, the realization is sought with the state space V being any differentiable manifold and system (1) obtained is complete (has solutions forward and backward for all times). In this case the response map is required to have an extension to an "input group". The construction of the realization is obtained via a certain factorization (Nerode equivalence) of this group.

The main difficulty of the local problem is that it is hard to localize the group argument. Therefore, another construction has to be used. The construction used in this paper is completely elementary and given explicitly in terms of functions, which is an advantage from the point of view of possible practical computations. Another advantage of the approach presented in this paper is that it requires minimal hypotheses on the response map, only regularity and finiteness of the response map (the extension assumption of [8] is dropped). However, under these assumptions it is not possible to obtain a realization which is good for all inputs, at least in the C^∞ case. Thus, we construct a realization which is good after some transient behavior. In the analytic case this transient behavior can be dropped.

* Received by the editors February 2, 1984, and in revised form December 1, 1984.
† Institute of Mathematics, Polish Academy of Sciences, 00-950 Warsaw, Sniadeckich 8, Poland.

Another approach to the local problem was presented in [5], where the operator F is given by a series with iterated integrals and f is analytic of the form $f(x, u) = f_0(x) + \sum_i u_i g_i(x)$. The analytic version of our result (Theorem 2) can be used to derive the existence result in [5].

A strictly causal operator F can be uniquely represented by a map $P: \mathcal{S} \to Y$ (called "response map", "input-output map" or "causal functional" if $Y = R$), where $\mathcal{S}$ is the set of all functions being restrictions to $[0, t)$, $0 \leqq t \leqq T$ of functions in $\mathcal{U}$, and

$$P(a) = (F(u))(t), \qquad a = u|_{[0, t)}.$$

We shall formulate our results in terms of the map P. Note that P defines F by the formula $(F(u))(t) = P(u|_{[0, t)})$, $u \in \mathcal{U}$.

In the paper we use the norm $|x| = |x_1| + \cdots + |x_n|$ in R^n and the norm $\|A\|$ of a matrix A denotes the norm of the corresponding operator in such normed spaces. Composition of functions is denoted by $\circ$.

2. Main results. Assume that $\mathcal{U} = \mathcal{U}_{pc}$, where $\mathcal{U}_{pc}$ consists of all piecewise constant functions $[0, T) \to \Omega$, with $T > 0$ fixed, and let $\mathcal{S}_{pc}$ be the corresponding class of restricted function. Let $R_+ = [0, \infty)$. For $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_p)$, $\mathbf{t} = (t_1, \cdots, t_p)$ we denote by

$$a = \boldsymbol{\alpha}(\mathbf{t}) = (t_1, \alpha_1) \cdots (t_p, \alpha_p), \qquad p \geqq 1, \quad t_i \in R_+, \quad \alpha_i \in \Omega, \tag{2}$$

the function in $\mathcal{S}_{pc}$ given by $a(t) = \alpha_i$ for $t \in [t_1 + \cdots + t_{i-1}, t_1 + \cdots + t_i)$. Denote $|a| = t_1 + \cdots + t_p$.

The elements $a, b \in \mathcal{S}_{pc}$ can be multiplied; by ab we mean the concatenation of a and b (writing one sequence (2) after the other). If $|a| + |b| \leqq T$, then $ab \in \mathcal{S}_{pc}$. We include the function e with empty domain (empty sequence (2)) in $\mathcal{S}_{pc}$, it plays the role of identity.

DEFINITION 1. Let $P = (P_1, \cdots, P_r)$ be a map $P: \mathcal{S}_{pc} \to R^r$. We say that $\Sigma = (V, f, h, x_0)$ is a *local realization of P of class C^k after function $c \in \mathcal{S}_{pc}$* if

(i) $x_0 \in V \subset R^n$ and $h: V \to R^r$ is a function of class C^k,

(ii) $f(\cdot, \alpha)$, $\alpha \in \Omega$, are vector fields on V of class C^{k-1} and their local flows $\Phi^f_{(t, \alpha)}(x)$ are of class C^k with respect to (t, x),

(iii) $P(ca) = h \circ \Phi^f_a(x_0)$

for $a = \boldsymbol{\alpha}(\mathbf{t})$ with any sequence $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_p)$ and $\mathbf{t}$ in a neighborhood of $0 \in R^p_+$. Here

$$\Phi^f_a = \Phi^f_{(t_p, \alpha_p)} \circ \cdots \circ \Phi^f_{(t_1, \alpha_1)}.$$

We also say that Σ is a *local realization around function $c = \boldsymbol{\beta}(\mathbf{s}^*)$* if (iii) is satisfied for all $c = \boldsymbol{\beta}(\mathbf{s})$ and $a = \boldsymbol{\alpha}(\mathbf{t})$ with $\mathbf{s}, \mathbf{t}$ in a neighborhood of $(\mathbf{s}^*, 0) \in R^m \times R^p_+$ where $x_0 = x_0(\mathbf{s})$. Finally, Σ is called a *realization of P on an interval* $[0, \sigma]$ if (iii) is satisfied for all $a \in \mathcal{S}_{pc}$ with $|a| \leqq \sigma$.

In case $k = 1$ one should read in (ii) "there exist local flows $\Phi^f_{(t, \alpha)}(x)$ of class C^1 of $f(\cdot, \alpha)$". The number n is called dimension of the realization, $n = \dim \Sigma$.

Let E_δ be the simplex $E_\delta = \{\mathbf{t} \in R^p_+ \mid |\mathbf{t}| < \delta\}$. We shall use the following assumptions on P.

(A1) Functions $\mathbf{t} \to P(\boldsymbol{\alpha}(\mathbf{t}))$ are of class C^k on E_T for any sequence $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_p)$, $\alpha_i \in \Omega$, and $p \geqq 1$.

Here and below a function defined on a (partially) closed polyhedron W is called of class C^k if all partial derivatives (directional partial derivatives) up to order k exist and are continuous on W. For $k = \omega$ we require that the Taylor series taken at any point of W is convergent in a neighborhood of this point.

(A2) The functions $P(\boldsymbol{\alpha}(\mathbf{t}))$ and their partial derivatives $(\partial/\partial t_i)P(\boldsymbol{\alpha}(\mathbf{t}))$ satisfy the Lipschitz conditions

$$|P(\boldsymbol{\alpha}(\mathbf{t}))-P(\boldsymbol{\alpha}(\mathbf{t}'))|\leqq L|\mathbf{t}-\mathbf{t}'|,$$

$$\left|\frac{\partial}{\partial t_i}P(\boldsymbol{\alpha}(\mathbf{t}))-\frac{\partial}{\partial t_i}P(\boldsymbol{\alpha}(\mathbf{t}'))\right|\leqq M|\mathbf{t}-\mathbf{t}'|,\qquad i=1,\cdots,p,$$

on E_T with constants L, M independent of the sequence $\boldsymbol{\alpha}$.

Let $\mathbf{b}=(b_1,\cdots,b_q)$, $b_i\in\mathcal{S}_{pc}$ and take $\delta=T-\max\{|b_1|,\cdots,|b_q|\}$. Define a map $\Psi^{\mathbf{b}}_{\boldsymbol{\alpha}}\colon E_\delta\to R^{rq}$,

$$\Psi^{\mathbf{b}}_{\boldsymbol{\alpha}}(\mathbf{t})=(P(\boldsymbol{\alpha}(\mathbf{t})b_1),\cdots,P(\boldsymbol{\alpha}(\mathbf{t})b_q)).$$

Define

$$\operatorname{rank} P=\sup_{\boldsymbol{\alpha},\mathbf{b},\mathbf{t}}\operatorname{rank} d\Psi^{\mathbf{b}}_{\boldsymbol{\alpha}}(\mathbf{t}),$$

where $d\Psi^{\mathbf{b}}_{\boldsymbol{\alpha}}$ denotes the Jacobian of the map $\Psi^{\mathbf{b}}_{\boldsymbol{\alpha}}$ and the supremum is taken over all finite sequences $\boldsymbol{\alpha}$, $\mathbf{b}$, $\mathbf{t}$ such that $\Psi^{\mathbf{b}}_{\boldsymbol{\alpha}}(\mathbf{t})$ makes sense.

We say that the *rank of P is attained at function $c\in\mathcal{S}_{pc}$* if there exist a representation $c=\boldsymbol{\beta}(\mathbf{s})=(s_1,\beta_1)\cdots(s_m,\beta_m)$ of c and a sequence $\mathbf{b}$ such that $\operatorname{rank} d\Psi^{\mathbf{b}}_{\boldsymbol{\beta}}(\mathbf{s})=\operatorname{rank} P$.

THEOREM 1. *If the map $P:\mathcal{S}_{pc}\to R^r$ satisfies condition* (A1) *and rank P is finite, then there exists a local realization Σ of P of class $C^k(k=1,2,\cdots,\infty,\omega)$, with* $\dim\Sigma=\operatorname{rank} P$, *after (as well as around) any function c at which the rank is attained.*

If, additionally, condition (A2) *is satisfied, then Σ is a realization on an interval $[0,\sigma]$ and $f(x,u)$ is bounded on $V\times\Omega$ and Lipschitzian with respect to x.*

Remark 1. From regularity properties of solutions of differential equations one can see that, if the map P is defined by a realization Σ via formula (iii), then (A1) is satisfied and $\operatorname{rank} P\leqq\dim\Sigma$.

In the (real) analytic case, $k=\omega$, we can strengthen condition (A1).

(A1)′ There exist a function $\rho:\Omega\to(0,\infty)$ and a $\delta>0$ such that the functions $\mathbf{t}\to\mathbf{P}(\boldsymbol{\alpha}(\mathbf{t}))$ are of class C^ω on E_T and have analytic extensions to $E^\rho_{\delta,\boldsymbol{\alpha}}=\{\mathbf{t}\in R^p\,|\,\rho(\alpha_1)|t_1|+\cdots+\rho(\alpha_p)|t_p|<\delta\}$.

THEOREM 2. *The map P has a local realization of class C^ω (after the identity e) if it satisfies condition* (A1)′ *and rank P is finite. If P satisfies also* (A2), *then it has a realization on an interval $[0,\sigma]$, $\sigma>0$. Conversely, if P is given by a C^w realization via* (iii) *with $c=e$, then* (A1)′ *is satisfied and* rank P *is finite.*

Remark 2. This theorem is a generalization of the analytic version of the existence theorem of Fliess [4], [5] (one can prove that our rank is equivalent to the rank used in [4], [5]).

To formulate a uniqueness result, we introduce the following definitions. Let $\mathcal{S}^\tau_{pc}$ denote the set of all $a\in\mathcal{S}_{pc}$ such that $|a|\leqq\tau$. For a map $P:\mathcal{S}_{pc}\to R^r$ define a local rank by

$$\operatorname{rank}_{\mathrm{loc}} P=\inf_{0<\tau\leqq T}\operatorname{rank} P|\mathcal{S}^\tau_{pc}.$$

The maximal value of τ at which the infimum is attained is denoted by τ^*; then $\operatorname{rank}_{\mathrm{loc}} P=\operatorname{rank} P|\mathcal{S}^{\tau^*}_{pc}$.

Any quadruple $\Sigma=(V,f,h,x_0)$ defines a map P_Σ given by

$$P_\Sigma(a)=h\circ\Phi^f_{(t_p,\alpha_p)}\circ\cdots\circ\Phi^f_{(t_1,\alpha_1)}(x_0),$$

for all $a=(t_1,\alpha_1)\cdots(t_p,\alpha_p)\in\mathcal{S}_{pc}$ such that the right-hand side makes sense. Assume

that P_Σ is well defined for all $a \in \mathscr{S}_{pc}^{\delta}$. Then $\operatorname{rank}_{\mathrm{loc}} P_\Sigma$ is well defined, with the infimum taken over $\tau \in [0, \delta]$.

DEFINITION 2. A realization Σ is called *locally minimal* if $\dim \Sigma = \operatorname{rank}_{\mathrm{loc}} P_\Sigma$.

DEFINITION 3. Realizations Σ and $\hat{\Sigma}$ are called *locally diffeomorphic* at points $x_1 \in V$ and $\hat{x}_1 \in \hat{V}$ if there is a local diffeomorphism χ of a neighborhood $U \subset V$ of the point x_1 onto a neighborhood $\hat{U} \subset \hat{V}$ of the point $\hat{x}_1$ such that $\chi(x_1) = \hat{x}_1$ and

$$h(x) = \hat{h} \circ \chi(x), \qquad d\chi(x) f(x, \alpha) = \hat{f}(\chi(x), \alpha), \tag{3}$$

for $x \in U$ and $\alpha \in \Omega$.

THEOREM 3. *Let Σ and $\hat{\Sigma}$ be realizations of class $C^k (k = 1, \cdots, \infty, \omega)$ on an interval $[0, \sigma]$ of a map $P: \mathscr{S}_{pc} \to R^r$ (after the identity e). If Σ and Σ are locally minimal, then they are locally diffeomorphic (of class C^k) at the points $x_1 = \Phi_c^f(x_0)$ and $\hat{x}_1 = \Phi_c^{\hat{f}}(\hat{x}_0)$, where $c \in \mathscr{S}_{pc}$ is any function at which the rank of the map $P|\mathscr{S}_{pc}^{\delta}$ is attained with $\delta = \min\{\sigma, \tau^*\}$.*

Let us formulate an existence result for a general class of functions $\mathscr{U}$. Assume that Ω is a metric space. Let $\mathscr{U}$ be a subfamily of $\mathscr{M}_c([0, T]; \Omega)$ (the space of measurable functions the image of each is contained in a compact subset of Ω a.e.) and contains all piecewise constant functions $\mathscr{U}_{pc} \subset \mathscr{U}$. As before, $\mathscr{S}$ denotes the set of restrictions to the intervals $(0, t)$, $0 \leqq t \leqq T$, of functions in $\mathscr{U}$. For $a = u|_{[0, t)}$ we denote $|a| = t$.

The term "realization on the interval $(0, \sigma]$" means now that (iii) is satisfied for all $a \in \mathscr{S}$ such that $|a| \leqq \sigma$, where $\Phi_a^f(x)$ denotes the point after time $t = |a|$ of the trajectory of system (1) starting from x, with $u = a$.

Additional problems which appear now are to guarantee that the function $f(x, u)$ is continuous with respect to u in order to (1) have a solution and to guarantee that (iii) is satisfied for nonpiecewise constant a. Thus we put the following additional conditions.

(A3) Functions $(\mathbf{t}, \boldsymbol{\alpha}) \to P(\boldsymbol{\alpha}(\mathbf{t}))$ are continuous as maps $E_T \times \Omega^p \to R^r$ together with partial derivatives $(\partial/\partial t_i) P(\boldsymbol{\alpha}(\mathbf{t}))$, $i = 1, \cdots, p$.

(A4) If $|a| = |a_1| = |a_2| = \cdots$, where $a, a_i \in \mathscr{S}$, the images of a and a_i are in a common compact subset of Ω and $a_i \to a$ pointwise on $[0, |a|)$, a.e., then $P(a_i) \to P(a)$.

THEOREM 4. *If the function $P: \mathscr{S} \to R^r$ satisfies assumptions* (A1), (A2), (A3), (A4) *and* rank P *is finite, then there exists a realization Σ of P of class $C^k (k = 1, \cdots, \infty, \omega)$ on an interval $[0, \sigma]$ after any function $c \in \mathscr{S}_{pc}$ at which the rank is attained. Additionally,* $\dim \Sigma = \operatorname{rank} P$ *and the function $f(x, u)$ is bounded, continuous with respect to (x, u) and satisfies the Lipschitz condition with respect to x uniformly on $u \in \Omega$.*

If $k = \omega$ and assumption (A1) *is replaced by* (A1)′, *then we can take $c = e$. If, additionally, Ω is compact then assumption* (A2) *can be replaced by* (A3) *strengthened to second order partial derivatives.*

3. Proofs.

Proof of Theorem 1. Let $c = \boldsymbol{\beta}(\mathbf{s}^*)$, $\mathbf{s}^* = (s_1^*, \cdots, s_m^*)$, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_m)$ and $\mathbf{b} = (b_1, \cdots, b_q)$ be such that $\operatorname{rank} d\Psi_{\boldsymbol{\beta}}^{\mathbf{b}}(\mathbf{s}^*) = \operatorname{rank} P = n$. Enlarging, possibly, the number of b_i in $\mathbf{b}$ we can assume that one of the b_i is identity e.

There exists a $n \times n$, nonsingular submatrix A of the Jacobian $J = d\Psi_{\boldsymbol{\beta}}^{\mathbf{b}}(\mathbf{s}^*)$. Assume that A consists of the first n rows and the first n columns of J. Define functions $\eta: R^n \to R^m$, $\P: R^\nu \to R^n$, $\nu = rq$, by

$$\eta(\tau_1, \cdots, \tau_n) = (\tau_1, \cdots, \tau_n, s_{n+1}^*, \cdots, s_m^*),$$

$$\P(y_1, \cdots, y_\nu) = (y_1, \cdots, y_n).$$

Then the composition

$$\Psi(\tau) = \P \circ \Psi^{\mathbf{b}}_{\boldsymbol{\beta}} \circ \eta(\tau) \tag{4}$$

has a nonsingular Jacobian at $\tau = \tau^*$ such that $\eta(\tau^*) = \mathbf{s}^*$. Thus, there exists a neighborhood V of τ^* such that Ψ is a diffeomorphism of V onto an open subset U of R^n, rank $d\Psi(\tau) = n$ for $\tau \in \operatorname{cl} V$ and Ψ is a homeomorphism $\Psi: \operatorname{cl} V \to \operatorname{cl} U$. For technical reasons it will be convenient to take V of the form

$$V = (\tau \in R^n \,|\, |\tau - \tau^*| < C\} \subset R^n_+ \tag{5}$$

where $0 < C < \min\{\tau_1^*, \cdots, \tau_n^*\}$ (the further construction can be also carried out for more general V).

If the submatrix A consists of the columns $i_1, \cdots, i_n$ and the rows $j_1, \cdots, j_n$ of J, then one defines η by putting the variables $\tau_1, \cdots, \tau_n$ in place of $s^*_{i_1}, \cdots, s^*_{i_n}$ and $\P$ is the projection onto the coordinates $y_{j_1}, \cdots, y_{j_n}$.

Now we can define our realization $\Sigma = (V, f, h, x_0)$. We take as the state space the set V defined above. Take $x_0 = \tau^* \in V$. We shall denote the state by τ instead of x. Define

$$h(\tau) = \mathbf{P}(\boldsymbol{\beta}(\eta(\tau))), \qquad \tau \in V.$$

From (A1) it follows that h is of class C^k. For $a = (t_1, \alpha_1) \cdots (t_p, \alpha_p) = \boldsymbol{\alpha}(t)$ define

$$\Phi_a(\tau) = \Psi^{-1} \circ \P \circ \Psi^{\mathbf{b}}_{\boldsymbol{\beta}, a} \circ \eta(\tau),$$

where

$$\Psi^{\mathbf{b}}_{\boldsymbol{\beta}, a}(\mathbf{s}) = (\mathbf{P}(\boldsymbol{\beta}(\mathbf{s}) a b_1), \cdots, \mathbf{P}(\boldsymbol{\beta}(\mathbf{s}) a b_q)). \tag{6}$$

Note that $\Psi^{\mathbf{b}}_{\boldsymbol{\beta}, e} = \Psi^{\mathbf{b}}_{\boldsymbol{\beta}, \boldsymbol{\alpha}(0)} = \Psi^{\mathbf{b}}_{\boldsymbol{\beta}}$.

The expression $\Phi_a(\tau) = \Phi_{\boldsymbol{\alpha}(\mathbf{t})}(\tau)$ is well defined for $(\tau, \mathbf{t})$ in a neighborhood W of the set $V \times \{0\} \subset R^n \times R^p_+$, defined by the condition $\P \circ \Psi^{\mathbf{b}}_{\boldsymbol{\beta}, a} \circ \eta(\tau) \subset \Psi(V)$. In particular, $\Phi_{\boldsymbol{\alpha}(0)}$ is the identity on V. The function $(\tau, \mathbf{t}) \to \Phi_{\boldsymbol{\alpha}(t)}(\tau)$ is of class C^k on W since $\Psi^{-1}, \P$ are of class C^k and the function $(\mathbf{t}, \mathbf{s}) \to \Psi^{\mathbf{b}}_{\boldsymbol{\beta}, \boldsymbol{\alpha}(\mathbf{t})}(\mathbf{s})$ is of class C^k (this follows from assumption (A1)).

Assume that we have proved that

$$\Phi_{a'} \circ \Phi_a(\tau) = \Phi_{aa'}(\tau) \tag{7}$$

for $a = \boldsymbol{\alpha}(\mathbf{t})$, $a' = \boldsymbol{\alpha}'(\mathbf{t}')$ and $(\tau, \mathbf{t}, \mathbf{t}')$ in a neighborhood of the set $V \times \{0\} \times \{0\}$ in $R^n \times R^p_+ \times R^{p'}_+$. Then we obtain that

$$\Phi_{(t, \alpha)} \circ \Phi_{(t', \alpha)}(\tau) = \Phi_{(t', \alpha)(t, \alpha)}(\tau) = \Phi_{(t'+t, \alpha)}(\tau)$$

for (τ, t, t') in a neighborhood of the set $V \times \{0\} \times \{0\}$ in $R^n \times R_+ \times R_+$. This together with $\Phi_{(0, \alpha)} = \mathrm{id}_V$ implies that $\Phi_{(t, \alpha)}$ is a local (semi) flow of class C^k. Thus, we can define the corresponding vector field

$$f(\tau, \alpha) = \frac{d}{dt} \Phi_{(t, \alpha)}(\tau)\big|_{t=0^+}, \qquad \tau \in V, \tag{8}$$

which is of class C^{k-1}. This completes the construction of the realization Σ.

From the construction it follows that $\Phi^f_{(t, \alpha)}(\tau) = \Phi_{(t, \alpha)}(\tau)$ for (τ, t) in a neighborhood of the set $V \times \{0\} \subset R^n \times R_+$. This and successive application of (7) gives that

$$\Phi^f_{(t_p, \alpha_p)} \circ \cdots \circ \Phi^f_{(t_1, \alpha_1)}(\tau) = \Phi_{(t_1, \alpha_1) \cdots (t_p, \alpha_p)}(\tau) \tag{9}$$

for $(\tau, \mathbf{t})$ in a neighborhood of the set $V \times \{0\} \subset R^n \times R^p_+$.

To prove that Σ is a local realization of P after the function c and to show (7) we use the following lemma.

LEMMA 1. *For any α, α' there exists a neighborhood N of the set $V\times\{0\}\times\{0\}$ in $R^n\times R_+^p\times R_+^{p'}$ such that for $a=\alpha(\mathbf{t})$, $a'=\alpha'(\mathbf{t}')$ and $(\tau,\mathbf{t},\mathbf{t}')\in N$ we have that*

$$\Psi^{\mathbf{b}}_{\beta,a'}\circ\eta\circ\Phi_a(\tau)=\Psi^{\mathbf{b}}_{\beta,aa'}\circ\eta(\tau), \tag{10}$$

$$\Psi^{\mathbf{b}}_{\beta,a'}\circ\eta\circ\Phi_a(\tau(\mathbf{s}))=\Psi^{\mathbf{b}}_{\beta,aa'}(\mathbf{s}). \tag{10}'$$

To show (7) it is enough to compose both sides of (10) with $\Psi^{-1}\circ\P$ on the left and use the definition of Φ_a.

Taking $\tau=\tau^*$, $a'=e$ in (10) and considering the component of the functions $\Psi^{\mathbf{b}}_{\beta}$ and $\Psi^{\mathbf{b}}_{\beta,a}$ corresponding to $b_i=e$, we obtain from the lemma that

$$\mathbf{P}(\beta(\eta(\Phi_a(\tau^*)))=\mathbf{P}(\beta(\eta(\tau^*))a) \tag{11}$$

i.e. $h\circ\Phi_a(\tau^*)=\mathbf{P}(ca)$ for $a=\alpha(\mathbf{t})$ and $\mathbf{t}$ in a neighborhood of $0\in R_+^p$. This together with (9) means that Σ is a local realization of P after the function c. The realization around c is obtained using equality (10)′ instead of (10). We postpone the proof of the second claim in Theorem 1 after the proof of Lemma 1.

Remark 2. In the above construction we used the lengths of time intervals of the input for the local coordinates of our realization. One can also use components of the output for the local coordinates, taking $\hat{V}=\Psi(V)$, $\hat{x}_0=\Psi(\tau^*)$ and

$$\hat{h}(x)=\mathbf{P}(\beta(\eta(\Psi^{-1}(x)))),$$

$$\hat{\Phi}_a(x)=\P\circ\Psi^{\mathbf{b}}_{\beta,a}\circ\eta\circ\Psi^{-1}(x).$$

Then $\hat{\Sigma}=(\hat{V},\hat{f},\hat{h},\hat{x}_0)$ is also a realization of P after the function c, where $\hat{f}$ is obtained from $\hat{\Phi}_a$ via (8). This construction is, perhaps, more natural when P is given by results of experiments which are points in the output space $Y=R^r$.

Lemma 1 will follow from the following lemma. For $C>0$, $\delta>0$ define the sets

$$W_{C,\delta}=\{(z,x,y)\in R^n\times R_+^p\times R_+^{p'}\,|\,|z-z_0|+\delta|x|+\delta|y|\leqq C\},$$

$$V_{C,\delta}=\{(z,x)\in R^n\times R_+^p\,|\,|z-z_0|+\delta|x|\leqq C\},\qquad V_C=\{z\in R^n\,|\,|z-z_0|\leqq C\}.$$

LEMMA 2. *Let $\varphi\colon W_{C,\delta}\to R^m$ be a function of class C^k, $k=1,2,\cdots,\infty,\omega$. Assume that*

$$\operatorname{rank} d\varphi=\operatorname{rank}\frac{\partial\varphi}{\partial z}=n=\operatorname{rank}\left(\frac{\partial\Psi}{\partial z},\frac{\partial\Psi}{\partial x}\right) \tag{12}$$

on $W_{C,\delta}$, where $\Psi\colon W_{C,\delta}\to R^{2m}$ is the function

$$\Psi(z,x,y)=(\varphi(z,x,0),\ \varphi(z,x,y))$$

and $(\partial\Psi/\partial z,\partial\Psi/\partial x)$ denotes its partial Jacobian with respect to z and x. Then, there exists a constant $0<\lambda\leqq\delta$ and a function $\gamma\colon V_{C,\lambda}\to R^n$ such that the equation

$$\varphi(\tilde{z},0,y)=\varphi(z,x,y) \tag{13}$$

has the solution $\tilde{z}=\gamma(z,x)$ for all $(z,x,y)\in W_{C,\lambda}$. If the function $z\to\varphi(z,0,0)$ is injective on V_C, then the function γ is unique and of class C^k.

Additionally, if each component φ_i, $i=1,\cdots,m$, of φ satisfies the Lipschitz condition on $W_{C,\delta}$

$$|\varphi_i(z,x,y)-\varphi_i(z',x',y')|\leqq L(|z-z'|+|x-x'|+|y-y'|) \tag{14}$$

and the same condition is satisfied for the partial derivatives $(\partial/\partial z_j)\varphi_i(z, x, y)$, $j = 1, \cdots, n$, *with constant* M, *then* λ *can be taken*

$$\lambda = \max\left\{\delta, (\Delta_{\min})^{-1}(n!)^2 L^{2n}\binom{m}{n} n(1+2CML^{-1})\right\}, \tag{15}$$

where $\Delta_{\min}$ *is the infimum of the sum of squares of all* n*-minors of* $(\partial/\partial z)\varphi(z, 0, 0)$ *over* $z \in V_C$.

The same result holds for the spaces R^p_+, $R^{p'}_+$ *replaced by* R^p *and* $R^{p'}$. *The lemma also holds if the condition* $\operatorname{rank} \partial\varphi/\partial z = n$ *is satisfied on* $W_{C,\delta} \cap \{(z, x, y) | x = 0, y = 0\}$ *only.*

Proof. Let $I = (i_1, \cdots, i_n)$ be any sequence of numbers $1 \leqq i_1 < \cdots < i_n \leqq m$. Denote all such sequences by $I_1, \cdots, I_\mu$, $\mu = \binom{m}{n}$. Define a function $\varphi_I : W_{C,\delta} \to R^n$,

$$\varphi_I(z, x, y) = \varphi_I(z, x) = (\varphi_{i_1}(z, x, 0), \cdots, \varphi_{i_n}(z, x, 0)). \tag{16}$$

Let $V_i \subset W_{C,\delta}$ be the set of points on which the function $\Delta_i = \det(\partial/\partial z)\varphi_{I_i}$ is different from zero. From the last equality of (12) and the definition of Ψ it follows that, on V_i,

$$\frac{\partial\Psi}{\partial z}\left(\frac{\partial}{\partial z}\varphi_{I_i}\right)^{-1}\frac{\partial}{\partial x}\varphi_{I_i} - \frac{\partial\Psi}{\partial x} = 0, \qquad i = 1, \cdots, \mu. \tag{17}$$

The second equality in (12) gives that the function $\Delta = \Delta_1^2 + \cdots + \Delta_\mu^2$ is nonzero on $W_{C,\delta}$. Define $\tilde{\Delta}_i = \Delta_i^2\Delta^{-1}$. Then we have that $\tilde{\Delta}_1 + \cdots + \tilde{\Delta}_\mu = 1$. Note that $\tilde{\Delta}_i((\partial/\partial z)\varphi_{I_i})^{-1}$ is well defined and of class C^k on $W_{C,\delta}$. Multiplying equations (17) by $\tilde{\Delta}_i$ and summing them up over i, we obtain that

$$\frac{\partial\Psi}{\partial z}A - \frac{\partial\Psi}{\partial x} = 0,$$

on $W_{C,\delta}$, where A is the matrix

$$A = \{A_{ij}\} = \sum_{r=1}^{\mu}\tilde{\Delta}_r\left(\frac{\partial}{\partial z}\varphi_{I_r}\right)^{-1}\frac{\partial}{\partial x}\varphi_{I_r}.$$

In particular, we have that

$$\sum_{i=1}^{n}\frac{\partial\Psi}{\partial z_i}A_{ij} - \frac{\partial\Psi}{\partial x_j} = 0, \qquad j = 1, \cdots, p.$$

This means that the function Ψ is constant along trajectories of the vector fields

$$g_j = \sum_{i=1}^{n} A_{ij}\frac{\partial}{\partial z_i} - \frac{\partial}{\partial x_j}, \qquad j = 1, \cdots, p.$$

which are of class C^{k-1} on $W_{C,\delta}$. Denote the local flow of g_j by Φ^j_t. Denote

$$D_j = \sup_{V_{C,\delta}}(|A_{1j}| + \cdots + |A_{nj}|), \qquad D = \max\{D_1, \cdots, D_p\}, \tag{18}$$

where the supremum over $V_{C,\delta}$ equals the supremum over $W_{C,\delta}$ since the functions φ_I given by (16) do not depend on y. Take $\lambda = \max\{D, \delta\}$. The vector fields g_j point in the set $W_{C,\lambda}$ (or are tangent to the boundary) at the boundary points (z, x, y) such that $|z - z_0| + \lambda|x| + \lambda|y| = C$ and $x_i \neq 0 \neq y_i$. Thus the trajectory $\Phi^j_t(z, x, y)$ of g_j starting in $W_{C,\delta}$ does not leave $W_{C,\lambda}$ until $t = x_j$. Combining p trajectories, we obtain, by the form of g_j, that

$$\Phi^p_{x_p} \circ \cdots \circ \Phi^1_{x_1}(z, x, y) = (\tilde{z}, 0, y) \tag{19}$$

where $(\tilde{z}, 0, y) \in W_{C,\lambda}$. So defined $\tilde{z}$ gives the solution $\tilde{z} = \tilde{\gamma}(z, x, y)$ of (13) for $(z, x, y) \in W_{C,\lambda}$ since the function Ψ is constant along the trajectories of g_j and in particular, $\Psi(\tilde{z}, 0, y) = \Psi(z, x, y)$. The function $\tilde{\gamma}$ is independent of y as the vector fields g_j do not depend on y.

The function $\gamma(z, x) = \tilde{\gamma}(z, x, 0)$ gives the desired solution. If the function $z \to \varphi(z, 0, 0)$ is injective, then the solution of $\varphi(\tilde{z}, 0, 0) = \varphi(z, x, 0)$ determines the function $\gamma(z, x)$ uniquely and assures that it is of class C^1 because, locally, $\gamma(z, x) = \Lambda_I^{-1} \circ \varphi_I(z, x, 0)$, where $\Lambda_I(z) = \varphi_I(z, 0, 0)$.

To prove the last part of the lemma, it is enough to majorize the constant D in (18) using constants L and M. The constant D defined as a supremum (18) over $V_{C,\lambda}$ (instead of $V_{C,\delta}$) is also good and we shall majorize such D computing at the same time λ. We have that

$$D \leqq n \max_{i,j} \sup_{V_{C,\lambda}} |A_{ij}| \tag{20}$$

and $\lambda = \max\{D, \delta\}$. From the definition of A it follows that

$$A = \Delta^{-1} \sum_{r=1}^{\mu} \Delta_r \left(\frac{\partial}{\partial z} \varphi_{I_r}\right)^* \frac{\partial}{\partial x} \varphi_{I_r},$$

where by B^* we denote the adjoint matrix of B (it consists of $(n-1)$-minors of B^T). From inequality (14) it follows that $|(\partial/\partial z_j)\varphi_i| \leqq L$ and $(\partial/\partial x_j)\varphi_i| \leqq L$, on $V_{C,\lambda}$. Therefore, from the definition of determinant it follows that $|\Delta_r| \leqq n! L^n$. Thus, we obtain easily that

$$|A_{ij}| \leqq \mu \Delta^{-1} n! L^n n(n-1)! L^{n-1} L = \mu (n!)^2 L^{2n} \Delta^{-1}. \tag{21}$$

To majorize Δ^{-1}, we shall compute the Lipschitz constant for $\Delta = \Delta_1^2 + \cdots + \Delta_\mu^2$. Note that nKN^{n-1} is the Lipschitz constant for the product $f_1 f_2 \cdots f_n$, if K is the Lipschitz constant for $f_1, \cdots, f_n$ and $N = \max_i \sup |f_i|$. Thus the definition of determinant and $|(\partial/\partial z_j)\varphi_i| \leqq L$ gives that the Lipschitz constant for Δ_r can be taken $n! nML^{n-1}$. Therefore, the Lipschitz constant for Δ is

$$L_\Delta = \mu 2n! nML^{n-1} n! L^N = 2\mu n (n!)^2 ML^{2n-1}.$$

Now we can majorize Δ^{-1} over $V_{C,\lambda}$ for $\lambda > \Delta_{\min}^{-1} L_\Delta C$

$$\sup_{V_{C,\lambda}} \Delta^{-1} = \left(\inf_{V_{C,\lambda}} \Delta\right)^{-1} \leqq \left(\inf_{V_C} \Delta - L_\Delta C \lambda^{-1}\right)^{-1} = (\Delta_{\min} - L_\Delta C \lambda^{-1})^{-1}. \tag{22}$$

Taking into account (20), (21) and (22), we obtain that

$$D \leqq n\mu (n!)^2 L^{2n} (\Delta_{\min} - C\lambda^{-1} 2\mu n (n!)^2 ML^{2n-1})^{-1}.$$

If we put $\lambda = D$, we obtain

$$D \leqq \Delta_{\min}^{-1} (n!)^2 L^{2n} \binom{m}{n} n(1 + 2CML^{-1}).$$

This means that λ given by (15) satisfies the assertions of the lemma.

The proof in case of R_+^p, $R_+^{p'}$ replaced by R^p, $R^{p'}$ is analogous.

If the condition $\operatorname{rank} \partial\varphi/\partial z = n$ is satisfied on $W_{C,\delta} \cap \{(z, x, y) | x = 0, y = 0\}$ only, then Δ is positive on this set. From its compactness it follows that Δ is also positive on a set $W_{C,\delta'}$ with a $\delta' \geqq \delta$. Thus (12) is satisfied on $W_{C,\delta'}$. In the Lipschitzian case it is enough to take any $\delta' \geqq \delta$ such that $\delta' > \Delta_{\min}^{-1} CL_\Delta$.

Proof of Lemma 1. Consider the function

$$\varphi(\tau, \mathbf{t}, \mathbf{t}') = \Psi^{\mathbf{b}}_{\beta, \alpha(\mathbf{t})\alpha'(\mathbf{t}')} \circ \eta(\tau). \tag{23}$$

We shall see that φ satisfies the assumptions of Lemma 1 with $z=\boldsymbol{\tau}$, $x=\mathbf{t}$, $y=\mathbf{t}'$ and $z_0=\boldsymbol{\tau}^*$. In fact, φ is well defined on the set $W_{C,\delta}$, where C is taken as in the definition (5) of the set V and $\delta=1$. Since $\varphi(\boldsymbol{\tau},0,0)=\Psi^{\mathbf{b}}_{\boldsymbol{\beta}}\circ\eta(\boldsymbol{\tau})$, then it follows from the choice of V that $\operatorname{rank}\partial\varphi/\partial\boldsymbol{\tau}(\boldsymbol{\tau},\mathbf{t},\mathbf{t}')=n$ for $\boldsymbol{\tau}\in\operatorname{cl}V$, $\mathbf{t}=0$, $\mathbf{t}'=0$ and so for $(\boldsymbol{\tau},\mathbf{t},\mathbf{t}')\in W_{C,\delta}$ with some $\delta\geqq 1$.

From condition (A1) it follows that φ is of class C^k.

We have that $\operatorname{rank}\partial\varphi/\partial\boldsymbol{\tau}\leqq\operatorname{rank}d\varphi\leqq\operatorname{rank}P$ and $\operatorname{rank}\partial\varphi/\partial\boldsymbol{\tau}\leqq\operatorname{rank}(\partial\Psi/\partial\boldsymbol{\tau},\partial\Psi/\partial\mathbf{t})\leqq\operatorname{rank}P$, where Ψ is defined as in Lemma 2. The last inequality above follows from the definition of rank P and the equality

$$\Psi(\boldsymbol{\tau},\mathbf{t},\mathbf{t}')=\Psi^{\mathbf{bb}'}_{\boldsymbol{\beta}\boldsymbol{\alpha}}(\eta(\boldsymbol{\tau}),\mathbf{t}),$$

where $\mathbf{bb}'=(b_1,\cdots,b_q,a'b_1,\cdots,a'b_q)$, $a'=\boldsymbol{\alpha}'(\mathbf{t}')$ and $\boldsymbol{\beta}\boldsymbol{\alpha}=(\beta_1,\cdots,\beta_m,\alpha_1,\cdots,\alpha_p)$. The above inequalities together with $\operatorname{rank}\partial\varphi/\partial\boldsymbol{\tau}=\operatorname{rank}P$ on $W_{C,\delta}$ imply that φ satisfies condition (12) in Lemma 2.

From Lemma 2 we deduce that there exists a unique function $\gamma(\boldsymbol{\tau},\mathbf{t})$ such that

$$\Psi^{\mathbf{b}}_{\boldsymbol{\beta},\boldsymbol{\alpha}'(\mathbf{t}')}\circ\eta(\gamma(\boldsymbol{\tau},\mathbf{t}))=\Psi^{\mathbf{b}}_{\boldsymbol{\beta},\boldsymbol{\alpha}(\mathbf{t})\boldsymbol{\alpha}'(\mathbf{t}')}\circ\eta(\boldsymbol{\tau}) \tag{24}$$

for $(\boldsymbol{\tau},\mathbf{t},\mathbf{t}')$ in a neighborhood of the set $U\times\{0\}\times\{0\}$ in $R^n\times R^p_+\times R^{p'}_+$. Taking $\mathbf{t}'=0$, we obtain a simpler equality

$$\Psi^{\mathbf{b}}_{\boldsymbol{\beta}}\circ\eta(\gamma(\boldsymbol{\tau},\mathbf{t}))=\Psi^{\mathbf{b}}_{\boldsymbol{\beta},a}\circ\eta(\boldsymbol{\tau}),\qquad a=\boldsymbol{\alpha}(\mathbf{t}). \tag{25}$$

Composing both sides with $\Psi^{-1}\circ\P$ on the left and taking into account the definitions of Ψ and Φ_a, we find that

$$\gamma(\boldsymbol{\tau},\mathbf{t})=\Phi_a(\boldsymbol{\tau})$$

and so formula (10) is proved.

The proof of formula (10)′ is analogous, except that we take the function φ of the form

$$\varphi(\boldsymbol{\tau},\boldsymbol{\tau}',\mathbf{t},\mathbf{t}')=\Psi^{\mathbf{b}}_{\boldsymbol{\beta},\boldsymbol{\alpha}(\mathbf{t})\boldsymbol{\alpha}'(\mathbf{t}')}(\boldsymbol{\tau},\boldsymbol{\tau}'+s'^*),$$

where $\boldsymbol{\tau}'=\mathbf{s}'-\mathbf{s}'^*$ and $z=\boldsymbol{\tau}$, $x=(\boldsymbol{\tau}',\mathbf{t})$, $y=\mathbf{t}'$, $z_0=\boldsymbol{\tau}^*$, where $\mathbf{s}'$ denotes the $m-n$ variables in $\mathbf{s}$ complementary to $\boldsymbol{\tau}$ (see the definition of the function η at the beginning of the proof of Theorem 1). Then instead of equality (25) we obtain

$$\Psi^{\mathbf{b}}_{\boldsymbol{\beta}}\circ\eta(\gamma(\boldsymbol{\tau},\boldsymbol{\tau}',\mathbf{t}))=\Psi^{\mathbf{b}}_{\boldsymbol{\beta},a}(\boldsymbol{\tau},\boldsymbol{\tau}'+\mathbf{s}'^*),\qquad a=\boldsymbol{\alpha}(\mathbf{t}),$$

and so

$$\begin{aligned}\gamma(\boldsymbol{\tau},\boldsymbol{\tau}',\mathbf{t})&=\Psi^{-1}\circ\pi\circ\Psi^{\mathbf{b}}_{\boldsymbol{\beta},a}(\boldsymbol{\tau},\boldsymbol{\tau}'+\mathbf{s}'^*)\\&=\Psi^{-1}\circ\pi\circ\Psi^{\mathbf{b}}_{\boldsymbol{\beta},a}(\boldsymbol{\tau}(\mathbf{s}),\mathbf{s}'^*)\\&=\Phi_a(\boldsymbol{\tau}(\mathbf{s})),\qquad \mathbf{s}=(\boldsymbol{\tau},\boldsymbol{\tau}'+\mathbf{s}'^*).\end{aligned}$$

Here the implicit function $\boldsymbol{\tau}(\mathbf{s})$ exists since the function $\Psi^{-1}\circ\pi\circ\Psi^{\mathbf{b}}_{\boldsymbol{\beta},a}$ is regular (of full rank) with respect to the first argument. The proof of Lemma 1 is complete.

Now we shall prove the second claim in Theorem 1. Condition (A2) implies that the function φ in (23) satisfies the Lipschitz condition with constant L and the partial derivatives $\partial\varphi/\partial\tau_i$ satisfy this condition with constant M. Therefore, we can apply formula (15) in Lemma 2 to deduce that λ does not depend on the sequences $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}'$ in (23). Therefore, equality (10) in Lemma 1 holds for $(\boldsymbol{\tau},\mathbf{t},\mathbf{t}')$ in $W_{C,\lambda}$, in particular, for $(\boldsymbol{\tau}^*,\mathbf{t},\mathbf{t}')$ such that $|\mathbf{t}|+|\mathbf{t}|'\leqq C\lambda^{-1}=\sigma$ where C is taken as in (5). This implies that (7) holds for such $\mathbf{t},\mathbf{t}'$ with $\mathbf{a}=\boldsymbol{\alpha}(\mathbf{t})$, $a'=\boldsymbol{\alpha}'(\mathbf{t}')$ and $\boldsymbol{\tau}=\boldsymbol{\tau}^*$. A successive application of

(7) together with $\Phi^f_{(t,\alpha)} = \Phi_{(t,\alpha)}$ give (9) with $t_1 + \cdots + t_p \leqq \sigma$ and $\tau = \tau^*$. We have also that (11) holds for $|a| \leqq \sigma = C\lambda^{-1}$, so we obtain that the realization Σ constructed in the course of proving Theorem 1 is a realization of P on the interval $[0, \sigma]$, where $\sigma = C\lambda^{-1}$. If we take into account that $\varphi(\tau, 0, 0) = \Psi^b_\beta \circ \eta(\tau)$, then formula (15) gives the following formula for σ

$$\sigma = \min\{C, \sigma_1\}$$

where

$$\sigma_1 = \Delta_{min}(n(n!)^2\binom{m}{n}L^{2n})^{-1}(C^{-1} + 2ML^{-1})^{-1}, \qquad n = \operatorname{rank} P,$$

where $\Delta_{\min}$ is the infimum over cl V of the sum of squares of all the n-minors of the Jacobian matrix $d(\Psi^b_\beta \circ \eta)$.

Condition (A2) and the definition of Φ_a imply also that the map $\Phi_{(t,\alpha)}(\tau)$ satisfies the Lipschitz condition with respect to t with a constant independent of α and $\tau \in V$ (note that Ψ^{-1} and $\P$ are Lipschitzian with the constants $L_1 = \sup_{x \in \operatorname{cl} V} \|(d\Psi(x))^{-1}\|$ and $L_2 = 1$). Thus, definition (8) of $f(\tau, \alpha)$ gives that f is bounded by the constant qLL_1. The Lipschitz property of f can be proved analogously. The proof of Theorem 1 is complete.

Remark 3. In the proof of Theorem 1 we did not use any other property of the expression Φ_a besides the properties following from Lemma 1. Therefore, (10) can be used for the definition of Φ_a. Also, due to Lemma 2 and the proof of Lemma 1, the following simpler equation

$$\Psi^b_\beta \circ \eta(\Phi_a(\tau)) = \Psi^b_{\beta,a} \circ \eta(\tau)$$

determines uniquely $\Phi_a(\tau)$. The definition of Φ_a by the above equation gives, in general, a bigger domain V for our realization. Namely, here τ should be in a set V on which the map $\Psi^b_\beta \circ \eta$ is injective (to have uniqueness), while previously we required that there exists a projection $\P$ such that $\P \circ \Psi^b_\beta \circ \eta$ is a diffeomorphism on V.

Proof of Theorem 2. *Necessity.* Let (V, f, h, x_0) be a realization of P of class C^ω and dimension n. Let $\sigma > 0$ be a constant such that V contains the closed ball B of radius σ and the center at x_0. Define

$$\rho(\alpha) = \sup_{x \in B} \|f(x, \alpha)\|,$$

where $\|\cdot\|$ is the Euclidean norm in R^n. Then the solution of the equation

$$\dot{x} = f(x, a), \qquad x(0) = x_0$$

is well defined and stays in B for $t \leqq |a|$, for any $a = (t_1, \alpha_1) \cdots (t_p, \alpha_p)$ such that

$$|a|_\rho = \rho(\alpha_1)|t_1| + \cdots + \rho(\alpha_p)|t_p| < \sigma.$$

Moreover, $\Phi^f_a(x_0) \in B$ for such a and $h \circ \Phi^f_a(x_0) = h \circ \Phi^f_{(t_p,\alpha_p)} \circ \cdots \circ \Phi^f_{(t_1,\alpha_1)}(x_0)$ is analytic with respect to $(t_1, \cdots, t_p)$. Thus condition (A1)′ is satisfied.

Define functions Ψ_α and Ψ^b by $\Psi_\alpha(\mathbf{t}) = \Phi^f_{\alpha(\mathbf{t})}(x_0)$, $\Psi^b(x) = (h \circ \Phi^f_{b_1}(x), \cdots, h \circ \Phi^f_{b_q}(x))$. One can easily see that $\Psi^b_\alpha(\mathbf{t}) = \Psi^b \circ \Psi_\alpha$. Thus, $\operatorname{rank} d\Psi^b_\alpha(\mathbf{t}) \leqq \operatorname{rank} d\Psi_\alpha(\mathbf{t}) \leqq n$ and we conclude that rank P is finite.

Sufficiency. If (A1)′ is satisfied, then the expressions

$$P((t_1, \alpha_1) \cdots (t_p, \alpha_p)) \tag{26}$$

are well defined for $\mathbf{t} = (t_1, \cdots, t_p) \in E^\rho_{\sigma,\alpha}$. Now the argument $(t_1, \alpha_1) \cdots (t_p, \alpha_p) = \boldsymbol{\alpha}(\mathbf{t})$ is not treated as a function but as a finite sequence with values in $R \times \Omega$. We multiply

such sequences by concatenation. Before constructing the realization we note that

$$(27)\qquad P(a'(t_1, \alpha_1)\cdots(t_p, \alpha_p)(-t_p, \alpha_p)\cdots(-t_p, \alpha_p)\cdots(-t_1, \alpha_1)a'') = P(a'a'')$$

where $a' = \boldsymbol{\alpha}'(\mathbf{t}')$, $a'' = \boldsymbol{\alpha}''(\mathbf{t}'')$, and $\mathbf{t}' \in R^{p'}$, $\mathbf{t}'' \in R^{p''}$, $\mathbf{t} = (t_1, \cdots, t_p)$ are such that $2|\boldsymbol{\alpha}(\mathbf{t})|_\rho + |a'|_\rho + |a''|_\rho < \delta$. For $p = 1$ this follows from the following equalities, by taking $\hat{t}_1 = -t_1$ in the first one,

$$P(a'(t_1, \alpha_1)(\hat{t}_1, \alpha_1)a'') = P(a'(t_1 + \hat{t}_1, \alpha_1)a''),$$

$$P(a'(0, \alpha)a'') = P(a'a'').$$

Each of these equalities is true for $\mathbf{t}' \in R_+^{p'}$, $\mathbf{t}'' \in R_+^{p''}$ and $t_1, \hat{t}_1 \geqq 0$ since both arguments represent the same functions. They extend to all $\mathbf{t}'$, $\mathbf{t}''$, t_1, $\hat{t}_1$ by the analyticity. The proof of (27) for general p follows by induction with respect to p.

The extension of the domain of the expressions (26) gives that the functions $\Psi_\alpha^{\mathbf{b}}$ are well defined on the sets $E_{\lambda,\alpha}^\rho \subset R^p$, $\lambda = \delta - \max\{|b_1|_\rho, \cdots, |b_q|_\rho\}$, where b_i are finite sequences with values in $R \times \Omega$ such that $|b_i|_\rho < \delta$. The function $\Psi_{\beta,a}^{\mathbf{b}}(\mathbf{s})$ given by (6) is well defined for $a = \boldsymbol{\alpha}(\mathbf{t})$ and $\mathbf{s} \in R^m$, $\mathbf{t} \in R^p$ such that $|a|_\rho + |\boldsymbol{\beta}(\mathbf{s})|_\rho \leqq \delta - \max\{|b_1|_\rho, \cdots, |b_q|_\rho\}$. Thus, the whole construction in the proof of Theorem 1 can be carried out with replacing R_+ by R, and replacing the positive cones by the whole spaces.

Let rank $d\Psi_\beta^{\mathbf{b}}(s^*) = \text{rank } P$, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_m)$. From the analyticity of the function $\Psi_\beta^{\mathbf{b}}$ it follows that this equality is satisfied for $\mathbf{s}^* \in R_+^m$ arbitrarily close to 0, in particular, we can assume that $|\boldsymbol{\beta}(\mathbf{s}^*)|_\rho < \frac{1}{3}\lambda$. Take the sequences $\boldsymbol{\beta}' = (\beta_1, \cdots, \beta_m, \beta_m, \cdots, \beta_1)$ and $\mathbf{b}' = (cb_1, \cdots, cb_q)$, where $c = (s_1^*, \beta_1)\cdots(s_m^*, \beta_m)$. Consider the function $\Psi_{\beta'}^{\mathbf{b}'} = \Psi_{\beta'}^{\mathbf{b}'}(\mathbf{s}, \hat{\mathbf{s}})$. From (27) and the definition of the function $\Psi_{\beta'}^{\mathbf{b}'}$ it follows that

$$\Psi_{\beta'}^{\mathbf{b}'}(\mathbf{s}, \hat{\mathbf{s}}^*) = \Psi_\beta^{\mathbf{b}}(\mathbf{s}),$$

where $\hat{\mathbf{s}}^* = (-s_m^*, \cdots, -s_1^*)$. Therefore,

$$\text{rank } d\Psi_{\beta'}^{\mathbf{b}'}(\mathbf{s}^*, \hat{\mathbf{s}}^*) = \text{rank } d\Psi_\beta^{\mathbf{b}}(\mathbf{s}^*) = \text{rank } P.$$

Now we can construct our realization Σ as in the proof of Theorem 1, starting with the function $\Psi_{\beta'}^{\mathbf{b}'}$ instead of $\Psi_\beta^{\mathbf{b}}$ and taking the sequence

$$c' = (s_1^*, \beta_1)\cdots(s_m^*, \beta_m)(-s_m^*, \beta_m)\cdots(-s_1^*, \beta_1)$$

instead of $c = (s_1^*, \beta_1)\cdots(s_m^*, \beta_m)$. From the form of c' and property (27) it follows that $P(c'a) = P(a)$, so such realization Σ is a realization after the identity e. The proof is complete.

Remark 4. From the analyticity it follows that condition (iii) is satisfied for $a = \boldsymbol{\alpha}(\mathbf{t})$ for all $\mathbf{t} \in R^p$ such that $P(a)$ and $\Phi_a^f(x_0)$ are well defined.

Proof of Theorem 3. Consider the map $P_\sigma = P|_{\mathscr{S}_{pc}^\sigma}$ and restrict all further considerations to the domain $\mathscr{S}_{pc}^\sigma$. Let $c = \boldsymbol{\beta}(s^*)$ and $\mathbf{b}$ be such that $|c| + |b_i| \leqq \sigma$ and rank $d\Psi_\beta^{\mathbf{b}}(\mathbf{s}^*) = \text{rank } P_\sigma$. Define functions η and $\P$ as at the beginning of the proof of Theorem 1. Then the function

$$\Psi = \P \circ \Psi_\beta^{\mathbf{b}} \circ \eta$$

is a local diffeomorphism defined on a neighborhood of the point $\boldsymbol{\tau}^* \in R^n$, where $\eta(\boldsymbol{\tau}^*) = \mathbf{s}^*$ and $n = \text{rank } P_\sigma = \text{rank}_{\text{loc}} P$. Define functions Ψ_β and $\Psi^{\mathbf{b}}$ by

$$\Psi_\beta(\mathbf{s}) = \Phi_{\beta(\mathbf{s})}^f(x_0), \qquad \Psi^{\mathbf{b}}(x) = (h \circ \Phi_{b_1}^f(x), \cdots, h \circ \Phi_{b_q}^f(x))$$

and, analogously, functions $\hat{\Psi}_\beta$ and $\hat{\Psi}^{\mathbf{b}}$ corresponding to the second realization $\hat{\Sigma}$.

Then $\Psi_{\beta}(\mathbf{s}^*) = x_1 = \Phi_c^f(x_0)$, $\hat{\Psi}_{\beta}(\mathbf{s}^*) = \hat{x}_1 = \Phi_c^{\hat{f}}(\hat{x}_0)$ and we have that

$$\Psi^{\mathbf{b}} \circ \Psi_{\beta}(\mathbf{s}) = \Psi_{\beta}^{\mathbf{b}}(\mathbf{s}) = \hat{\Psi}^{\mathbf{b}} \circ \hat{\Psi}_{\beta}(\mathbf{s}),$$

and so

$$\P \circ \Psi^{\mathbf{b}} \circ \Psi_{\beta} \circ \eta(\tau) = \Psi(\tau) = \P \circ \hat{\Psi}^{\mathbf{b}} \circ \hat{\Psi}_{\beta} \circ \eta(\tau)$$

for τ in a neighborhood of $\tau^* \in R^n$. From the minimality of $\sum$ it follows that $\dim \sum = n$, i.e. the maps $\Psi_{\beta} \circ \eta$ and $\P \circ \Psi^{\mathbf{b}}$ are "into R^n" and "from a subset of R^n". Since their composition Ψ is a local diffeomorphism of R^n, they are local diffeomorphisms too. Define the local diffeomorphism

$$\chi = \hat{\Psi}_{\beta} \circ \eta \circ (\Psi_{\beta} \circ \eta)^{-1} = (\P \circ \hat{\Psi}^{\mathbf{b}})^{-1} \circ \P \circ \Psi^{\mathbf{b}}.$$

Then we have that $\chi(x_1) = \hat{x}_1$. From the definition of functions $\Psi^{\mathbf{b}}$, Ψ_{β}, $\hat{\Psi}^{\mathbf{b}}$, $\hat{\Psi}_{\beta}$ and definition (6) of $\Psi_{\beta,a}^{\mathbf{b}}$ it follows that

$$\P \circ \Psi^{\mathbf{b}} \circ \Phi_a^f \circ \Psi_{\beta} \circ \eta = \P \circ \Psi_{\beta,a}^{b} \circ \eta = \P \circ \hat{\Psi}^{\mathbf{b}} \circ \Phi_a^{\hat{f}} \circ \hat{\Psi}_{\beta} \circ \eta$$

locally around τ^*, for $a = \alpha(\mathbf{t})$ with $\mathbf{t}$ sufficiently small. Composing the above expressions with $(\P \circ \hat{\Psi}^{\mathbf{b}})^{-1}$ on the left and $(\Psi_{\beta} \circ \eta)^{-1}$ on the right, we obtain

$$\chi \circ \Phi_a^f = \Phi_a^{\hat{f}} \circ \chi.$$

Putting $a = (t, \alpha)$ and taking the derivative of both sides with respect to t at $t = 0$ gives the second equality in (3).

We have also that

$$h \circ \Psi_{\beta} \circ \eta(\tau) = P(\beta(\eta(\tau))) = \hat{h} \circ \hat{\Psi}_{\beta} \circ \eta(\tau)$$

for τ in a neighborhood of τ^*. Composing both sides with $(\Psi_{\beta} \circ \eta)^{-1}$ on the right gives that $h = \hat{h} \circ \chi$ in a neighborhood of x_1. The proof is complete.

Proof of Theorem 4. Assumptions (A1), (A2) and Theorem 1 imply that there exists a realization Σ of class C^k of the map $P|_{\mathscr{S}_{pc}}$ after a function c, on an interval $[0, \delta]$. From the definition of Φ_a in the proof of Theorem 1 and from (A3) it follows that the function $\Phi_{(t,\alpha)}(\tau)$ is continuous with respect to (τ, t, α) together with the partial derivative with respect to t. This and definition (8) of $f(\tau, \alpha)$ imply that f is continuous with respect to (τ, α).

Condition (A2) implies, additionally, that $f(\tau, \alpha)$ is bounded and Lipschitzian with respect to τ. To see the latter, it is enough to use the definitions of Φ_a and $f(\tau, \alpha)$ and to note that the maps Ψ^{-1}, $\P$ and η are Lipschitzian together with the first derivatives.

From the continuity and the Lipschitz property of f it follows that, if functions a_i, a have values in a compact subset of Ω and $a_i \to a$ pointwise almost everywhere, then $\Phi_{a_i}^f(\tau) \to \Phi_a^f(\tau)$ provided all the terms are well defined. Let $a \in \mathscr{S}$ be any function such that $|a| \leqq \delta$. There exists a sequence of functions $a_i \in \mathscr{S}_{pc}$, $|a_i| = |a|$ such that $a_i \to a$ pointwise. From the fact that Σ is a realization of $P|_{\mathscr{S}_{pc}}$, the continuity property mentioned above and condition (A4) it follows that $P(ca) = h \circ \Phi_a^f(\tau^*)$. Thus Σ is a realization of P after the function c, on the internal $[0, \delta]$.

The proof of the second claim in Theorem 4 is analogous to the proof of Theorem 2. If condition (A3) is strengthened to the second order partial derivatives, then $f(\tau, \alpha)$ is continuous together with the first order partial derivatives with respect to τ. This and the compactness of Ω implies the Lipschitz property of f.

Acknowledgment. The author is grateful to Professor Cz. Olech for helpful discussions and comments concerning the paper.

REFERENCES

[1] R. W. BROCKETT, *Volterra series and geometric control theory*, Automatica, 12 (1976), pp. 167-176.
[2] P. E. CROUCH, *Dynamical realizations of finite Volterra series*, this Journal, 19 (1981), pp. 177-202.
[3] P. D'ALESSANDRO, A. ISIDORI AND A. RUBERTI, *Realization and structure theory of bilinear dynamical systems*, this Journal, 12 (1974), pp. 517-535.
[4] M. FLIESS, *Realizations of nonlinear systems and abstract transitive Lie algebras*, Bull. Amer. Math. Soc. (N.S.), 2 (1980), pp. 444-446.
[5] ———, *Réalisation locale des systèmes non linéaires, algèbres de Lie filtrées transitives et séries génératrices non commutatives*, Inv. Math., 71 (1983), pp. 521-533.
[6] J. P. GAUTHIER AND G. BORNARD, *Existence and uniqueness of minimal realizations in the C^∞ case*, Syst. Contr. Letters, 1 (1982), pp. 395-398.
[7] B. JAKUBCZYK, *Existence and uniqueness of nonlinear realizations*, in Analyse des Systèmes, Proc. Conf. Bordeaux 1978, Astérisque, 75-76 (1980), pp. 141-147.
[8] ———, *Existence and uniqueness of realizations of nonlinear systems*, this Journal, 18 (1980), pp. 455-471.
[9] ———, *Construction of formal and analytic realizations of nonlinear systems*, in Feedback Control of Linear and Nonlinear Systems, Hinrichsen and Isidori, eds., Springer, Berlin, 1982, pp. 147-156.
[10] ———, *Réalisations locales des opérateures causales non linéaires*, C.R. Acad. Sci. Paris, 299 Ser. I (1984), pp. 787-789.
[11] B. JAKUBCZYK AND B. KAŚKOSZ, *Realizability of Volterra series with constant kernels*, Nonlinear Anal., Th. Meth. Appl., 5 (1980), pp. 167-183.
[12] E. D. SONTAG, *Polynomial Response Maps*, Springer, Berlin, 1979.
[13] H. J. SUSSMANN, *A generalization of the closed subgroup theorem to quotients of arbitrary manifolds*, J. Diff. Geom., 10 (1975), pp. 151-166.
[14] ———, *Existence and uniqueness of minimal realizations of nonlinear systems*, Math. Systems Theory, 10 (1977), pp. 263-284.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 2, March 1986

006

THE VALUE FUNCTION IN OPTIMAL CONTROL: SENSITIVITY, CONTROLLABILITY, AND TIME-OPTIMALITY*

FRANK H. CLARKE† AND PHILIP D. LOEWEN†

Abstract. We consider a general optimal control problem in which the constraints depend on a parameter α, and the resulting value function $V(\alpha)$. A formula for the generalized gradient of V is proven and then used to obtain results on stability and controllability of the problem. A special study is made of the time-optimal control problem, one consequence of which is a new criterion assuring local null-controllability of the system and continuity of the minimal time function at the origin.

Key words. sensitivity analysis, perturbation, generalized gradients, time-optimality, controllability

1. Introduction. The value function plays a central role in optimization. To illustrate this, let us consider a standard problem $\mathrm{P}(\alpha)$ in optimal control theory, indexed by a parameter α as follows:

$$\mathrm{P}(\alpha)\qquad \begin{aligned} &\text{minimize} && \int_0^1 L(x(t), u(t), \alpha)\, dt \\ &\text{subject to} && \dot{x}(t) = \varphi(x(t), u(t), \alpha), \\ & && u(t) \in U(\alpha), \\ & && x(0) \in C_0(\alpha), \quad x(1) \in C_1(\alpha). \end{aligned}$$

The minimum in the problem $\mathrm{P}(\alpha)$ defines $V(\alpha)$. The first role of V is evident: it measures the sensitivity of the problem (more precisely, of the problem's optimal outcome) to perturbations of the objective function and the various constraints. Knowledge of how V behaves is therefore useful in problems of design or error analysis. Particularly interesting is the derivative of V, a measure of what has been called "differential stability." In the quest for differential properties of V, however, we must face the fact that not only will V not generally be differentiable, it can easily fail to be continuous or even everywhere finite. (By the usual convention, the minimum over the empty set is $+\infty$, so that $V(\alpha) = +\infty$ whenever $\mathrm{P}(\alpha)$ admits no feasible (x, u).) The very fact that V is finite near a given α is in many cases a coveted conclusion, implying as it does a type of local controllability of the constraints.

In studying differential properties of V then, whether for sensitivity analysis, or to explain results such as the maximum principle geometrically, or in connection with the well-known role that V plays in verification techniques (i.e., the Hamilton-Jacobi equation as a sufficient condition, see [4, § 3.7]), or in constructing feedback laws (as in dynamic programming), one must consider derivatives in other than the classical sense.

Following a seminal paper by Gauvin [6], a great deal of progress has recently been made in studying the value function V of a perturbed mathematical programming problem based on its generalized gradient ∂V. (See [2], [4, § 6.5], [7], and especially Rockafellar [20], in which the issue is surveyed.) In [4, § 3.4], Clarke obtains a formula for ∂V for a general fixed-time control problem, subject however only to additive

* Received by the editors April 1, 1984, and in revised form November 15, 1984. The authors gratefully acknowledge the support of the Natural Sciences and Engineering Research Council of Canada under grant number 9082 and a 1967 Science Scholarship.

† CRM, Université de Montréal, Montréal, Québec, Canada H3C 3J7.

perturbations of the endpoint constraints. See [1], [3], [5], [13], [15], [18], [19] for results related to stability of control problems.

Our first purpose in this article is to obtain a formula for the generalized gradient of the value function of a general fixed- or free-time control problem in which all of the constraints (including the dynamics) are subject to (nonadditive) peturbation. Secondly, we will explore some of the major consequences of this characterization. In § 3 we derive a series of conditions under which, successively, V is locally finite (which, as mentioned above, has implications for controllability), locally Lipschitz, admits directional derivatives, or actually differentiable. Section 4 is devoted to an example illustrating the main theorem. In § 5 we undertake a special study of the time-optimal control problem. The special case of this problem in which the dynamics are linear and V reduces to the minimal time function T has been extensively studied [8], [9], [11], [16], [17]. Using the general theorem of § 3, we are able to extend to nonlinear systems the main results of the linear theory. In particular, we derive a new criterion assuring the continuity of T at the origin and the local null-controllability of the system.

We now proceed to describe the context in which we shall work.

The Problem. The object of our study is a relative of the standard optimal control problem known as the *differential inclusion problem*:

$$\text{(P)}\quad \min\{f(T, x(0), x(T)):\ \dot{x}(t)\in F(x(t)) \text{ a.e. on } [0, T], \text{ and } (T, x(0), x(T))\in S\}.$$

The objective of problem (P) is to choose a (nondegenerate) time interval $[0, T]$ and an arc (i.e. an absolutely continuous function) $x:[0, T]\to X\subseteq \mathbf{R}^n$ satisfying the endpoint constraints $(T, x(0), x(T))\in S$ and the dynamic constraint $\dot{x}(t)\in F(x(t))$ a.e. on $[0, T]$, and providing a minimum for the function $f(T, x(0), x(T))$. Here $F: X\to \mathbf{R}^n$ is a *multifunction*, a mapping which carries a point x in $\mathbf{R}^n$ into a nonempty compact subset $F(x)\subseteq \mathbf{R}^n$. To see the relationship between (P) and a more conventional optimal control problem, suppose the Mayer problem below is given:

$$\text{(Q)}\quad \begin{aligned}\min\{f(T, x(0), x(T)):\ &\dot{x}(t)=\varphi(x(t), u(t)) \text{ a.e. on } [0, T],\\ &\text{where } u(t)\in U \text{ a.e. on } [0, T], \text{ and } (T, x(0), x(T))\in S\}.\end{aligned}$$

We may then define a version of (P) with the same f and S, and with $F(x) := \varphi(x, U)$. Any arc admissible for (Q) is also admissible for (P), where it is assigned the same value. Conversely, reasonable hypotheses on φ imply that if $x:[0, T]\to \mathbf{R}^n$ is an admissible arc for (P), then there is a measurable function $u:[0, T]\to U$ such that $\dot{x}(t)=\varphi(x(t), u(t))$ a.e. on $[0, T]$. (This is Filippov's lemma. In the sequel, any φ and U for which $F(x)=\varphi(x, U)$ for all $x\in X$, will be termed a *classical representation* of F.) Thus the Mayer problem (Q) is a special case of (P). Of course, nonautonomous versions of both problems exist. The techniques presented below require very few modifications to be applied to such problems: we have chosen the simple form of (P) to ease the exposition.

Hypotheses. The following mild conditions govern the data comprising problem (P). Throughout this paper, B denotes the open unit ball of the appropriate Euclidean space.

(h_1) There are a closed set $X\subseteq \mathbf{R}^n$ and an $\varepsilon>0$ such that the multifunction F is defined on $X+\varepsilon B$; the values of F are nonempty compact convex subsets of $\mathbf{R}^n$.

(h_2) F is locally Lipschitz on $X+\varepsilon B$. (A multifunction $\Gamma:\mathbf{R}^m\to\mathbf{R}^m$ is *Lipschitz of rank* K on a set $C\subseteq \mathbf{R}^m$ if $\Gamma(x_2)\subseteq \Gamma(x_1)+K|x_2-x_1|B$ for all x_1, x_2 in C.)

(h_3) F obeys a linear growth condition: for some $c\geqq 0$, $k\geqq 0$, one has $F(x)\subseteq (k|x|+c)B$ for all $x\in X+\varepsilon B$.

(h_4) The constraint set $S \subseteq \mathbf{R} \times \mathbf{R}^n \times \mathbf{R}^n$ is a closed set, and $\{(t, x) : (t, x, y) \in S\}$ is compact.

(h_5) The objective function $f : S \to \mathbf{R}$ is locally Lipschitz on S.

Remark. The requirement in (h_1) that F be convex-valued means that we are working in this article with the "relaxed" problem in the sense of Warga [22]. (See also [4, § 5.5], in which classical representations of standard relaxed problems are discussed.) The state constraint $x(t) \in X$ is admitted to allow consideration of local minima. Our hypotheses will limit attention to arcs lying in int X, hence excluding binding "unilateral" state constraints.

The growth condition (h_3) allows us to use Gronwall's inequality to estimate the modulus of continuity of any trajectory for F. Here is the result.

LEMMA 1.1. *Suppose $x : [0, T] \to \mathbf{R}^n$ is an arc with $\dot{x}(t) \in F(x(t))$ a.e. on $[0, T]$. Then*

$$|x(t) - x(a)| \leqq \left(|x(a)| + \frac{c}{k}\right)(e^{k(t-a)} - 1)$$

for all $0 \leqq a \leqq t \leqq T$.

The following result concerning the sequential compactness of trajectories is crucial to the theory of existence of solutions to (P) and to the sensitivity analysis of Part 3. It is the autonomous case of [4, Thm. 3.1.7].

PROPOSITION 1.2. *Let $\Gamma : \mathbf{R}^m \to \mathbf{R}^m$ be an upper semicontinuous multifunction with nonempty compact convex values defined on a set of the form $C + \varepsilon B$, where $C \subseteq \mathbf{R}^m$ is closed and $\varepsilon > 0$. Let $x_j : [0, T_j] \to \mathbf{R}^m$ be a sequence of arcs for which $T_j \to T > 0$, and satisfying the following hypotheses*:

(i) *$x_j(t) \in C$ for all $t \in [0, T_j]$;*

(ii) *there is a constant $M > 0$ such that $\Gamma(x_j(t)) \subseteq M\bar{B}$ for all $t \in [0, T_j]$, and $|\dot{x}_j(t)| \leqq M$ a.e. on $[0, T_j]$;*

(iii) *$\dot{x}_j(t) \in \Gamma(x_j(t)) + r_j(t)B$ a.e. on $[0, T] \cap [0, T_j]$, where $\{r_j\}$ is a sequence of measurable functions on $[0, T]$ converging uniformly to zero*;

(iv) *the sequence $\{x_j(0)\}$ is bounded.*

There there is a subsequence of $\{x_j\}$ along which $x_j|_{[0,T]}$ converges uniformly to an arc $x : [0, T] \to C$ obeying $\dot{x}(t) \in \Gamma(x(t))$ a.e. on $[0, T]$. Along this subsequence, $x_j(T_j) \to x(T)$. (If $T_j < T$, then $x_j|_{[0,T]}$ should be interpreted as the extension $\tilde{x}_j$ of x_j which obeys $\tilde{x}_j(t) := x_j(T_j)$ for $t \in (T_j, T]$.)

Proof. Conditions (ii) and (iv) imply that $\tilde{x}_j = x_j|_{[0,T]}$ is a uniformly bounded equicontinuous family of functions, whence it has a uniformly convergent subsequence. By (i), the limit function x lies in C. Moreover, the Dunford-Pettis criterion applies (along the subsequence) to show that $\{\dot{\tilde{x}}_j\}$ is weakly compact in $L^1[0, T]$, so that x is actually an arc. Finally, a support function argument shows that $\dot{x}(t) \in F(x(t))$ a.e. on $[0, T]$, just as in [4]. □

The existence theory for problem (P) is a simple consequence of Proposition 1.2.

PROPOSITION 1.3. *Suppose that the data of problem* (P) *satisfy* (h_1)-(h_5), *and that there is at least one feasible arc for* (P). *Suppose also that degenerate arcs are ruled out, i.e.*

$$(h_6) \qquad S \cap \{(0, x, x) : x \in \mathbf{R}^n\} = \varnothing.$$

Then problem (P) *has a solution.*

Proof. Let $x_j : [0, T_j] \to \mathbf{R}^n$ be a sequence of feasible arcs for (P) such that

$$f(T_j, x_j(0), x_j(T_j)) \to \inf (\mathrm{P}) < +\infty.$$

Then $\{(T_j, x_j(0))\}$ lies in a compact set by (h_4), so we may pass to a subsequence if necessary and assume that $x_j(0) \to x_0 \in \mathbf{R}^n$ and $T_j \to T \geqq 0$.

Note that $T = 0$ is impossible, since Lemma 1.1 implies that

$$|x_j(T_j) - x_0| \leqq |x_j(T_j) - x_j(0)| + |x_j(0) - x_0| \leqq \left(|x_j(0)| + \frac{c}{k}\right)(e^{kT_j} - 1) + |x_j(0) - x_0|.$$

So if $T_j \to 0$, then $x_j(T_j) \to x_0$ and the sequence $\{(T_j, x_j(0), x_j(T_j))\}$ in the closed set S would converge to a point $(0, x_0, x_0)$ outside S by (h_6). This is absurd: we must have $T > 0$.

Hypotheses (i), (ii), (iv) of Proposition 1.2 are now evident; condition (iii) follows readily from Lemma 1.1 and the growth condition (h_3). We deduce that a subsequence of $\{x_j\}$ converges to a feasible arc $x: [0, T] \to \mathbf{R}^n$ which solves (P). □

2. Techniques. This section summarizes the techniques of nonsmooth analysis which are central to the sensitivity results of § 3.

Generalized gradients. The calculus of generalized gradients, presented in detail in [4], has a well-developed geometric aspect which is the foundation of the arguments in § 3.

Let C be a nonempty closed subset of $\mathbf{R}^m$, with $x \in C$. The vector $v \in \mathbf{R}^m$ is *perpendicular to C at x*, written $v \perp C$ at x, if $|(x+v) - x| \leqq |(x+v) - c|$ for all $c \in C$, with strict inequality for $c \neq x$. (I.e. the closed ball of radius $|v|$ and centre $x+v$ meets C in the single point x.) In terms of the Euclidean distance function $d_C(z) := \inf\{|z - c| : c \in C\}$, $v \perp C$ at x if x is the unique point of C at which the infimum defining $d_C(x+v)$ is attained. This hints at a profitable link between perpendicularity and a problem of minimization.

LEMMA 2.1. *Let C be a nonempty closed subset of $\mathbf{R}^m$, with $x \in C$. If $v \perp C$ at x, then $\langle v, c - x\rangle \leqq \frac{1}{2}|c - x|^2$ for all c in C.*

Proof. Write the definition of $v \perp C$ at x in terms of the inner product. □

The collection of all perpendiculars to C at base points near x defines the *normal cone to C at x*, denoted $N_C(x)$, as follows: $N_C(x)$ is the closed convex cone generated by the set

$$\left\{\lim_{i \to \infty} \frac{v_i}{|v_i|} : v_i \to 0,\ v_i \perp C \text{ at } x_i \to x\right\} \cup \{0\}.$$

The first set in this union consists of all unit vectors which can be obtained by computing limits of normalized perpendiculars: the perpendiculars must diminish in length and their corresponding base points x_i in C must converge to the given point x.

Now let an extended-real-valued function f be defined on $\mathbf{R}^m$. Suppose $f(x)$ is finite, and that epi f is locally closed near $(x, f(x))$. Then the *generalized gradient of f at x* is the (closed, convex) set

$$\partial f(x) = \{\zeta \in \mathbf{R}^m : (\zeta, -1) \in N_{\text{epi } f}(x, f(x))\}.$$

Extremely bad behaviour of f near x is captured by the *asymptotic generalized gradient of f at x*, namely

$$\partial^\infty f(x) = \{\zeta \in \mathbf{R}^m : (\zeta, 0) \in N_{\text{epi } f}(x, f(x))\}.$$

Note that $\partial^\infty f(x)$ is a closed convex cone containing the origin. It reduces to $\{0\}$ if and only if f is Lipschitz near x ([4, Prop. 2.9.7]).

Necessary conditions. The calculus of generalized gradients allows the formulation of general first-order necessary conditions for problem (P) without any assumptions

of smoothness or differentiability beyond (h_1)-(h_5). The principal figure in these conditions is the *Hamiltonian* $H: \mathbf{R}^n \times \mathbf{R}^n \to \mathbf{R}$, defined by

$$H(x, p) := \sup \{\langle p, v\rangle : v \in F(x)\}.$$

When $F(x)$ admits a classical representation $\varphi(x, U)$, H is a function sometimes referred to as the "maximized Hamiltonian" in the literature:

$$H(x, p) := \sup \{\langle p, \varphi(x, u)\rangle : u \in U\}.$$

THEOREM 2.2 (Necessary Conditions). *Suppose that the arc* $x: [0, T] \to \operatorname{int} X$ *solves* (P). *Then for all* r *sufficiently large, there exist a scalar* $\lambda \in \{0, 1\}$, *an arc* $p: [0, T] \to \mathbf{R}^n$, *and a vector* $\zeta \in \mathbf{R} \times \mathbf{R}^n \times \mathbf{R}^n$ *such that conclusions* (a)-(d) *below hold with* $\lambda + \|p\| > 0$.

(a) $(-\dot{p}(t), \dot{x}(t)) \in \partial H(x(t), p(t))$ *a.e. on* $[0, T]$.

(b) *There is a constant* h *such that* $H(x(t), p(t)) = h$ *for all* $t \in [0, T]$.

(c) $\zeta \in \partial f(T, x(0), x(T))$.

(d) $(h, p(0), -p(T)) \in \lambda\zeta + r\, \partial d_S(T, x(0), x(T))$.

Proof. Evidently $x: [0, T] \to \mathbf{R}^n$ is an interior solution for (P) if and only if the arc $(x(0), x(\cdot)): [0, T] \to \mathbf{R}^n \times \mathbf{R}^n$ is an interior solution for the related problem

$$\min \{f(\tau, y_0(\tau), y(\tau)) : (\dot{y}_0(t), \dot{y}(t)) \in \{0\} \times F(y) \text{ a.e. on } [0, \tau],$$
$$(y_0(0), y(0)) \in D \cap M\bar{B}, (\tau, y_0(\tau), y(\tau)) \in S\},$$

where $D = \{(y, y) : y \in \mathbf{R}^n\}$ and $M > 0$ is chosen so large that MB includes the whole second component of S (which can be done by (h_4)). This related problem has precisely the form for which [4, Thm. 3.6.1, Corollary], provides necessary conditions. A straightforward translation of terms yields the version of the result stated above. □

Remarks. Since (in general notation) $N_C(z)$ is the closed convex cone generated by $\partial d_C(z)$, condition (c) is often replaced at very slight expense by

$$(c') \qquad (h, p(0), -p(T)) \in \lambda\zeta + N_S(T, x(0), x(T)).$$

The pair (p, ζ) satisfying conditions (a), (b), and (c') is called an *index* λ *multiplier* corresponding to x. The multiplier is *normal* if $\lambda = 1$ and *abnormal* if $\lambda = 0$; if no solution of (P) admits a nontrivial abnormal multiplier, we say that problem (P) itself is *normal.* It follows from condition (a) [4, Prop. 3.2.4(b)] that the arc p satisfies $|\dot{p}(t)| \leqq K|p(t)|$ for some constant K. This observation, together with Gronwall's lemma, implies that an arc p satisfying (a) is either identically zero or else nonvanishing on $[0, T]$.

We refer to [4] for a complete discussion of the relationship between this theorem and the celebrated Maximum Principle of Pontryagin.

3. General sensitivity analysis. We are now prepared to investigate the sensitivity of the differential inclusion problem (P) to small changes in its objective function, dynamics, and constraints. Specifically, we will compute the generalized gradient at 0 of the *value function* $V: \mathbf{R}^m \to \mathbf{R} \cup \{+\infty\}$ defined in terms of the following perturbed optimization problem:

$$\text{P}(\alpha) \qquad V(\alpha) := \min\{f(T, x(0), x(T), \alpha) : \dot{x}(t) \in F(x(t), \alpha) \quad \text{a.e. on } [0, T],$$
$$\text{and } (T, x(0), x(T), \alpha) \in S\}.$$

The main result, presented in Theorem 3.3 below, is that $\partial V(0)$ is captured by an extension of the multiplier rule, Theorem 2.2, which incorporates elements which monitor the effects of the perturbation along the optimal arc.

Multipliers. An *index* λ *multiplier* corresponding to an admissible arc $x: [0, T] \to \operatorname{int} X$ for P(0) is a triple (p, q, ζ) satisfying conditions (a)-(d) below.

(a) $(p, q): [0, T] \to \mathbf{R}^n \times \mathbf{R}^m$ is an arc obeying the *Hamiltonian inclusion*

$$(-\dot{p}(t), -\dot{q}(t), \dot{x}(t)) \in \partial \mathscr{H}(x(t), 0, p(t)) \quad \text{a.e. on } [0, T],$$

where the *Hamiltonian* for the perturbed problem $\mathrm{P}(\alpha)$ is

$$\mathscr{H}(x, \alpha, p) := \sup \{\langle p, v\rangle : v \in F(x, \alpha)\}.$$

(b) There is a constant h such that $\mathscr{H}(x(t), 0, p(t)) = h$ for all t in $[0, \mathrm{T}]$.

(c) $\zeta \in \partial f(T, x(0), x(T), 0)$.

(d) $(h, p(0), -p(T), -q(T)) \in \lambda\zeta + N_S(T, x(0), x(T), 0)$.

The collection of all such triples is the set $M^\lambda(x)$, the *index* λ *multiplier set corresponding to* x. We shall later be concerned with the set Y of all solutions to P(0) and the corresponding collection of multipliers

$$M^\lambda(Y) = \bigcup_{x \in Y} M^\lambda(x).$$

Hypotheses. The following modifications of hypotheses (h_1)-(h_6) above ensure that for each $\alpha \in \mathbf{R}^n$ near the nominal value $\alpha = 0$, problem $\mathrm{P}(\alpha)$ is amenable to the techniques of §§ 1 and 2. In particular, they imply that whenever $V(\alpha) < +\infty$, then $\mathrm{P}(\alpha)$ has a solution. We retain them throughout § 3.

(H1) The $\mathbf{R}^n$-valued multifunction F is defined on $X \times \varepsilon B \subseteq \mathbf{R}^n \times \mathbf{R}^m$ for some $\varepsilon > 0$, where $X \subseteq \mathbf{R}^n$ is closed. The values of F are nonempty compact convex sets.

(H2) The multifunction F is (jointly) locally Lipschitz on $X \times \varepsilon B$.

(H3) There are constants $c \geqq 0$, $k \geqq 0$ such that $F(x, \alpha) \subseteq (k|x| + c)B$ for all $(x, \alpha) \in X \times \varepsilon B$.

(H4) The constraint set $S \subseteq \mathbf{R} \times \mathbf{R}^n \times \mathbf{R}^n \times \mathbf{R}^m$ is closed, and its projection $\{(t, x) : (t, x, y, \alpha) \in S\}$ is compact.

(H5) The objective function $f: S \to \mathbf{R}$ is locally Lipschitz on S.

(H6) Degenerate arcs are inadmissible, i.e.

$$S \cap \{(0, x, x, \alpha) : x \in X, \alpha \in \varepsilon \bar{B}\} = \varnothing.$$

In addition to the hypotheses stating that $\mathrm{P}(\alpha)$ is well-behaved for each fixed α, we demand that the perturbation structure of the problem vary reasonably as α changes. This requires three more hypotheses.

(H7) Problem P(0) has a feasible arc. By Proposition 1.2, it follows that P(0) has a solution. We also assume that every arc solving P(0) lies entirely inside int X.

(H8) For every arc $x: [0, T] \to \mathbf{R}^n$ in Y, the multifunction $(t, x, y, \alpha) \to N_S(t, x, y, \alpha)$ is closed at the point $(T, x(0), x(T), 0)$.

(H9) The perturbation structure is *nondegenerate*, i.e. every triple (p, q, ζ) in $M^0(Y)$ with $q(0) = 0$ also has $p \equiv 0$.

To understand (H9), suppose first that the unperturbed data f, F, and S of problem (P) in § 1 are forced into the shape of $\mathrm{P}(\alpha)$ by imposing trivial α-dependence. Then an arc $x: [0, T] \to \mathbf{R}^n$ is admissible for P(0) if and only if it is admissible for (P), and every index λ multiplier (p, q, ζ_0) corresponding to x must have $q \equiv 0$ and $\zeta_0 = (\zeta, 0) \in \partial f(T, x(0), x(T)) \times \{0\}$. These multipliers are in one-to-one correspondence with the mulipliers (p, ζ) for (P) defined by conclusions (a), (b), (c'), (d) of Theorem 2.2. Hence in the case of trivial α-dependence, (H9) reduces to the assumption that P(0) is normal. If $\mathrm{P}(\alpha)$ has nontrivial α-dependence, (H9) is strictly weaker than the assumption that P(0) is normal because of the additional condition $q(0) = 0$. Indeed, there are perturba-

tion schemes (such as the additive endpoint perturbation studied in [4, § 3.4]) which are always nondegenerate but in which the normality of p(0) is determined by other features of the problem's data.

Hypothesis (H8) is a mild condition which is satisfied automatically if,for example, the cone $N_S(T, x(0), x(T), 0)$ is pointed. In many cases, such as the additive endpoint perturbations studied in [4, § 3.4] or the minimal time problem of § 5, (H8) can be verified directly. A detailed discussion of conditions which imply (H8) is given in [14].

Hypothesis (H7) is required because problems whose solution arcs spend some time on the boundary of X obey a more complicated multiplier rule than that described by Theorem 2.2. (This amounts to allowing abstract state constraints in problem (P): the interested reader may consult [4, Thm. 3.6.1].) The proof techniques presented below can be extended to this setting, but (H7) eases the exposition considerably. Proposition 3.2 below shows that (H7) implies that the perturbed problem is *tame* in the sense of [20].

LEMMA 3.1. *Assume* (H1)-(H6). *Then there exists a positive constant* T_0 *such that if* $x:[0, T]\to\mathbf{R}^n$ *is admissible for some* $\mathrm{P}(\alpha)$, $\alpha\in\varepsilon B$, *then* $T\geqq T_0$.

Proof. If this statement were false, then there would be a sequence $\{\alpha_i\}$ in εB with corresponding trajectories $x_i:[0, T_i]\to\mathbf{R}^n$ admissible for $\mathrm{P}(\alpha_i)$ while $T_i\to 0$. By passing to a subsequence if necessary, we may assume $x_i(0)\to x_0$ (by (H4)) and $\alpha_i\to\alpha\in\varepsilon\bar{B}$. But then Lemma 1.1 would imply that $x_i(T_i)\to x_0$, and we would reach the absurd conclusion that the sequence $\{(T_i, x_i(0), x_i(T_i), \alpha_i)\}$ in the closed set S converges to a point $(0, x_0, x_0, \alpha)$ which lies outside S by (H6). □

PROPOSITION 3.2. *Assume* (H1)-(H7). *Then there is some* $\delta\in(0, \varepsilon)$ *such that* $\alpha\in\delta B$ *and* $V(\alpha)<V(0)+\delta$ *imply that all solutions of* $\mathrm{P}(\alpha)$ *lie in* int X.

Proof. Suppose not. Then there is a sequence $\{\alpha_j\}$ with $|\alpha_j|<\frac{1}{j}$ and $V(\alpha_j)<V(0)+\frac{1}{j}$ for which there exist corresponding solutions $x_j:[0, T_j]\to\mathbf{R}^n$ of $\mathrm{P}(\alpha_j)$ which all fail to lie in int X. By passing to a subsequence if necessary, we may assume that $x_j(0)\to x_0$ and $T_j\to T>0$. According to Proposition 1.2 (with $\Gamma(x):=F(x, 0)$ and $r_j(t):=K|\alpha_j|$ for some fixed $K>0$), a further subsequence (which we do not relabel) converges uniformly to an arc $x:[0, T]\to\mathbf{R}^n$ admissible for P(0) and obeying

$$V(0)\leqq f(T, x(0), x(T), 0)\leqq\lim_{j\to\infty} f(T_j, x_j(0), x_j(T_j), \alpha_j)$$

$$=\lim_{j\to\infty} V(\alpha_j)\leqq\lim_{j\to\infty}\left(V(0)+\frac{1}{j}\right)=V(0).$$

Hence the arc x is a solution of P(0). But this contradicts (H7), since any arc which is the uniform limit of arcs not contained in int X cannot lie in int X. □

Theorem 3.3 below is the main result of this section. It relates the differential properties of V to the arcs q in the multiplier sets introduced above. For any index λ multiplier (p, q, ζ) corresponding to an arc $x:[0, T]\to\mathbf{R}^n$ for P(0), we define $Q(p, q, \zeta):=-q(0)$. The notation $Q[M^\lambda(x)]$ designates the set of all possible values of $-q(0)$ obtained in this way, and $Q[M^\lambda(Y)]$ denotes $\cup_{x\in Y} Q[M^\lambda(x)]$. The proof of Theorem 3.3 occupies the remainder of § 3.

THEOREM 3.3. *Under hypotheses* (H1)-(H9), *the function* V *is lower semicontinuous near* 0, *and one has*

$$\partial V(0)=\overline{\mathrm{co}}\,\{Q[M^1(Y)]\cap\partial V(0)+Q[M^0(Y)]\cap\partial^\infty V(0)\}.$$

If the cone $Q[M^0(Y)]$ *is pointed, the closure operation is superfluous and one also has*

$$\partial^\infty V(0)=\mathrm{co}\,\{Q[M^0(Y)]\cap\partial^\infty V(0)\}.$$

(*A subset of* $\mathbf{R}^m$ *is* pointed *if zero cannot be obtained as a positive linear combination of its nonzero elements.*)

COROLLARY 1 (Lipschitz behaviour of V). *If* $Q[M^0(Y)]=\{0\}$, *then* V *is finite and Lipschitz near* 0. *In particular, if* P(0) *is normal* (*i.e.*, $Q[M^0(Y)]=\{0\}$) *then* P(0) *is locally controllable, in the sense that* P(α) *admits feasible arcs for all* α *near* 0.

Proof. The cone $\{0\}$ is certainly pointed, so $\partial^\infty V(0)=\{0\}$ by Theorem 3.3. This implies that V is finite and Lipschitz near 0 [4, Prop. 2.9.7]. □

COROLLARY 2 (Existence of nontrivial multipliers). *If* $x:[0, T]\to\mathbf{R}^n$ *solves* P(0) *then it has an index* λ *multiplier* (p, q, ζ) *with* $\lambda+|q(0)|>0$.

Proof. If $Y=\{x\}$ the proof is simple: either $Q[M^0(x)]$ is different from $\{0\}$, whence the conclusion is immediate, or $Q[M^0(x)]=\{0\}$. In the latter case, $\partial^\infty V(0)=\{0\}$ implies that $\partial V(0)$ is nonempty [4, Prop. 2.9.7]: thus

$$\varnothing\neq\partial V(0)=\operatorname{co}\{Q[M^1(x)]\cap\partial V(0)\},$$

and we find $M^1(x)\neq\varnothing$.

If x is not the only solution to P(0), a bookkeeping device allows the formulation of a problem related to P(0) for which x is the unique solution. Transposition of the preceding argument into this case yields the desired result, as in [4, Thm. 3.5.2]. □

The following immediate consequences of the formulas in the theorem are proven just as in [4, Thm. 6.5.2, Corollaries 2, 3]. As in that reference, we use the notation V', V^+, V_+ to denote (respectively) the usual one-sided directional derivative, upper right and lower right Dini derivatives (see [4, p. 242]).

COROLLARY 3 (Directional derivatives of V). *Suppose that* $Q[M^0(Y)]=\{0\}$. *Then one has for each* u *in* $\mathbf{R}^m$:

$$V^+(0; u)\leq\inf_{x\in Y}\sup\langle u, Q[M^1(x)]\rangle,$$

$$V_+(0; u)\geqq\inf_{x\in Y}\inf\langle u, Q[M^1(x)]\rangle.$$

If $Q[M^1(x)]$ *is a singleton* $Q(x)$ *for each* x *in* Y, *then* $V'(0; u)$ *exists for each* u *in* $\mathbf{R}^m$ *and one has*

$$V'(0; u)=\inf_{x\in Y}\langle Q(x), u\rangle.$$

COROLLARY 4 (Differentiability of V). *If* Y *is a singleton* $\{x\}$, *and if for this* x *one has* $Q[M^0(x)]=\{0\}$ *and* $Q[M^1(x)]=\{\zeta\}$, *then* V *is strictly differentiable at* 0 *with* $DV(0)=\zeta$. (*Strict differentiability is defined in* [4, § 2.2].)

Proof of Theorem 3.3. To prove Theorem 3.3, we will calculate the generalized gradients of V by appealing to the definitions given in § 2. The first step is to establish that epi V is locally closed near $(0, V(0))$.

LEMMA 3.4. *The function* V *is lower semicontinuous on* δB, *where* δ *is given by Proposition* 3.2.

Proof. Fix any α in δB. Given any sequence $\alpha_j\to\alpha$, we must show that $V(\alpha)\leqq\liminf_{j\to\infty}V(\alpha_j)$. If the right side is $+\infty$, there is nothing to prove; otherwise, we may pass to a subsequence if necessary and assume that $\lim_{j\to\infty}V(\alpha_j)$ exists. Let $x_j:[0, T_j]\to\mathbf{R}^n$ solve P(α_j): Proposition 1.2 implies that a subsequence of these arcs converges to an F-trajectory $x:[0, T]\to\mathbf{R}^n$ admissible for P(α), with $T\geqq T_0>0$ by Lemma 3.1. Then

$$V(\alpha)\leqq f(T, x(0), x(T), \alpha)=\lim_{j\to\infty}f(T_j, x_j(0), x_j(T_j), \alpha_j)=\lim_{j\to\infty}V(\alpha_j),$$

as required. □

The proof of Theorem 3.3 is built on the following geometric proposition of Rockafellar [20, Prop. 15].

PROPOSITION 3.5. *Let D and D^∞ be closed subsets of $\mathbf{R}^m$ such that D^∞ is a cone containing the point 0 and the recession cone of D. For the closed cone*

$$N := \{r(\zeta, -1) : r > 0,\ \zeta \in D\} \cup \{(\zeta, 0) : \zeta \in D^\infty\}$$

in $\mathbf{R}^m \times \mathbf{R}$, one has

$$\{\zeta : (\zeta, -1) \in \overline{\text{co}}\ N\} = \overline{\text{co}}\ (D + D^\infty).$$

If the cone D^∞ is pointed, the closure operation on the right-hand side is superfluous, and one also has

$$\{\zeta : (\zeta, 0) \in \overline{\text{co}}\ N\} = \text{co}\ D^\infty.$$

To complete the proof, we must successfully identify D with $Q[M^1(Y)] \cap \partial V(0)$ and D^∞ with $Q[M^0(Y)] \cap \partial^\infty V(0)$. Both these sets are closed, being the intersections of closed sets. ($Q[M^\lambda(Y)]$ is closed by Proposition 1.2.) Moreover, D^∞ is the intersection of two cones containing 0, and is therefore such a cone itself. Thus we only need to show that

$$N_{\text{epi}\, V}(0, V(0)) = \overline{\text{co}}\ (N_1 \cup N_2),$$

where

$$N_1 := \{r(\zeta, -1) : r > 0,\ \zeta \in Q[M^1(Y)] \cap \partial V(0)\}$$
$$N_2 := \{(\zeta, 0) : \zeta \in Q[M^0(Y)] \cap \partial^\infty V(0)\},$$

and then check that D^∞ is the recession cone of D.

The definitions of $\partial V(0)$ and $\partial^\infty V(0)$ imply that

$$N_{\text{epi}\, V}(0, V(0)) \supseteq \overline{\text{co}}\ (N_1 \cup N_2).$$

To prove the reverse inclusion, we will use the definition of the normal cone in terms of limits of normalized perpendicular vectors given in § 2. It motivates the following lemma.

LEMMA 3.6. *Suppose $(\beta, -u) \perp \text{epi}\ V$ at (α, v), where $v < V(0) + \delta$ and $\alpha \in \delta B$. Then there exists a solution $x : [0, T] \to \mathbf{R}^n$ to $\text{P}(\alpha)$ (for some $T \geqq T_0$) to which there correspond a scalar $\lambda \in \{0, 1\}$, an arc $(p, q) : [0, T] \to \mathbf{R}^n \times \mathbf{R}^m$, and a vector $\zeta \in \mathbf{R} \times \mathbf{R}^n \times \mathbf{R}^n \times \mathbf{R}^m$ such that $\|(p, q)\| + \lambda / |(\beta, -u)| > 0$ and* (a)-(e) *hold.*

(a) $(-\dot{p}(t), -\dot{q}(t), \dot{x}(t)) \in \partial \mathcal{H}(x(t), \alpha, p(t))$ *a.e. on* $[0, T]$.

(b) *There is a constant h such that $\mathcal{H}(x(t), \alpha, p(t)) = h$ on* $[0, T]$.

(c) $\zeta = (\tau, \xi, \eta, \eta_1) \in \partial f(T, x(0), x(T), \alpha)$.

(d) $(h, p(0), -p(T), -q(T)) \in \lambda (u / |(\beta, -u)|) \zeta + N_S(T, x(0), x(T), \alpha)$.

(e) $q(0) = -\lambda (\beta / |(\beta, -u)|)$.

Proof. By Proposition 3.2, $V(\alpha) \leqq v < V(0) + \delta$ implies that problem $\text{P}(\alpha)$ has a solution $x : [0, T] \to \mathbf{R}^n$ lying entirely inside int X. Now for any $\alpha' \in \delta \bar{B}$, every F-trajectory $y : [0, T'] \to \mathbf{R}^n$ admissible for $\text{P}(\alpha')$ obeys

$$\begin{aligned} V(\alpha') &\leqq f(T', y(0), y(T'), \alpha') \\ &\leqq f(T', y(0), y(T'), \alpha') + v - f(T, x(0), x(T), \alpha), \end{aligned}$$

i.e. $(\alpha', f(T', y(0), y(T'), \alpha') + v - f(T, x(0), x(T), \alpha)) \in \text{epi } V$. Replacing c in Lemma 2.1 with the left side of this expression yields the inequality

$$uf(T, x(0), x(T), \alpha) - \langle \alpha, \beta \rangle \leqq uf(T', y(0), y(T'), \alpha') - \langle \alpha', \beta \rangle + \tfrac{1}{2}|(\alpha - \alpha', f(T', y(0), y(T'), \alpha') - f(T, x(0), x(T), \alpha))|^2.$$

Note that equality holds when $\alpha' = \alpha$ and $y = x$ (so $T' = T$). Thus $(x(\cdot), \alpha)$ is an interior solution for the (unperturbed) problem in which

$$\tilde{f}(t, x, x_1, y, y_1) := uf(t, x, y, y_1) - \langle \beta, x_1 \rangle + \tfrac{1}{2}|(y_1 - \alpha, f(t, x, y, y_1) - f(T, x(0), x(T), \alpha))|^2,$$

$$\tilde{F}(x, x_1) := F(x, x_1) \times \{0\},$$

$$\tilde{S} := \{(t, x, x_1, y, y_1) : (t, x, y, y_1) \in S, x_1 \in \delta \bar{B}\}.$$

The Hamiltonian for this new problem is

$$\tilde{H}(x, x_1, p, q) = \sup \{\langle (p, q), (v, 0) \rangle : v \in F(x, x_1)\} = \mathcal{H}(x, x_1, p).$$

Hypotheses (h_1)-(h_5) of § 1 are easy to verify, so by Theorem 2.2, there exist a scalar $\lambda \in \{0, 1\}$, an arc $(p, q) : [0, T] \to \mathbf{R}^n \times \mathbf{R}^m$, and a vector $\tilde{\zeta} \in \mathbf{R} \times \mathbf{R}^n \times \mathbf{R}^m \times \mathbf{R}^n \times \mathbf{R}^m$ such that $\lambda + \|(p, q)\| > 0$ and conclusions (i)-(iv) below hold.

(i) $(-\dot{p}(t), -\dot{q}(t), \dot{x}(t), 0) \in \mathcal{H}(x(t), \alpha, p(t)) \times \{0\}$ a.e. on $[0, T]$.

(ii) There is a constant h such that $\mathcal{H}(x(t), \alpha, p(t)) = h$ on $[0, T]$.

(iii) $\tilde{\zeta} = u(\tau, \xi, 0, \eta, \eta_1) + (0, 0, -\beta, 0, 0)$ for some $(\tau, \xi, \eta, \eta_1)$ in $\partial f(T, x(0), x(T), \alpha)$.

(iv) $(h, p(0), q(0), -p(T), -q(T)) \in \lambda \tilde{\zeta} + N_{\tilde{S}}(T, x(0), \alpha, x(T), \alpha)$.

Condition (iv) implies that $q(0) = -\lambda\beta$ and

(iv′) $(h, p(0), -p(T), -q(T)) \in \lambda u(\tau, \xi, \eta, \eta_1) + N_S(T, x(0), x(T), \alpha)$.

Replacing the arc (h, p, q) by the scaled version $(h, p, q)/|(\beta, -u)|$ yields the desired conclusions. □

Suppose now that a sequence $(\beta_i, -u_i)$ of vectors perpendicular to epi V at corresponding points (α_i, v_i) is given, where $(\beta_i, -u_i) \to (0, 0)$, $(\alpha_i, v_i) \to (0, V(0))$, and the limit of $(\beta_i, -u_i)/|(\beta_i, -u_i)|$ exists and equals (say) (β_0, u_0). Since $N_{\text{epi } V}(0, V(0))$ is precisely the closed convex cone generated by such limit vectors, we need only prove $(\beta_0, -u_0) \in N_1 \cup N_2$. This is a consequence of the following lemma.

LEMMA 3.7. *Let $(\beta, -u_0)$ be the unit vector introduced above. Then there is a solution $x : [0, T] \to \mathbf{R}^n$ to* P(0), *an arc (p, q), and a vector $\zeta \in \mathbf{R} \times \mathbf{R}^n \times \mathbf{R}^n \times \mathbf{R}^m$ such that*

(a) $(-\dot{p}(t), -\dot{q}(t), \dot{x}(t)) \in \partial \mathcal{H}(x(t), 0, p(t))$ *a.e. on* $[0, T]$,

(b) *there is a constant h such that $\mathcal{H}(x(t), 0, p(t)) = h$ on* $[0, T]$.

(c) $\zeta = (\tau, \xi, \eta, \eta_1) \in \partial f(T, x(0), x(T), 0)$,

(d) $(h, p(0), -p(T), -q(T)) \in u_0 \zeta + N_S(T, x(0), x(T), 0)$,

(e) $q(0) = -\beta_0$.

Proof. For all i sufficiently large, Lemma 3.6 applies to the perpendicular vectors $(\beta_i, -u_i)$ to provide solutions $x_i : [0, T_i] \to \mathbf{R}^n$ (with $T_i \geqq T_0 > 0$) for $\text{P}(\alpha_i)$ with corresponding quantities $\lambda_i, (h_i, p_i, q_i)$, and ζ_i satisfying Lemma 3.6(a)-(e): note that

$$f(T_i, x_i(0), x_i(T_i), \alpha_i) = V(\alpha_i) \leqq v_i \to V(0).$$

Observe that $\lambda_i \neq 0$ for all i sufficiently large. For if this were false, then there would be a subsequence with $\lambda_i = 0$ for all i (we do not relabel). Then condition (d) implies $q_i(0) = 0$ for all i and the nontriviality condition forces $p_i(0) \neq 0$ for all i. We may therefore scale the arcs (h_i, p_i, q_i) by $|p_i(0)|^{-1}$: the result is a sequence of arcs

$(h_i, p_i, q_i): [0, T] \to \mathbf{R} \times \mathbf{R}^n \times \mathbf{R}^m$ such that

$$(-\dot{p}_i(t), -\dot{q}_i(t), \dot{x}_i(t)) \in \partial \mathcal{H}(x_i(t), \alpha_i, p_i(t)) \quad \text{a.e. on } [0, T],$$

$$\mathcal{H}(x_i(t), \alpha_i, p_i(t)) = h_i \quad \text{on } [0, T],$$

$$(h_i, p_i(0), -p_i(T_i), -q_i(T_i)) \in N_S(T_i, x_i(0), x_i(T_i), \alpha_i), \quad \text{and}$$

$$|p_i(0)| = 1 \quad \text{while } q_i(0) = 0.$$

For these arcs, $\{(p_i(0), q_i(0), x_i(0))\}$ is a bounded set (by (h_4)). Since $\partial \mathcal{H}$ is a multifunction satisfying the hypotheses of Proposition 1.2, a subsequence (which we do not relabel) of $\{(p_i, q_i, x_i)\}$ converges uniformly to an arc (p, q, x) on some interval $[0, T]$ $(T \geqq T_0)$ such that x is admissible for P(0) and $(-\dot{p}, -\dot{q}, \dot{x}) \in \partial \mathcal{H}(x, 0, p)$ a.e. on $[0, T]$. Thus the sequence h_i also converges to some constant h such that $\mathcal{H}(x(t), 0, p(t)) = h$ on $[0, T]$. The note above implies that $f(T, x(0), x(T), 0) \leqq V(0)$, so that x actually solves problem P(0): the problem with this is that hypothesis (H8) allows us to take the limit of the transversality condition above and deduce that

$$(h, p(0), -p(T), q(T)) \in N_S(T, x(0), x(T), 0).$$

This contradicts the nondegeneracy of problem P(α), since $|p(0)| = 1$ while $q(0) = 0$ and $\lambda = 0$. So we must indeed have $\lambda_i = 1$ for all i sufficiently large.

Likewise, the sequence $\{p_i(0)\}$ must be bounded. Otherwise, there would be a subsequence along which $|p_i(0)| \to +\infty$ (we do not relabel). Scaling the arcs (h_i, p_i, q_i) by the factor $|p_i(0)|^{-1}$ gives a sequence of solutions and "multipliers" obeying

$$(-\dot{p}_i(t), -\dot{q}_i(t), \dot{x}_i(t)) \in \partial \mathcal{H}(x(t), \alpha_i, p_i(t)) \quad \text{a.e. on } [0, T],$$

$$\mathcal{H}(x_i(t), \alpha_i, p_i(t)) = h_i \quad \text{on } [0, T_i],$$

$$(h_i, p_i(0), -p_i(T_i), -q_i(T_i)) \in \lambda \frac{u_i}{|(\beta_i, u_i)|} \frac{\zeta_i}{|p_i(0)|} + N_S(T_i, x_i(0), x_i(T_i), \alpha_i),$$

$$|p_i(0)| = 1 \quad \text{and} \quad q_i(0) = \frac{-\lambda u_i}{|(\beta_1, -u_i)|} \frac{1}{|p_i(0)|} \to 0.$$

Passing to a uniformly convergent subsequence of $\{(p_i, q_i, x_i)\}$ by Proposition 1.2 leads once again to a solution $x: [0, T] \to \mathbf{R}^n$ for problem P(0) and an arc (p, q) with $|p(0)| = 1$ but $q(0) = 0$. The limit of the transversality condition above shows that (H9) is violated. Thus $\{p_i(0)\}$ is bounded.

Therefore the sequence $\{(p_i(0), q_i(0), x_i(0))\}$ is bounded as it stands, and $\lambda_i = 1$ for all sufficiently large i. A final application of Proposition 1.2 yields a subsequence of $\{(p_i, q_i, x_i)\}$ which converges uniformly to a solution x for problem P(0), and the adjoint arc (p, q) described in the conclusions of the lemma. (This last limiting argument depends upon the fact that since f is locally Lipschitz, the multifunction ∂f is upper semicontinuous, hence closed. See [4, Prop. 2.1.5].) □

Lemma 3.7 allows us to prove that $(\beta_0, -u_0)$ lies in $N_1 \cup N_2$. Suppose first that $u_0 \neq 0$. Upon scaling the arc (h, p, q) by $1/u_0$ in conclusions (a)-(e) and renaming the resulting adjoint arcs as (h, p, q), we obtain a multiplier of index 1 for the solution $x(\cdot)$, with $-q(p) = \beta_0/u_0$. But by construction,

$$u_0\left(\frac{\beta_0}{u_0}, -1\right) = (\beta_0, -u_0) \in N_{\text{epi } V}(0, V(0)),$$

so $-q(0) = \beta_0/u_0$ lies in $\partial V(0)$. In other words,

$$\frac{\beta_0}{u_0} \in Q[M^1(Y)] \cap \partial V(0),$$

i.e. $(\beta_0, -u_0) \in N_1$.

Second, assume $u_0 = 0$. Then (h, p, q) is an adjoint arc associated with an index 0 multiplier for the solution $x(\cdot)$, and $-q(0) = \beta_0$. Since $(\beta_0, 0)$ lies in $N_{\text{epi } V}(0, V(0))$ by construction, we also have $\beta_0 \in \partial^\infty V(0)$. Hence $\beta_0 \in Q[M^0(Y)] \cap \partial^\infty V(0)$ and $(\beta_0, 0) \in N_2$.

We have finally established that $N_{\text{epi } V}(0, V(0)) \subseteq \overline{\text{co}}(N_1 \cup N_2)$. Since the reverse inclusion is obvious, equality holds. Only one technical verification remains in the proof of Theorem 3.3.

LEMMA 3.8. *The set* $Q[M^0(Y)] \cap \partial^\infty V(0)$ *contains the recession cone of the set* $Q[M^1(Y)] \cap \partial V(0)$.

Proof. Recall that the recession cone of a set C in $\mathbf{R}^m$ is the cone

$$0^+C := \left\{ \lim_{i\to\infty} \delta_i y_i : y_i \in C, \delta_i \downarrow 0 \right\}.$$

Since $\partial^\infty V(0)$ automatically contains $0^+ \partial V(0)$, it suffices to prove that $Q[M^0(Y)]$ contains $0^+Q[M^1(Y)]$. Any element q of the latter set must be obtained as $q = \lim_{i\to\infty} -\delta_i q_i(0)$ for a sequence of solutions $x_i : [0, T_i] \to \mathbf{R}^n$ for P(0) with corresponding multipliers (p_i, q_i, ζ_i) of index 1 and a sequence of scalars $\delta_i \downarrow 0$. When we replace $(\delta_i h_i, \delta_i p_i, \delta_i q_i)$ by (h_i, p_i, q_i) in conditions (a)-(d) defining a multiplier, the Hamiltonian inclusion is unchanged and the transversality condition takes the form

$$(h_i, p_i(0), -p_i(T_i), -q_i(T_i)) \in \lambda \delta_i \zeta_i + N_S(T_i, x_i(0), x_i(T_i), 0).$$

Under this scaling, we have $q = \lim_{i\to\infty} -q_i(0)$. In particular, $\{q_i(0)\}$ is a bounded sequence. Now the sequence $\{x_i(0)\}$ is bounded by (H4), and the sequence $\{p_i(0)\}$ is bounded by an argument identical to that of Lemma 3.7. Hence the arcs $\{(p_i, q_i, x_i)\}$ have a uniformly convergent subsequence by Proposition 1.2. Since $\delta_i \downarrow 0$, the limit of this subsequence is an arc (p, q, x) on some interval $[0, T]$ for which x solves P(0) and (p, q) is an adjoint pair satisfying the limiting transversality condition

$$(h, p(0), -p(T), -q(T)) \in N_S(T, x(0), x(T), 0).$$

Moreover, one has $q = \lim_{i\to\infty} -q_i(0) = -q(0)$. That is, q lies in $Q[M^0(Y)]$ as required. □

The proof of Theorem 3.3 is complete.

4. Example: sensitivity to discount rate. In this section we apply Theorem 3.3 to a well-known optimal control problem. The discussion illustrates the theorem's utility and also demonstrates how it can be applied to many nonautonomous, fixed-time problems by introducing auxiliary state variables.

The model. Consider a factory whose productive capacity at time t is $x(t)$. At each instant, a certain fraction $u(t)$ of the factory's output may be reinvested to increase its capacity for production: the capacity then grows according to the law $\dot{x}(t) = u(t)x(t)$. The remaining output is to be sold at a fixed price: hence the total profits over the preassigned time interval $[0, \tau]$ amount to $\int_0^\tau (1 - u(t))x(t)\, dt$. The manager's objective is to maximize his company's profit over the time period $[0, \tau]$ by judicious choice of

the fraction $u(t)$ in $[0, 1]$. It can be posed as follows:

$$\min\left\{\int_0^\tau (u(t)-1)x(t)\,dt : \dot{x}(t)=u(t)x(t) \text{ a.e. on } [0,\tau],\right.$$
$$\left. 0\leqq u(t)\leqq 1 \text{ on } [0,\tau],\ x(0)=c>0\right\}.$$

To keep the problem interesting, we assume the given stopping time t exceeds 1.

We will use Theorem 3.3 to investigate the marginal loss of revenue incurred when a discount rate $\alpha>0$ forces the inclusion of the factor $e^{-\alpha t}$ in the objective integrand. This loss is measured by the function

$$V(\alpha) := \min\left\{\int_0^\tau e^{-\alpha t}(u(t)-1)x(t)\,dt : \dot{x}(t)=u(t)x(t) \text{ a.e. on } [0,\tau],\right.$$
$$\left. 0\leqq u(t)\leqq 1,\ x(0)=c>0\right\},$$

which measures the negative of the factory's profit under the optimal reinvestment policy.

Sensitivity analysis. The control problem defining $V(\alpha)$ can be solved directly by the Maximum Principle. After some calculation, one can define a switching time $\tau_s=\tau+1/\alpha \log(1-\alpha)$ and discover the optimal strategy $u(t)=1_{[0,\tau_s]}(t)$. The value of the optimal policy comes to $V(\alpha)=-c(1-\alpha)^{(1-\alpha)/\alpha}\,e^{(1-\alpha)\tau}$. L'Hospital's rule gives $V(0)=-c\,e^{\tau-1}$ and $V'(0)=c(\tau-\frac{1}{2})\,e^{\tau-1}$.

To calculate $\partial V(0)$ using Theorem 3.3, we must express V as the value function of a differential inclusion problem. This can be done by introducing new states x_0 (which measures the time) and x_1 (which accumulates the objective integral) so that the state space is $X=[-1,\tau+1]\times\mathbf{R}^2\subseteq\mathbf{R}^3$, and defining

$$F(x_0,x_1,x,\alpha) := \{(1, e^{-\delta x_0}(1-v)x, vx) : v\in[0,1]\},$$
$$f(T,x_0,x_1,x,y_0,y_1,y,\alpha) := -y_1,$$
$$S := \{\tau\}\times\{(0,0,c)\}\times\mathbf{R}^3\times\mathbf{R}.$$

Then clearly

$$V(\alpha)=\min\{f(T,x_0(0),x_1(0),x(0),x_0(T),x_1(T),x(T),\alpha):$$
$$(\dot{x}_0,\dot{x}_1,\dot{x})\in F(x_0,x_1,x,\alpha) \text{ a.e. on } [0,T],$$
$$(T,x_0(0),x_1(0),x(0),x_0(T),x_1(T),x(T),\alpha)\in S\}.$$

Hypotheses (H1)-(H8) are evidently satisfied by these data, and (H9) will be verified shortly. The Hamiltonian for this perturbed problem is

$$\mathcal{H}(x_0,x_1,x,\alpha,p_0,p_1,p)=\max\{\langle(p_0,p_1,p),(1,e^{-\alpha x_0}(1-v)x,vx)\rangle : 0\leqq v\leqq 1\}$$
$$=p_0+\max\{px, p_1 x\,e^{-\alpha x_0}\}.$$

At any base point $s\in S$, $N_S(s)=\mathbf{R}\times\mathbf{R}^3\times\{(0,0,0)\}\times\{0\}$; also $\partial f(r)=(0,0,0,0,0,-1,0,0)$ for all r.

The formula for $\partial V(0)$ given by Theorem 3.3 involves only multipliers corresponding to solutions of P(0)—the easy problem in which $\alpha=0$. Suppose, therefore, that $(x_0,x_1,x):[0,\tau]\to\mathbf{R}^3$ is an admissible arc for P(0) with a corresponding index λ

multiplier (p_0, p_1, p, q, ζ). Then the transversality condition

$$(h, p_0(0), p_1(0), p(0), -p_0(\tau), -p_1(\tau), -p(\tau), -q(\tau))$$
$$\in \lambda(0, 0, 0, 0, 0, -1, 0, 0) \times \mathbf{R} \times \mathbf{R}^3 \times \{(0, 0, 0)\} \times \{0\}$$

implies that $p_1(\tau) = \lambda$ while $p_0(\tau) = p(\tau) = q(\tau) = 0$. Since $\mathscr{H}$ has no x_1-dependence, the Hamiltonian inclusion implies that $\dot{p}_1 = 0$ a.e. on $[0, \tau]$, whence $p_1 \equiv \lambda$ on $[0, \tau]$. Thus the Hamiltonian inclusion becomes

$$(-\dot{p}_0, -\dot{p}_1, -\dot{p}, -\dot{q}, \dot{x}_0, \dot{x}_1, \dot{x}) \in \begin{cases} (0, 0, \lambda, -\lambda x_0 x, 1, x, 0) & \text{if } p < \lambda, \\ (0, 0, p, 0, 1, 0, x) & \text{if } p > \lambda, \\ \text{convex hull of these} & \text{if } p = \lambda. \end{cases}$$

Hence $p_0(t) \equiv p_0(\tau) = 0$.

Note that $\lambda = 0$ implies $\dot{q} = 0$ so $q \equiv 0$; also, $p_1 \equiv \lambda = 0$. By Gronwall's inequality, $p \equiv 0$ when $\lambda = 0$. Thus $\lambda = 0$ implies $(p_0, p_1, p) \equiv 0$ for any arc x: in particular, (H9) holds and $Q[M^0(Y)] = \{0\}$.

Now that (H1)-(H9) are verified, Theorem 3.3 applies. Corollary 2 asserts that any solution for P(0) has an index λ multiplier with $\lambda + |q(0)| > 0$. We have just seen that this condition must fail if $\lambda = 0$, so any multiplier corresponding to a solution must have $\lambda = 1$. Then $p(\tau) = 0 < \lambda$ implies that $\dot{p} = -1$ on $(\tau - 1, \tau]$ and $\dot{p} = -p$ on $[0, \tau - 1)$. Consequently $\dot{x} = x$ on $[0, \tau - 1)$ and $\dot{x} = 0$ on $(\tau - 1, \tau]$, while $\dot{x}_1 = 0$ on $[0, \tau - 1)$ and $\dot{x}_1 = x(\tau)$ on $(\tau - 1, \tau]$. We also obtain $\dot{q} = 0$ on $[0, \tau - 1)$ and $\dot{q} = tx(t)$ on $(\tau - 1, \tau]$. Integrating these equations gives the unique multiplier of index 1 for problem P(0):

$$p(t) = \begin{cases} e^{(\tau-1)-t} & \text{on } [0, \tau - 1), \\ \tau - t & \text{on } [\tau - 1, \tau], \end{cases}$$

$$q(t) = \begin{cases} c(\tfrac{1}{2} - \tau)\, e^{\tau-1} & \text{on } [0, \tau - 1), \\ \tfrac{1}{2} c\, e^{\tau-1}(t^2 - \tau^2) & \text{on } [\tau - 1, \tau]. \end{cases}$$

Thus multiplier corresponds to an arc (x_0, x_1, x) which must be the unique solution to P(0), where $x_0(t) = t$ and

$$x_1(t) = \begin{cases} 0 & \text{on } [0, \tau - 1), \\ c\, e^{\tau-1}(t - \tau + 1) & \text{on } [\tau - 1, \tau], \end{cases}$$

$$x(t) = \begin{cases} c\, e^{t} & \text{on } [0, \tau - 1), \\ c\, e^{\tau-1} & \text{on } [\tau - 1, \tau]. \end{cases}$$

Hence $V(0) = -x_1(\tau) = -c\, e^{\tau-1}$ and $Q[M^1(Y)] = \{-q(0)\} = \{c(\tau - \frac{1}{2})\, e^{\tau-1}\}$. Theorem 3.3 asserts that $\partial V(0) = \{c(\tau - \frac{1}{2})\, e^{\tau-1}\}$, as expected.

5. The minimal time function.

The problem. The standard minimum time problem in optimal control theory requires that the state of a given dynamical system be steered to its nominal value, say 0, from some given initial state $\alpha \neq 0$ in the least possible time. Of course, different initial values α generally require different control strategies and take different times to reach the origin. In a differential inclusion formulation, this variation defines a function $T: \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ via

$$T(\alpha) := \min \{T : \dot{x}(t) \in F(x(t)) \text{ a.e. on } [0, T],\ x(0) = \alpha,\ x(T) = 0\}.$$

$T(\alpha)$ assumes the value $+\infty$ iff there is no possibility of steering α to 0. Thus the very

finiteness of T is related to controllability properties of the system, a fact which we shall exploit presently. The other issues we shall discuss are the differential properties of T (or equivalently, the sensitivity of the problem with respect to the initial state), and the behaviour of T at the origin.

The function T has received attention especially in the *linear case* [1], [8], [9], [16], [17], in which $F(x)$ is of the form $Ax+BU$ for an $n\times n$ matrix A, an $n\times m$ matrix B, and a compact convex subset U of $\mathbf{R}^m$ (i.e., the case of a standard linear control system $\dot{x}=Ax+Bu$, $u\in U$). We shall recover in this section many of the results for the linear problem, but more importantly we shall develop a methodology for the much more complex nonlinear case. (See [3], [11], [18], [19], [21] for somewhat related results in a nonlinear setting.) An important issue here is the continuity of T at the origin. It is possible for T to be finite everywhere and yet not go to zero as α approaches zero.

An example of this situation can be based upon the vector field $\psi:\mathbf{R}^2\to\mathbf{R}^2$ defined by $\psi(x,y)=(\theta(y),0)$, where

$$\theta(y):=\max\{0,\min(1,2-y)\};$$

the nature of ψ is indicated in Fig. 1. We define a multifunction consistent with our hypotheses via

$$F(x,y):=\{\psi(x,y)+(u,v):|u|\leqq 1,|v|\leqq 1\}.$$

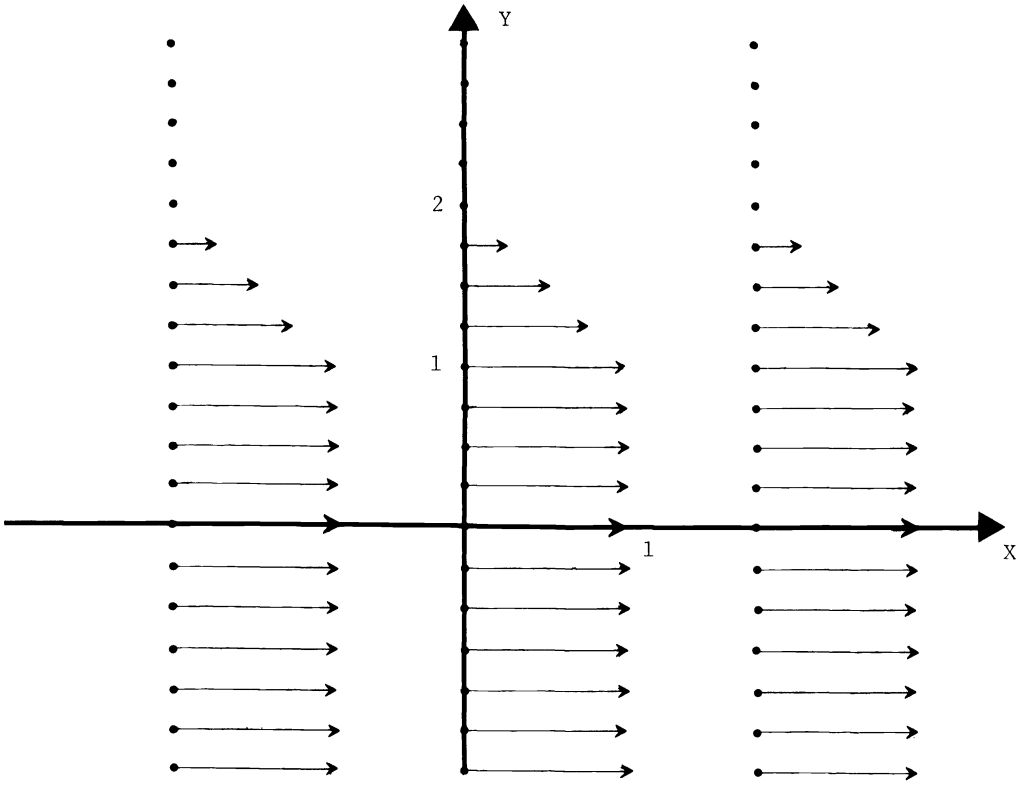


FIG. 1

We can think of ψ as the current in a river, and of (u, v) as the natural velocity of a boat; thus $F(x, y)$ is the set of possible resultant velocities depending on the steering angle. It is easy to see that starting from points $(0, -r)$ (for $r > 0$) the origin can be reached in time r, whereas, for some positive m, it is the case that from points $(\varepsilon, -r)$, for any $\varepsilon > 0$, it will require time in excess of $r + m$ to reach the origin. Thus T is discontinuous at 0 despite being finite everywhere. (Note also that $0 \in F(0)$: this hypothesis figures in some of the results below.)

The generalized gradient of T. Let us fix $\alpha_0 \in \mathbf{R}^n \backslash \{0\}$ with the property that $T(\alpha_0) < +\infty$. We posit the usual hypotheses (H1)-(H3) regarding F; these simplify somewhat because now F does not depend on α. Upon fixing any $M > T(\alpha_0)$, we may define

$$S := \{(m, \alpha_0 + \alpha, 0, \alpha) : m \in [0, M],\ \alpha \in \bar{B}\}$$

and in doing so recognize that the minimal time problem is a special case of P(0) for (H4)-(H6) are satisfied. The problem defining $T(\alpha_0)$ (namely, P(0)) admits a solution by Proposition 1.3; we adopt the hypothesis that all solutions lie in the interior of X, which gives (H7) (and implies that $0 \in \operatorname{int} X$ necessarily). It turns out that (H8) is automatically satisfied for this problem, since $N_S(s)$ is given by $\{(0, \beta, v, -\beta) : v \in \mathbf{R}^n, \beta \in \mathbf{R}^n\}$ for any s in rel int S.

Let us now examine multipliers corresponding to a trajectory $x: [0, T] \to \mathbf{R}^n$ obeying $x(0) = \alpha_0$, $x(T) = 0$. A multiplier of index λ corresponding to x consists of a triple (p, q, ζ) obeying

(a) $\dot{q}(t) = 0$ and $(-\dot{p}(t), \dot{x}(t)) \in \partial H(x(t), p(t))$ a.e. on $[0, T]$, where $H(x, p) := \sup\{\langle p, v\rangle : v \in F(x)\}$;

(b) $H(x(t), p(t)) = h$ on $[0, T]$, for some constant h;

(c) $\zeta \in \{(1, 0, 0, 0)\} = \partial f(T, \alpha_0, 0, 0)$;

(d) $(h, p(0), -p(T), -q(T)) \in \lambda(1, 0, 0, 0) + \{(0, \beta, v, -\beta) : \beta \in \mathbf{R}^n,\ v \in \mathbf{R}^n\}$.

Condition (d) implies $h = \lambda$ and $q(T) = p(0)$, so we have $q \equiv p(0)$ by (a). Moreover, $q = 0$ implies $p(0) = 0$, whence $p \equiv 0$ by Gronwall's inequality: thus (H9) holds.

We have now shown that an element of $M^\lambda(x)$, a multiplier of index λ corresponding to x, is in essence an arc p satisfying

$$(-\dot{p}(t), \dot{x}(t)) \in \partial H(x, p) \quad \text{a.e. on } [0, T], \tag{1}$$

$$H(x(t), p(t)) = \lambda \quad \text{on } [0, T]. \tag{2}$$

Since we have verified all the hypotheses, we may deduce from Theorem 3.3 the following result.

THEOREM 5.1. *One has*

$$\partial T(\alpha_0) = \overline{\operatorname{co}}\{Q[M^1(Y)] \cap \partial T(\alpha_0) + Q[M^0(Y)] \cap \partial^\infty T(\alpha_0)\},$$

where Y is the set of solutions to the minimal time problem beginning at α_0, and where $Q[M^\lambda(Y)]$ signifies the set

$$\bigcup_{x \in Y} \{-p(0) : \text{the arc } (x, p) \text{ satisfies (1) and (2)}\}.$$

If $Q[M^0(Y)]$ is pointed, the closure operation is unnecessary and one also has

$$\partial^\infty T(\alpha_0) = \operatorname{co}\{Q[M^0(Y)] \cap \partial^\infty T(\alpha_0)\}.$$

Normality. As in the case of Theorem 3.3, a variety of consequences can be drawn from these formulas. We shall limit ourselves here to those bearing upon the Lipschitz character of T. An arc x admissible for P(α_0) is temed *normal* if whenever conditions (1) and (2) hold with $\lambda = 0$, then $p \equiv 0$; i.e. if the only multiplier of index 0 for x is

the trivial one. The problem $P(\alpha_0)$ itself is termed *normal* if every solution x in Y is normal; i.e., if $Q[M^0(Y)] = \{0\}$. Theorem 5.1 then yields the following.

COROLLARY 1. *If* $P(\alpha_0)$ *is normal, then* T *is Lipschitz near* α_0, *and* $\partial T(\alpha_0)$ *is a compact convex set obeying*

$$\operatorname{ext}\{\partial T(\alpha_0)\} \subseteq Q[M^1(Y)].$$

(In the terminology of [4], [5], this implies that "normal problems are calm.")

DEFINITION 5.2. The *abnormal set* S is the set of all points $x(\tau)$ where, for some $\tau > 0$, there exists an arc $(x, p):[0, \tau] \to \mathbf{R}^n \times \mathbf{R}^n$ with p not identically zero, such that

(i) $x(0) = 0$, $x(t) \in \operatorname{int} X$;

(ii) $(-\dot{p}(t), \dot{x}(t)) \in -\partial H(x(t), p(t))$ a.e. on $[0, \tau]$;

(iii) $H(x(t), p(t)) = 0$ on $[0, \tau]$.

By convention, we shall also allow $\tau = 0$ above, so that S always contains the origin. Observe that after time reversal, S is the set of states reachable from the origin via abnormal F-trajectories (condition (ii) implies $\dot{x}(t) \in -F(x(t))$ a.e.).

COROLLARY 2. *If the point* α_0 *of Theorem* 5.1 *lies in the complement of* S, *then* T *is Lipschitz near* α_0.

Proof. The problem $P(\alpha_0)$ is normal, for if x on $[0, T]$ were any abnormal solution to $P(\alpha_0)$ and p its associated multiplier, then the arc $(\tilde{x}(t), \tilde{p}(t)) := (x(T-t), p(T-t))$ would satisfy the conditions of Definition 5.2, contradicting $\alpha_0 \notin S$. The result follows from Corollary 1. □

The set S thus provides an estimate of the initial conditions for which T is badly behaved (i.e., non-Lipschitz). We shall be able to study the "smallness" of S following a discussion of the behaviour of T at the origin.

Behaviour at the origin. We shall henceforth assume that $0 \in F(0)$. This implies that the zero arc is an F-trajectory.

DEFINITION 5.3. We shall say *the origin is normal* if for all positive T sufficiently small, the zero arc on $[0, T]$ is normal.

To paraphrase, the origin is normal if for T small enough, it is not possible to find a nontrivial arc p satifying on $[0, T]$ the conditions

$$(-\dot{p}(t), 0) \in \partial H(0, p(t)) \quad \text{a.e.,} \tag{3}$$

$$H(0, p(t)) = 0. \tag{4}$$

(By reversing time, it is equivalent to consider $(\dot{p}, 0) \in \partial H(0, p)$ in (3).) We recall [4] that a locally Lipschitz function φ is called *regular* at u provided that for all v the usual one-sided directional derivative $\varphi'(u; v)$ exists and coincides with the generalized directional derivative $\varphi^\circ(u; v)$. Among others, smooth and convex functions are regular. It follows from [4, Thm. 2.8.2] that H is regular when F admits a smooth classical relaxed representation.

The following result asserts that by and large the abnormal set will be nontrivial unless 0 lies in the interior of $F(0)$.

PROPOSITION 5.4. *If* $0 \in \operatorname{int} F(0)$, *then* $S = \{0\}$ *and the origin is normal*; *moreover,* T *is Lipschitz on a neighbourhood of* 0. *Conversely, when the origin is normal and* H *is regular on* $\{0\} \times \mathbf{R}^n$, *then* $S = \{0\}$ *implies* $0 \in \operatorname{int} F(0)$.

Proof. Suppose that $0 \in \operatorname{int} F(0)$. Then $H(0, p(0)) = 0$ implies $p(0) = 0$, so that (since $|\dot{p}| \leqq K|p|$) any arc (x, p) satisfying (i)-(iii) of Definition 5.2 has p identically zero. Thus $S = \{0\}$ and the origin is normal.

To verify that T is Lipschitz near 0, observe that the Lipschitz character of F implies that there exists $\delta > 0$ for which $\delta \bar{B} \subseteq F(\alpha)$ for all α in δB. Hence there is a

constant $M>0$ such that for all α in δB, $H(\alpha, p(0))=1$ implies $|-p(0)| \leqq M$. By Theorem 5.1, $\partial T(\alpha) \subseteq M\bar{B}$ for all nonzero α in δB. We will see below (Thm. 5.6) that T is continuous at 0. This allows us to verify that T is Lipschitz of rank $M+1$ throughout δB.

For the converse, suppose that 0 fails to lie in int $F(0)$. We need to show that S does not reduce to $\{0\}$. The normal cone to $F(0)$ contains a nonzero element ζ: $H(0, \zeta)=0$. Consider the differential inclusion initial-value problem

$$(-\dot{p}(t), \dot{x}(t)) \in -\partial H(x(t), p(t)), \qquad (x(0), p(0))=(0, \zeta).$$

Since ∂H is compact convex-valued and upper semicontinuous [4], standard existence results imply the local existence of a solution (x necessarily remains in int X for t near 0.) Then the regularity of H applies [4, Prop. 7.7.1] to yield $H(x, p)=$ *constant* $=H(0, p(0))=H(0, \zeta)=0$. But the origin is normal, so $x(t) \neq 0$ for $t>0$: this proves that S contains nonzero points. □

The following shows that under certain smoothness hypotheses S can be shown to consist of a curve through the origin. (It seems that S often coincides with the "switching curve" of an optimal synthesis of the prolem, but no formal result of that type is yet known to us.) A function is said to be C^{1+} if it is differentiable and its derivative is locally Lipschitz. Zeidan [23, Thm. 3] has given conditions assuring that the Hamiltonian of a control problem be C^{1+}.

PROPOSITION 5.5. *Suppose that $H(x, p)$ is C^{1+} on $X \times (\mathbf{R}^n \backslash \{0\})$, and that the normal cone to $F(0)$ at 0 is contained in the line $\mathbf{R}\zeta$ for some ζ in $\mathbf{R}^n$. Then for some (non-degenerate) interval $[a, b]$ containing 0, S is of the form*

$$\{s(t): a \leqq t \leqq b\},$$

where $s(0)=0$, s is continuous at 0, and s is C^1 in (a, b) except perhaps at 0. (We allow the cases $a=-\infty$ and $b=+\infty$.)

Proof. Consider (x, p) as in Definition 5.2. (If no such (x, p) exist, then $S=\{0\}$ and we simply take $s \equiv 0$.) Since $H(0, p(0))=0$, we can assume (by scaling p) that $p(0)$ is either ζ or $-\zeta$. If $\zeta=0$ then $0 \in$ int $F(0)$, which implies $S=\{0\}$ (Prop. 5.4). So suppose $\zeta \neq 0$, and consider first the possible case $p(0)=\zeta$. Then there exists a local solution to the initial-value problem

$$(-\dot{p}(t), \dot{x}(t))=\nabla H(x(t), p(t)), \qquad (x(0), p(0))=(0, \zeta).$$

Since the Hamiltonian has the constant value $H(0, \zeta)=0$ along such a solution, the curve $x(t)$ traces out points of S until it leaves X (if ever). In a similar fashion, the case $p(0)=-\zeta$ may produce another branch of S. The required representation of S now follows readily. □

We now see that normality of the origin rules out the type of behaviour exhibited by the example illustrated by Fig. 1.

THEOREM 5.6. *If the origin is normal, then T is finite in a neighbourhood of* 0, *and continuous at* 0.

Proof. We shall define a new family of problems P(α) of the type discussed in § 3, one for which x belongs to $\mathbf{R}^{n+1}$. We shall use the notation (y, z) to express x as a point in $\mathbf{R}^n \times \mathbf{R}$. The data of problem (P) are:

$f(T, x_0, x_T) := z_T+(T-b)^2$, where $x_T=(y_T, z_T)$ and b is a given positive number chosen so small that the zero is normal on $[0, b]$;

$\tilde{F}(x) := F(y) \times \{|y|\}, \qquad \tilde{X}=X \times \mathbf{R},$

$S := \{(T, \alpha, 0, 0, \beta, \alpha) \in \mathbf{R} \times \mathbf{R}^{n+1} \times \mathbf{R}^{n+1} \times \mathbf{R}^n : \frac{1}{2}b \leqq T \leqq 2b, |\alpha| \leqq 1, \text{any } \beta\}.$

(The reader may find it helpful to view the problem $P(\alpha)$ in an equvalent form: to minimize $(b-T)^2+\int_0^T |y(t)|\,dt$ over the trajectories y for F on an interval $[0, T]$ satisfying $y(0)=\alpha$, $y(T)=0$.) It is evident that the unique solution to $P(0)$ is $x\equiv 0$ on $[0, b]$.

We now claim that this arc is normal. We calculate (for $p=(r, s)$ in $\mathbf{R}^n\times\mathbf{R}$)

$$\tilde{H}(x, \alpha, p)=H(y, r)+s|y|.$$

A multiplier of index 0 corresponding to the zero arc on $[0, b]$ therefore consists of an arc (r, s, q) satisfying on $[0, b]$ the conditions:

$$(-\dot{r}(t), 0)\in\partial H(0, r(t))+s\bar{B}\times\{0\},\quad \dot{s}(t)=0,\quad \dot{q}(t)=0 \quad \text{a.e.},$$

$$H(0, r(t))=h,\qquad (h, r(0), s(0), -r(b), -s(b), -q(b))\in N_S(b, x(0), x(b), 0).$$

The last of these conditions gives $h=0$, $r(0)=q(b)$, $s(b)=0$. It follows that s is identically zero, so that r satisfies on $[0, b]$

$$(-\dot{r}(t), 0)\in\partial H(0, r) \quad \text{a.e.},\qquad H(0, r(t))=0.$$

Since the origin is normal, our choice of b implies that r is zero on $[0, b]$. Then $q(b)=r(0)=0$, which together with $\dot{q}=0$ a.e. implies that q is 0 too. Hence (r, s, q) is zero, which confirms that the only multiplier of index 0 is the trivial one, as stated. (This incidentally implies hypothesis (H9) of § 3; the other hypotheses are readily confirmed.) We now apply Corollary 1 of Theorem 3.3 to deduce that the value function $V(\alpha)$ corresponding to the above problem is finite and Lipschitz near 0. For some $\delta>0$, this certainly implies the existence, for all $|\alpha|\leqq\delta$, of an admissible trajectory y for F on an interval $[0, \tau]$ satisfying $\tau\leqq 2b$, $y(0)=\alpha$, $y(\tau)=0$, whence we deduce $T(\alpha)\leqq 2b$. (Of course, δ depends on b.) Given $\varepsilon>0$, choose any b in $(0, \varepsilon/2)$, and let the above conclusions hold for that b, for all $|\alpha|\leqq\delta$. Then $|\alpha|<\delta$ implies $T(\alpha)\leqq 2b<\varepsilon$ and we have shown that T is continuous at 0. □

Remark. A simple example in which the origin is normal yet for which T is not Lipschitz at 0 (albeit continuous) is the well-known instance of the linear case $F(x)=Ax+BU$ in which

$$A=\begin{pmatrix}0 & 1\\ 0 & 0\end{pmatrix},\quad B=\begin{pmatrix}0\\ 1\end{pmatrix},\quad U=[-1, 1].$$

The origin is normal by Corollary 2 below; calculation shows

$$T(x, y)=\begin{cases}-y+[2y^2-4x]^{1/2} & \text{if } 2x\leqq -y|y|,\\ y+[2y^2+4x]^{1/2} & \text{if } 2x\geqq -y|y|.\end{cases}$$

This function fails to be Lipschitz along

$$S=\{(x, y)\colon 2x+y|y|=0\},$$

the abnormal set and also the switching curve for the optimal synthesis.

Controllability. We now proceed to explore some controllability implications of Theorem 5.6. The proof showed that T is finite near 0, whence:

COROLLARY 1. *If the origin is normal, then there is a neighbourhood of* 0 *all of whose points can be steered to* 0 *in finite time.*

We now show that Corollary 1 can be viewed as an extension of the well-known result in linear control theory (see for example [12]) that gives a criterion for controllability in terms of the rank of the controllability matrix

$$C=[B\quad AB\quad A^2B\quad \cdots\quad A^{n-1}B].$$

Suppose that $F(x)$ is of the form $\varphi(x)+BU$, where $\varphi: \mathbf{R}^n \to \mathbf{R}^n$ is C^1, $\varphi(0)=0$, B is $n \times m$, and U is a compact convex subset of $\mathbf{R}^m$ containing 0 in its interior. Construct the $n \times mn$ matrix C as indicated above, where $A = D\varphi(0)$.

COROLLARY 2. *The controllability matrix C has full rank (i.e., n) if and only if the origin is normal. When this is the case, all points in a neighbourhood of 0 can be steered to 0 in finite time, and T is continuous at 0.*

Proof. The origin fails to be normal if and only if there is a nonvanishing arc p defined on some interval $[0, T]$ and obeying

$$(-\dot{p}(t), 0) \in \partial H(0, p(t)) \quad \text{a.e.}, \qquad H(0, p(t)) = 0.$$

Since $H(x, p)$ is given by $\langle \varphi(x), p\rangle + \max\{\langle p, Bu\rangle : u \in U\}$ in this setting, these conditions are equivalent to

$$-\dot{p}(t) = A^* p(t) \quad \text{a.e.}, \qquad B^* p(t) \equiv 0.$$

Defining $p_0 = p(0)$, we find that the origin fails to be normal if and only if there is a nonzero vector p_0 such that for some $T > 0$,

$$B^* e^{-A^* t} p_0 \equiv 0 \quad \text{on } [0, T]. \tag{5}$$

Now if the origin is not normal, then alternately setting $t = 0$ and computing d/dt in (5) shows that $p_0^* C = 0$, i.e. rank $(C) < n$. Conversely, if rank $(C) < n$ then there exists a nonzero vector p_0 such that $p_0^* C = 0$. This implies (5) via the Cayley-Hamilton theorem. □

Normality of the origin has been seen to be a useful property of a system; it is one that can often be confirmed on an ad hoc basis in specific cases. We now give an example of a criterion applicable to fully nonlinear systems. We suppose that $F(x)$ has a classic representation $\varphi(x, U)$, where $D_x\varphi(x, u)$ exists, φ and $D_x\varphi$ are continuous in (x, u), and U is compact.

COROLLARY 3. *Suppose that the normal cone to $F(0)$ at 0 is contained in the line $\mathbf{R}\zeta$ for some ζ, that 0 is an extreme point of $F(0)$, and that a unique $\hat{u} \in U$ exists such that $\varphi(0, \hat{u}) = 0$. Then, if ζ is not an eigenvector for $D_x^* \varphi(0, \hat{u})$, the origin is normal. Thus T is finite near 0 and continuous at 0. If in addition H is C^{1+} for $p \neq 0$, then T is Lipschitz near 0 except perhaps on a curve through the origin.*

Proof. If the origin is not normal, then there exists on some interval $[0, T]$ a nonvanishing arc p satisfying

$$(-\dot{p}(t), 0) \in \partial H(0, p(t)) \quad \text{a.e.}, \qquad H(0, p(t)) = 0.$$

Here we have $H(x, p) = \max\{\langle p, \varphi(x, u)\rangle : u \in U\}$, and by [4, Thm. 2.8.2] we deduce the existence of $n+1$ points u_i in U such that

$$0 = \sum_{i=1}^{n+1} \lambda_i \varphi(0, u_i),$$

$$-\dot{p}(t) = \sum_{i=1}^{n+1} \lambda_i D_x^* \varphi(0, u_i) p(t),$$

$$\langle p(t), \varphi(0, u_i)\rangle = H(0, p(t)) = 0 \text{ for each } i,$$

where the λ_i are nonnegative and sum to 1. Since each $\varphi(0, u_i)$ belongs to $F(0)$ and 0 is an extreme point, each $\varphi(0, u_i)$ equals 0, whence each u_i equals $\hat{u}$. The condition $H(0, p(t)) = 0$ implies that $p(t)$ belongs to $N_{F(0)}(0)$ for each t, whence $p(t) = \alpha(t)\zeta$

for some nonvanishing scalar function α. These conclusions lead to

$$\alpha(t)A^*\zeta = -\dot{\alpha}(t)\zeta,$$

which contradicts the hypothesis that ζ is not an eigenvector for A^*. The final assertion follows from Proposition 5.5. □

When F is globally defined (i.e., $X = \mathbf{R}^n$), we can combine Lyapunov stability and controllability near 0 in the usual way (see for example [12]) to obtain a global conclusion as follows.

COROLLARY 4. *Let there be a locally Lipschitz selection $f(x)$ for $F(x)$ such that $f(x) \cdot \nabla L(x) < 0$ for $x \neq 0$, where $L: \mathbf{R}^n \to \mathbf{R}$ is differentiable for $x \neq 0$ and the level sets $\{x: L(x) \leqq \alpha\}$ are compact for each α. Suppose in addition that the origin is normal. Then any point $\mathbf{R}^n$ can be steered to the origin in finite time.*

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. 567-587.

[2] A. AUSLENDER, *Differential Stability in Nonconvex and Nondifferentiable Programming*, Math. Programming Stud., 10, P. Huard, ed., North-Holland, Amsterdam.

[3] A. V. BOLTYANSKII, *The continuity of Bellman's functions*, Differentsialnye Uravneniya, 15 (1979), pp. 131-133. (In Russian.)

[4] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, Wiley-Interscience, New York, 1983.

[5] F. H. CLARKE AND R. B. VINTER, *Local optimality conditions and Lipschitzian solutions to the Hamilton-Jacobi equation*, this Journal, 21 (1983), pp. 856-870.

[6] J. GAUVIN, *The generalized gradient of a marginal function in mathematical programming*, Math. Oper. Res., 4 (1979), pp. 458-463.

[7] B. GOLLAN, *Perturbation theory for abstract optimization prolems*, J. Optim. Theory Appl., 35 (1981), pp. 417-441.

[8] O. HAJEK, *Geometric theory of time-optimal control*, this Journal, 9 (1971), pp. 339-350.

[9] ———, *On differentiability of the minimal time function*, Funkcial. Ekvac., 20 (1976), pp. 97-114.

[10] H. HERMES, *On local controllability*, this Journal, 20 (1982), pp. 211-220.

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

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

[13] F. LEMPIO AND H. MAURER, *Differential stability in infinite-dimensional nonlinear programming*, Appl. Math. Optim., 6 (1980), pp. 139-152.

[14] P. D. LOEWEN, *The sensitivity of optimal value functions in differential inclusion problems*, Thesis, Univ. of British Columbia, Canada.

[15] H. MAURER, *Differential stability in optimal control problems*, Appl. Math. Optim., 5 (1979), pp. 283-295.

[16] F. MIGNANEGO AND G. PIERI, *On a generalized Bellman's equation for the optimal-time problem*, Systems Control Lett., to appear.

[17] G. OLDSER, *Time-optimal control of multivalued systems near the origin*, J. Optim. Theory Appl., 16 (1975), pp. 497-517.

[18] N. N. PETROV, *On the Bellman function for the time-optimal process problem*, J. Appl. Math. Mech., 34 (1970), pp. 785-791.

[19] ———, *On the continuity of the Bellman function with respect to a parameter*, Vestnik Leningrad Univ. Math., 7 (1979), pp. 169-176.

[20] R. T. ROCKAFELLAR, *Lagrange multipliers and subderivatives of optimal value functions in nonlinear programming*, Math. Programming Stud., 17 (1982), pp. 28-66.

[21] E. ROXIN, *A geometric interpretation of Pontryagin's maximum principle*, in Nonlinear Differential Equations and Nonlinear Mechanics, J. P. LaSalle and S. Lefschetz, eds., Academic Press, New York, 1963, pp. 303-324.

[22] J. WARGA, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.

[23] V. ZEIDAN, *Sufficient conditions for the generalized problem of Bolza*, Trans. Amer. Math. Soc., 275 (1983), pp. 561-586.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 2, March 1986

007

STRONG CONTROLLABILITY OF NONLINEAR SYSTEMS*

RONALD M. HIRSCHORN†

Abstract. The shape of the reachable set for an affine control system is studied by introducing trace vector fields, freely-controlled and semi-controlled vector fields. Properties of these classes of vector fields are established and they are related to the structure of the reachable set of states at time t. A method for generating semi-controlled vector fields from a trace vector fields and a freely-controlled vector field is presented.

Key words. controllability, affine systems

1. Introduction. The purpose of this paper is to study the shape of the reachable set for system models of the form

$$\frac{dx}{dt}(t) = A(x(t)) + \sum_{i=1}^{m} u_i(t) B_i(x(t)), \qquad x(0) = x_0 \in M \tag{1.1}$$

where M is a real analytic manifold, $A, B_1, \cdots, B_m$ are real analytic vector fields and the controls u_i are piecewise constant functions from $R^+ = [0, \infty)$ into $R = (-\infty, \infty)$. A state x_1 is said to be *reachable from* x_0 *at time* t_1 if for some choice of controls there is a solution $x_u(t)$ to (1.1) with $x_u(0) = x_0$ and $x_u(t_1) = x_1$. Let $\mathcal{R}_t(x_0)$ denote the set of states reachable from x_0 at time t, and set $\mathcal{R}(x_0) = \cup_{t \geqq 0} \mathcal{R}_t(x_0)$. In order to investigate the shape of the reachable set $\mathcal{R}(x_0)$, we introduce the idea of a *trace vector field* X, a vector field with the property that $X_t(x) \in \overline{\mathcal{R}_t(x)} \forall x \in M$ and $\forall t > 0$ where $X_t(x)$ is defined, i.e., the integral curve for X at time t is in the closure of reachable set at time t. The set $\mathcal{T}$ of trace vector fields of the system (1.1) yields information about the shape of the reachable set. The standard linear controllability results asserts that for the system $\dot{x} = Ax + Bu$ a vector field X belongs to $\mathcal{T}$ iff $X(x) = Ax + d(x)$ where $d(x) \in \text{range}\,[B \mid AB \mid \cdots \mid A^{n-1}B]$. It follows that $\mathcal{R}_t(x_0)$ includes $e^{tA}x_0 + \text{range}\,[B \mid AB \mid \cdots \mid A^{n-1}B]$.

For the system (1.1) it is clear that the "drift term" A belongs to $\mathcal{T}$, and the "controlled" vector fields B_i have the property that $A + \alpha B_i$ belongs to $\mathcal{T}$ for all real α (i.e., set $u_i \equiv \alpha_i$ and $u_j \equiv 0$ for $j \neq i$). The purpose of this paper is to describe a method for generating more vector fields in $\mathcal{T}$ by finding more "controlled" vector fields. A vector field X is said to be a *freely-controlled* vector field for the system (1.1) if $A + \alpha X$ belongs to $\mathcal{T}$ for all real α and *semi-controlled* if $A + \alpha X$ belongs to $\mathcal{T}$ for all $\alpha \geqq 0$. Thus in the linear case X is freely-controlled iff $X(x) \in \text{range}\,[B \mid AB \mid \cdots \mid A^{n-1}B]$ for all x in $M = R^n$, and every semi-controlled vector field is also freely controlled. In the nonlinear case there are often semi-controlled vector fields which are not freely-controlled. The problem of finding the reachable set for nonlinear systems has been studied by a number of authors (cf. Brockett [1], Hirschorn [5], Kunita [8], Sussmann and Jurdjevic [12], Jurdjevic and Kupka [9]). The work of Kunita [8] provided the motivation for this paper.

In § 2 some of the basic properties of semi and freely-controlled vector fields and their connections with the reachable set are presented. In § 3 a technique is described for generating semi-controlled vector fields from a freely-controlled vector field and a trace vector field.

* Received by the editors February 21, 1984, and in final revised form January 14, 1985.
† Department of Mathematics and Statistics, Queen's University, Kingston, Ontario, Canada K7L 3N6.

2. Controlled vector fields and controllability. Let $V(M)$ denote the set of all real-analytic vector fields on M and $C^\omega(M)$ the ring of real analytic functions on M. We regard $V(M)$ as a Lie algebra over R with the Lie bracket $[X, Y] = XY - YX$ and let ad_X denote the linear operator on $V(M)$ defined by $\mathrm{ad}_X Y = [X, Y]$. If X is a vector field on M and $x \in M$, then $t \to X_t(x)$ denotes the *integral curve* of X through x, i.e., $X_0(x) = x$ and $(d/dt)X_t(x) = X(X_t(x))$. If $X_t(x)$ is defined for all t in R and for all $x \in M$, then X is said to be *complete*. If D is a subset of $V(M)$ then $\{D\}_{LA}$ will denote the smallest Lie subalgebra of $V(M)$ containing D and $D(x)$ is the subset $\{X(x) | X \in D\}$ of the tangent space of M at x.

The system (1.1) gives rise to the triple of Lie subalgebras $(\mathscr{L}, \mathscr{L}_0, \mathscr{B})$ where $\mathscr{B} = \{B_1, \cdots, B_m\}_{LA}$, $\mathscr{L} = \{A, B_1, \cdots, B_m\}_{LA}$ and $\mathscr{L}_0$ is the smallest ideal in $\mathscr{L}$ which contains $\mathscr{B}$. In [12] Sussmann and Jurdjevic relate some of basic qualitative properties of the reachable set to these Lie algebras. Let $\mathscr{D}$ be a Lie subalgebra of $V(M)$. Then $x \to \mathscr{D}(x)$ defines an involutive distribution on M and $I(\mathscr{D}, x)$ will denote the unique maximal connected integral manifold for this distribution containing x. The existence of $I(\mathscr{D}, x)$ is a consequence of a global version of Frobenius' theorem (cf. [11]). Sussmann and Jurdjevic [12] showed that the reachable set $\mathscr{R}(x)$ is contained in $I(\mathscr{L}, x)$ and, with respect to the topology of $I(\mathscr{L}, x)$, $\mathscr{R}(x)$ is contained in the closure of its interior. Also $\mathscr{R}_t(x)$ is contained in $I(\mathscr{L}_0, A_t(x))$ and $\mathscr{R}_t(x)$ is contained in the closure of its interior relative to the topology of $I(\mathscr{L}_0, A_t(x_0))$. This means that the closure of the reachable set has basically the same shape as the reachable set. The system (1.1) will be called *controllable* if $\mathscr{R}(x) = I(\mathscr{L}, x)$ for all $x \in M$ and *strongly controllable* if $\mathscr{R}_t(x) = I(\mathscr{L}_0, A_t(x))$ for all $x \in M$ and for all $t > 0$. For linear systems $I(\mathscr{L}_0, x) = x + \mathrm{range}\,[B \vdots AB \vdots \cdots \vdots A^{n-1}B]$, $A_t(x) = e^{tA}x$, so that every time-invariant linear system is strongly-controllable. A few nonlinear systems are known to be strongly-controllable (cf. [5], [8]) but this is not the usual case. In fact for most systems little is known about $\mathscr{R}_t(x)$ beyond the fact that the interior of $\mathscr{R}_t(x)$ is nonempty in $I(\mathscr{L}_0, A_t(x))$.

A more detailed picture of $\mathscr{R}_t(x)$ and $\mathscr{R}(x)$ is possible. A vector field $Y \in V(M)$ is called a *trace vector field* for the system (1.1) if $Y_t(x) \in \overline{\mathscr{R}_t(x)}$ for all $x \in M$ and for all $t > 0$ such that $Y_t(x)$ is defined, where $\overline{\mathscr{R}_t(x)}$ is the closure of $\mathscr{R}_t(x)$ relative to the topology of $I(\mathscr{L}_0, A_t(x))$. A vector field X is *semi (freely)-controlled* if $A + \alpha X$ is a trace vector field for all $\alpha > 0$ ($\alpha \in R$).

Let $\mathscr{T}$ denote the collection of all trace vector fields for (1.1), $\mathscr{F}$ the collection of freely-controlled vector fields, and $\mathscr{F}^+$ the semi-controlled vector fields for (1.1). One can obtain more quantitative information about the reachable set by examining the triple $(\mathscr{T}, \mathscr{F}^+, \mathscr{F})$. Some of the basic properties of $\mathscr{T}$, $\mathscr{F}^+$ and $\mathscr{F}$ are derived in the following:

LEMMA 2.1. *Consider the system* (1.1). *Then*:

a) *A vector field X belongs to $\mathscr{F}^+$ ($\mathscr{F}$) if and only if $A_t(X_s(x))$ is in the closure of $\mathscr{R}_t(x)$ for all $t > 0$, $s > 0$ ($s \in R$) and for all $x \in M$.*

b) *If $f_1, f_2, \cdots, f_k \in C^\omega(M)$ and $X_1, \cdots, X_k \in \mathscr{F}$ then $\sum_{i=1}^k f_i X_i \in \mathscr{F}$.*

c) *If $f_1, \cdots, f_k \in C^\omega(M)$ are nonnegative and $X_1, \cdots, X_k \in \mathscr{F}^+$ then $\sum_{i=1}^k f_i X_i \in \mathscr{F}^+$.*

d) *If $c_1, c_2, \cdots, c_k$ are nonnegative real numbers with $\sum_{i=1}^k c_i = 1$ and if $Y_1, \cdots, Y_k \in \mathscr{T}$ then $\sum_{i=1}^k c_i Y_i \in \mathscr{T}$.*

Proof. a) Let ε, $t > 0$ and let $X \in \mathscr{F}^+$. Then $(A + \alpha X)_t(x) \in \overline{\mathscr{R}_t(x)}$ for all $\alpha > 0$ so for $\alpha = s/\varepsilon$ one has $(A + sX/\varepsilon)_\varepsilon(x) \in \overline{\mathscr{R}_\varepsilon(x)}$ and $A_{t-\varepsilon} \circ (A + sX/\varepsilon)_\varepsilon(x) \in \overline{\mathscr{R}_t(x)}$. Letting ε tend to zero, one has $A_t \circ X_s(x) \in \overline{\mathscr{R}_t(x)}$. Suppose $A_t \circ X_s(x) \in \overline{\mathscr{R}_t(x)}$ for s, $t > 0$ and

for all $x \in M$. The Campbell-Baker-Hausdorff formula asserts that if $F, G \in V(M)$ then

$$F_{t_1} \circ G_{t_2}(x) = Z_1(x) \quad \text{where } Z = t_1 F + t_2 G + \frac{t_1 t_2}{2}[F, G] + \cdots$$

is a formal power series which converges for t_1 and t_2 in some neighbourhood of 0 in R^2. Thus $(A_{t/n} \circ X_{\alpha t/n})^n(x) = (t/nA + \alpha t/nX + t^2\alpha/2n^2[A, X] + \cdots)_1^n(x) \in \overline{\mathcal{R}_t(x)}$ where n is a positive integer and $\alpha > 0$. Letting $n \to \infty$, one has $(A + \alpha X)_t(x) \in \overline{\mathcal{R}_t(x)}$ so that $X \in \mathcal{F}^+$. If $X \in \mathcal{F}$, then $-X \in \mathcal{F}^+$ and the case where $X \in \mathcal{F}$ follows from the above.

b) The integral curve for $f_i X_i$ is a reparametrization of the integral curve of X_i, so using part a) it is clear that $f_i X_i \in \mathcal{F}$. Thus it is enough to check that if $X_1, \cdots, X_k \in \mathcal{F}$ then $\sum_{i=1}^k X_i \in \mathcal{F}$. For $\alpha \in R$ the well-known results on bang-bang controls (cf. [7]) assert that the integral curves for $A + \alpha(\sum_{i=1}^k X_i)$ can be approximated by compositions of integral curves for $\{A + \alpha X_i | i = 1, \cdots, k\}$ and it follows that $A + \alpha \sum_{i=1}^k X_i \in \mathcal{T}$, hence $\sum_{i=1}^k X_i \in \mathcal{F}^+$.

c) and d) are proved using the same method of proof as is used in part b).

THEOREM 2.2. *Consider the system* (1.1) *with associated triples* $(\mathcal{L}, \mathcal{L}_0, \mathcal{B})$ *and* $(\mathcal{T}, \mathcal{F}^+, \mathcal{F})$. *Then*:

a) $\mathcal{F}$ *is a Lie subalgebra of* $V(M)$ *containing* $\mathcal{B}$.

b) $\mathcal{T} + \mathcal{F}^+ \subset \mathcal{T}$.

c) $\mathcal{F}^+ \supset \mathcal{F}$ *and* $\mathcal{F}(x) \subset \mathcal{F}^+(x) \subset \mathcal{L}_0(x)$ *for all* $x \in M$.

Proof. a) $\mathcal{F}$ is a vector space by Lemma 2.1 part b). Let $X, Y \in \mathcal{F}$. To show that $[X, Y] \in \mathcal{F}$, one can proceed as follows: For $X^1, \cdots, X^k \in \mathcal{F}$, $s_1, \cdots, s_k \in R$ one can use Lemma 2.1 part a) to see that $A_{t-\varepsilon k} \circ (A_\varepsilon \circ X^1_{s_1}) \circ \cdots \circ (A_\varepsilon \circ X^k_{s_k})(x) \in \overline{\mathcal{R}_t(x)}$ for all $\varepsilon > 0$. Letting $\varepsilon \to 0$ it follows that $A_t \circ X^1_{s_1} \circ \cdots \circ X^k_{s_k}(x) \in \overline{\mathcal{R}_t(x)}$. From Chow's theorem (cf. [7]) one can see that for each $t > 0$ and n a positive integer one can express $([X, Y])_{t/n}(x)$ as $X_{r_1} \circ Y_{r_2} \circ \cdots \circ X_{r_{l-1}} \circ Y_{r_l}(x)$ for some set of real numbers $r_1, \cdots, r_l$ and thus $A_{t/n} \circ ([X, Y])_{t/n}(x) \in \overline{\mathcal{R}_{t/n}(x)}$. It follows that $(A_{t/n} \circ ([X, Y])_{t/n})^n(x) \in \overline{\mathcal{R}_t(x)}$, and letting $n \to \infty$ the Campbell-Baker-Hausdorff formula can be used to show that $(A + [X, Y])_t(x) \in \overline{\mathcal{R}_t(x)}$. If X is replaced by a multiple αX, one has $(A + \alpha[X, Y])_t(x) \in \overline{\mathcal{R}_t(x)}$ or $[X, Y] \in \mathcal{F}$.

b) Let $X \in \mathcal{F}^+$ and $Y \in \mathcal{T}$ so that from Lemma 2.1 $A_t \circ X_s(x) \in \overline{\mathcal{R}_t(x)}$ for $t, s > 0$ and by definition $Y_t(x) \in \overline{\mathcal{R}_t(x)}$ for $t > 0$. Thus $(A_\varepsilon \circ X_s) \circ (Y_{t-\varepsilon})(x) \in \overline{\mathcal{R}_t(x)}$ for all $\varepsilon > 0$ and in the limit as $\varepsilon \to 0$ one sees that $X_s \circ Y_t(x) \in \overline{\mathcal{R}_t(x)}$ for all $s > 0$. In particular $X_{t/n} \circ Y_{t/n}(x) \in \overline{\mathcal{R}_t(x)}$ for n a positive integer, and using the Campbell-Baker-Hausdorff formula one sees that $(X_{t/n} \circ Y_{t/n})^n(x) \to (X + Y)_t(x) \in \overline{\mathcal{R}_t(x)}$. This means that $X + Y \in \mathcal{T}$.

c) Clearly $\mathcal{F}^+ \supset \mathcal{F}$. Fix $x \in M$ and $X \in \mathcal{F}^+$. Then $A_t \circ X_s(x) \in \overline{\mathcal{R}_t(x)}$ for $s > 0$. Now $\overline{\mathcal{R}_t(x)} \subset I(\mathcal{L}_0, A_t(x))$ and in [12] it is shown that $I(\mathcal{L}_0, A_t(x)) = A_t(I(\mathcal{L}_0, x))$. This means that $A_t \circ X_s(x) \subset A_t(I(\mathcal{L}_0, x))$ so that $X_s(x) \subset I(\mathcal{L}_0, x)$ for all $s > 0$, and thus $X(x) \subset \mathcal{L}_0(x)$, which completes the proof.

The next results show that a knowledge of $\mathcal{F}^+$ translates into information on the structure of the reachable set at time t.

DEFINITION. Let $\mathcal{S}$ be a subset of $V(M)$ and let $t > 0$. An *integral curve of* $\mathcal{S}$ *on* $[0, t]$ is a continuous mapping $\alpha : [0, t] \to M$ such that there exists a partition $0 = t_0 < t_1 < \cdots < t_k = t$ and vector fields $X_1, \cdots, X_k \in \mathcal{S}$ with the property that on $[t_{i-1}, t_i]$ the curve α is an integral curve of X_i for $i = 1, 2, \cdots, k$. A point y in M is $\mathcal{S}$*-reachable* from $x \in M$ if $\exists$ a $t > 0$ and an integral curve α of $\mathcal{S}$ on $[0, t]$ such that $\alpha(0) = x$ and $\alpha(t) = y$. By Theorem 2.7 $\mathcal{F}^+(x) \subset \mathcal{L}_0(x)$ for all $x \in M$ so that points in M which are

$\mathcal{F}^+$-reachable from x will be contained in $I(\mathcal{L}_0, x)$. Note that $I(\mathcal{L}_0, x)$ is precisely the set of points in M which are $\mathcal{L}_0$-reachable from x_0 (cf. [14]).

THEOREM 2.3. *Consider the system* (1.1) *with triples* $(\mathcal{L}, \mathcal{L}_0, \mathcal{B})$ *and* $(\mathcal{T}, \mathcal{F}^+, \mathcal{F})$. *Suppose that* $Y \in \mathcal{T}$ *is a trace vector field and that* y *is* $\mathcal{F}^+$*-reachable from* x. *Then* $Y_t(y)$ *is contained in* $\overline{\mathcal{R}_t(x)}$ *for all* $t>0$ *and for all* $x \in M$, *where* $\overline{\mathcal{R}_t(x)}$ *is the closure of* $\mathcal{R}_t(x)$ *relative to the topology of* $I(\mathcal{L}_0, A_t(x))$.

COROLLARY 1. *The system* (1.1) *is strongly controllable if and only if* $\mathcal{F}^+(x) = \mathcal{L}_0(x)$ *for all* $x \in M$ *if and only if* $\mathcal{F}(x) = \mathcal{F}^+(x) = \mathcal{L}_0(x)$ *for all* $x \in M$.

COROLLARY 2. *The system* (1.1) *is strongly controllable if and only if for each* $x \in M \exists$ *vector fields* $X_1, \cdots, X_k \in \mathcal{F}^+$ *whose positive linear span at* x *is all of* $\mathcal{L}_0(x)$ (i.e. *for each* $v \in \mathcal{L}_0(x)$ $\exists$ *nonnegative constants* $c_1, \cdots, c_k$ *such that* $v = \sum_{i=1}^k c_i X_i(x)$).

Proof. (Theorem 2.3). Let y be $\mathcal{F}^+$-reachable from x so that $y = X^1_{s_1} \circ X^2_{s_2} \circ \cdots \circ X^k_{s_k}(x)$ for some $X^i \in \mathcal{F}^+$ and $s_i > 0$. Then $(A_\varepsilon \circ X^1_{s_1}) \circ \cdots \circ (A_\varepsilon \circ X^k_{s_k})(x) \in \overline{\mathcal{R}_{k\varepsilon}(x)}$ by Lemma 2.1, and if $Y \in \mathcal{T}$ than $Y_{t-k\varepsilon} \circ (A_\varepsilon \circ X^1_{s_1}) \circ \cdots \circ (A_\varepsilon \circ X^k_{x_k})(x) \in \overline{\mathcal{R}_t(x)}$. Letting $\varepsilon \to 0$, one finds that $Y_t(y) \in \overline{\mathcal{R}_t(x)}$, which completes the proof.

Proof. (Corollary 1). Suppose that (1.1) is strongly controllable. Then $\mathcal{R}_t(x) = I(\mathcal{L}_0, A_t(x)) = A_t(I(\mathcal{L}_0, x))$ and if $X \in \mathcal{L}_0$ then $A_t \circ X_s(x) \in \mathcal{R}_t(x)$ for $s>0$. From Lemma 2.1 $X \in \mathcal{F}^+$, so that $\mathcal{F}^+ \supset \mathcal{L}_0$. From Theorem 2.2 $\mathcal{F}^+(x) \subset \mathcal{L}_0(x)$ so $\mathcal{F}^+(x) = \mathcal{L}_0(x)$. Suppose $\mathcal{L}_0(x) = \mathcal{F}^+(x)$ for all $x \in M$. Then the points in M which are $\mathcal{F}^+$-reachable from x will be $I(\mathcal{L}_0, x)$, and Theorem 2.3 implies that for $y \in I(\mathcal{L}_0, x)$, $A_t(y) \in \overline{\mathcal{R}_t(x)}$, so $\overline{\mathcal{R}_t(x)} = I(\mathcal{L}_0, A_t(x))$ for all $t>0$ and $x \in M$. This implies that $\mathcal{R}_t(x) = I(\mathcal{L}_0, A_t(x))$ (cf. [5]).

Proof. (Corollary 2). This result follows from Corollary 1 and part c) of Lemma 2.1.

THEOREM 2.4. *Consider the system* (1.1) *and suppose that the vector field* $A \in \mathcal{F}^+$. *Then* $\overline{\mathcal{R}_t(x)} = \overline{\mathcal{R}(x)}$ *for all* $t>0$ *and for all* $x \in M$. *In particular the system is strongly controllable if and only if it is controllable.*

Proof. In [5] it is shown that the closure of $\mathcal{R}_t(x)$ is equal to the closure of the set

$$\left\{ A_{t_1} \circ B_{s_1} \circ \cdots \circ A_{t_k} \circ B_{s_k}(x) \mid k = 1, 2, \cdots;\ t_i > 0;\ s_i \in R;\ \sum_{i=1}^{k} t_i = t \right\}$$

and $\overline{\mathcal{R}(x)}$ is equal to the closure of

$$\{A_{t_1} \circ B_{s_1} \circ \cdots \circ A_{t_k} \circ B_{s_k}(x) \mid k = 1, 2, \cdots;\ t_i > 0;\ s_i \in R\}.$$

If $A \in \mathcal{F}^+$ then $A_t \circ A_s(x) \in \overline{\mathcal{R}_t(x)}$ for all $s>0$ so that $A_{t+s}(x) \in \overline{\mathcal{R}_t(x)}$ for all $s>0$. This implies that $\overline{\mathcal{R}_t(x)} = \overline{\mathcal{R}(x)}$ for all $t>0$.

COROLLARY 1. *If* (1.1) *is a right-invariant system on a compact Lie group* G (cf. [13]) *and* $A \in \mathcal{F}^+$, *then the system is strongly controllable.*

Proof. For systems which are described by right-invariant vector fields and whose state space is a compact Lie group G it is known that $\mathcal{R}(x) = I(\mathcal{L}, x)$ (cf. [13]). If $A \in \mathcal{F}^+$ then $\mathcal{L} = \mathcal{L}_0$ and Theorem 2.4 implies that $\overline{\mathcal{R}_t(x)} = \overline{I(\mathcal{L}_0, x)} = I(\mathcal{L}_0, x) = I(\mathcal{L}_0, A_t, x)$ so that the system is strongly controllable.

To take advantage of the above results, one needs a way to find vector fields in $\mathcal{F}^+$ for the system (1.1). At this stage one knows that

$$\mathcal{F}^+ \supset \mathcal{F} \supset \mathcal{B}$$

and that $\mathcal{T} \supset \{A + \mathcal{F}^+\} \supset \{A + \mathcal{B}\}$. The next section provides a way of generating more vector fields in $\mathcal{F}^+$ given a complete vector field $X \in \mathcal{F}$ and a vector field $Y \in \mathcal{T}$.

3. Construction of controlled vector fields. Suppose that $X \in \mathcal{F}$ is a complete vector field and $Y \in \mathcal{T}$ is a trace vector field. Since X is complete, the map $x \to X_s(x)$ is a

diffeomorphism from M onto itself for each fixed $s \in R$. If $\phi(x) = X_s(x)$, then $d\phi$ is a linear isomorphism of tangent spaces and $x \to (\phi_* Y)(x) = d\phi_{\phi^{-1}(x)} Y(\phi^{-1}(x))$ defines a real analytic vector field on M which will be denoted by $dX_s Y$. Furthermore $dX_s Y \in \mathcal{T}$ for all $s \in R$. To see this, note that $t \to X_s \circ Y_t \circ X_{-s}(x)$ is the integral curve for $dX_s Y$ through x. If $y = X_{-s}(x)$, then y is $\mathcal{F}^+$-reachable from x (here $X \in \mathcal{F} \subset \mathcal{F}^+$) so that $Y_t(X_{-s}(x)) \in \overline{\mathcal{R}_t(x)}$ by Theorem 2.3. If $0 < \varepsilon < t$, then it follows that $z = Y_{t-\varepsilon}(X_{-s}(x)) \in \overline{\mathcal{R}_{t-\varepsilon}(x)}$, $Y_\varepsilon(X_s(z)) \in \overline{\mathcal{R}_\varepsilon(z)}$ so that $Y_\varepsilon \circ X_s \circ Y_{t-\varepsilon} \circ X_{-s}(x) \in \overline{\mathcal{R}_t(x)}$ for all ε between zero and t. Letting $\varepsilon \to 0$, one finds that $X_s \circ Y_t \circ X_{-s}(x)$ is in the closure of $\mathcal{R}_t(x)$ for any $t > 0$ and any $s \in R$, and thus $dX_s Y \in \mathcal{T}$ for all real s.

The observation that if $X \in \mathcal{F}$ is complete and $Y \in \mathcal{T}$ then $dX_s Y \in \mathcal{T}$ for real s forms the basis for many of the results of Kunita in [8]. In the case where $\mathcal{L} = \mathcal{L}_0$ his results can be used to show that if, for some positive integer k, one has $\operatorname{ad}_X^{k+1} Y = Z \in \mathcal{F}$ then $\operatorname{ad}_X^k Y \in \mathcal{F}^+$. More generally one knows that, at least locally, $\exists$ some positive integer k and functions $a_0, a_1, \cdots, a_k \in C^\omega(M)$ such that

$$\operatorname{ad}_X^{k+1} Y = a_0 Y + a_1 \operatorname{ad}_X Y + \cdots + a_k \operatorname{ad}_X^k Y + Z \tag{3.1}$$

where $Z \in \mathcal{F}$ (i.e., take $Z = 0$ and $k = n-1$ where $n = \dim M$). The case where the functions a_i vary with x but are constant along the integral curves of X is considered later in this section. For now it will be assumed that X and Y are such that $a_0, \cdots, a_k$ are constants. In this case the least positive integer k for which (3.1) holds with a_i constant functions on M and $Z \in \mathcal{F}$ is called the *controllability index of* (X, Y), $k(X, Y)$. If (3.1) never holds with a_i constants then we set $k(X, Y) = \infty$.

If $k(X, Y) < \infty$ then one can construct from $dX_s Y$ a new family of vector fields W^s with the property that $W^s \in \mathcal{T}$ for all real s and $\operatorname{span}\{W^s | s \in R\}$ is a finite dimensional subspace of $V(M)$. Since X, Y are real analytic vector fields one can expand $dX_{-s} Y$ in a Taylor series about $s = 0$. It is well known that $dX_{-s} Y = \sum_{j=0}^{\infty} (s^j/j!) \operatorname{ad}_X^j Y$ for all s in R, where ad_X is a linear operator on $V(M)$, an infinite dimensional real vector space. For $k = k(X, Y)$ it is not difficult to verify that

$$V = \operatorname{span}\{Y, \operatorname{ad}_X Y, \cdots, \operatorname{ad}_X^k Y\}$$

is a $(k+1)$-dimensional subspace of $V(M)$ and that

$$T(\operatorname{ad}_X^i Y) = \begin{cases} \operatorname{ad}_X^{i+1} Y & \text{for } 1 \leqq i < k, \\ \sum_{i=0}^{k} a_i \operatorname{ad}_X^i Y & \text{for } i = k \end{cases} \tag{3.2}$$

define a linear operator on V. It follows that $\exp sT = I + sT + (s^2/2!) T^2 + \cdots$ defines a linear operator on V for all real s, and

$$W^s = (\exp sT)(Y) = \sum_{i=0}^{\infty} \frac{s^i}{i!} T^i(Y) \tag{3.3}$$

defines a vector field which is in V for all real s. To see that $W^s \in \mathcal{T}$, note that the power series expansions for $dX_{-s} Y$ and W^s have the same coefficients for $1, s, \cdots, s^k$ as a consequence of (3.2). For $dX_{-s} Y$ the coefficient of $s^{k+1}/(k+1)!$ is

$$\operatorname{ad}_X^{k+1} Y = a_0 Y + a_1 \operatorname{ad}_X Y + \cdots + a_k \operatorname{ad}_X^k Y + Z$$

and for W^s the coefficient of $s^{k+1}/(k+1)!$ is

$$T^{k+1}(Y) = a_0 Y + a_1 \operatorname{ad}_X Y + \cdots + a_k \operatorname{ad}_X^k Y$$

by (3.2). These coefficients differ by Z, a freely-controlled vector field in $\mathcal{F}$. Similarly

the coefficients of $s^{k+2}/(k+2)!$ in the power series expansions for $dX_{-s}Y$ and W^s differ by $(\mathrm{ad}_X Z + a_k Z)$, which is also freely controlled, since $X, Z \in \mathcal{F}$ and $\mathcal{F}$ is a Lie algebra by Theorem 2.2. A simple induction argument shows that the coefficients of an arbitrary power of s in the series expansions for $dX_{-s}Y$ and W^s differ by a vector field in $\mathcal{F}$. Thus $W^s = dX_{-s}Y + Z^s$ where $Z^s \in \mathcal{F}$, $dX_{-s}Y \in \mathcal{T}$. By Theorem 2.2 $\mathcal{F} \subset \mathcal{F}^+$ and $\mathcal{F}^+ + \mathcal{T} \subset \mathcal{T}$ so that $W^s \in \mathcal{T}$ for all real s.

The fact that $W^s \in \mathcal{T}$ for all real s can be used to construct semi-controlled vector fields for the system (1.1).

Since $W^s \in \mathcal{T}$ and $A \in \mathcal{T}$, then by Theorem 2.2 it follows that $(1-b)A + bW^s \in \mathcal{T}$ for all $b > 0$. It turns out that one can choose s to be a function of b in such a way that as b approaches 0, the vector field $bW^{s(b)}$ approaches a nonzero vector field W. This means that $A + W$ will be in $\mathcal{T}$, and if $s(b)$ is replaced by a positive multiple $\alpha s(b)$, then $A + \alpha W \in \mathcal{T}$ for all $\alpha \geqq 0$. In other words W is, by definition, a semi-controlled vector field in $\mathcal{F}^+$. The method for extracting W from W^s will now be explained.

LEMMA 3.1. *Suppose that X is a complete freely-controlled vector field for the system* (1.1) *and Y is a trace vector field such that $k(X, Y) < \infty$. Then for each s in R*

$$W^s = \sum_{i=0}^{k} b_i^s \, \mathrm{ad}_X^i \, Y \tag{3.4}$$

is a trace vector field for (1.1) *and the real numbers b_i^s are the entries in the first column of the matrix exponential $e^{sQ(X,Y)}$ where $Q(X, Y)$ is the $(k+1)\times(k+1)$ matrix*

$$Q(X, Y) = \begin{bmatrix} 0 & 0 & \cdots & 0 & a_0 \\ 1 & 0 & \cdots & 0 & a_1 \\ 0 & 1 & \cdots & 0 & a_2 \\ \vdots & & \vdots & & \vdots \\ 0 & \cdots & 0 & 1 & a_k \end{bmatrix} \tag{3.5}$$

with $k = k(X, Y)$ and $a_0, a_1, \cdots, a_k$ defined by (3.1).

Proof. By definition $W^s = (\exp sT)(Y) \in V$ for all $s \in R$ where $V = \mathrm{span}\{Y, \mathrm{ad}_X Y, \cdots, \mathrm{ad}_X^k Y\}$. Thus $\beta = \{Y, \mathrm{ad}_X Y, \cdots, \mathrm{ad}_X^k Y\}$ is an ordered basis for V and from the definition of T, equation (3.2), it follows that the matrix representation for T with respect to β is $Q(X, Y)$. Thus, in β-coordinates, one can represent W^s by the $(k+1)$ by 1 coordinate matrix $[W^s]_\beta$ with entries b_i^s (i.e. $W^s = \sum_{i=0}^k b_i^s \, \mathrm{ad}_X^i \, Y$). Since $W^s = (\exp sT)(Y)$, one has

$$[W^s]_\beta = [(\exp sT)\, Y]_\beta = [\exp sT]_\beta [Y]_\beta.$$

Now $[Y]_\beta$ is the transpose of $[1, 0, \cdots, 0]$ and $[\exp sT]_\beta = e^{s[T]_\beta} = e^{sQ(X,Y)}$ so that the coordinate matrix of W^s with respect to β is the first column of the matrix $e^{sQ(X,Y)}$. This completes the proof.

The rate at which the terms b_i^s from (3.4) grow with respect to s can be studied by using the Cauchy integral formula to evaluate $e^{sQ(X,Y)}$ (cf. [2]). If $\lambda_1, \lambda_2, \cdots, \lambda_d$ are the distinct eigenvalues of $Q(X, Y)$ with multiplicities $m_1, m_2, \cdots, m_d$, then

$$e^{sQ(X,Y)} = \sum_{j=1}^{d} \sum_{l=0}^{m_j - 1} s^l \, e^{s\lambda_j} Q_{jl} \tag{3.6}$$

where

$$Q_{j,l} = \frac{1}{l!(m_j - 1 - l)!} \frac{d^{m_j - 1 - l}}{dz^{m_j - 1 - l}} [(z - \lambda_j)^{m_j} (Iz - Q(X, Y))^{-1}]_{z = \lambda_j}.$$

Let a_{MAX} (a_{MIN}) denote the maximum (minimum) real part of the eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_d$. Relabel these eigenvalues so that among those eigenvalues with real part a_{MAX} the ones with maximum multiplicity are $\lambda_1, \cdots, \lambda_q$. Then among all eigenvalues of $Q(X, Y)$ with real part a_{MAX} the maximum multiplicity is m_1 which is the multiplicity of $\lambda_1, \lambda_2, \cdots, \lambda_q$. The remaining eigenvalues are relabelled so that among those with real part a_{MIN} the ones with maximum multiplicity are $\lambda_p, \cdots, \lambda_d$. Then the sum (3.6) for $e^{sQ(X,Y)}$ can be rewritten as

$$e^{sQ(X,Y)} = \sum_{j=1}^{q} s^{m_1-1} e^{s\lambda_j} Q_{j(m_1-1)} + \sum_{j=p}^{d} s^{m_d-1} e^{s\lambda_j} Q_{j(m-1)} + R(s)$$

where $R(s)$ is the matrix formed by the remaining double sums in (3.6). The first two sums in the above expression can be simplified by noting that $\lambda_1, \cdots, \lambda_q$ are of the form $\lambda_j = a_{\mathrm{MAX}} + ib_j$ and $\lambda_p, \cdots, \lambda_d$ are of the form $\lambda_j = a_{\mathrm{MIN}} + ib_j$. Since

$$e^{s\lambda_1} = e^{s(a_{\mathrm{MAX}}+ib_1)} = e^{sa_{\mathrm{MAX}}} e^{sib_1}$$

the above sum can be written as

$$e^{sQ(X,Y)} = s^{m_1-1} e^{sa_{\mathrm{MAX}}} D_{\mathrm{MAX}}(s) + s^{m_d-1} e^{sa_{\mathrm{MIN}}} D_{\mathrm{MIN}}(s) + R(s) \tag{3.7}$$

where

$$D_{\mathrm{MAX}}(s) = \sum_{j=1}^{q} e^{sib_j} Q_{j(m_1-1)} \quad \text{and} \quad D_{\mathrm{MIN}}(s) = \sum_{j=P}^{d} e^{sib_j} Q_{j(m_d-1)}.$$

Note that $D_{\mathrm{MAX}}(s)$, $D_{\mathrm{MIN}}(s)$ and $R(s)$ are periodic with respect to s with period 2π because $e^{sib_j} = \cos sb_j + i \sin sb_j$. This decomposition results in a corresponding decomposition for W^s as a consequence of Lemma 3.1.

LEMMA 3.2. *Suppose that X is a complete, freely-controlled vector field for the system* (1.1) *and Y is a trace vector field with $k(X, Y) < \infty$. Then for each real s the trace vector field W^s can be decomposed as the sum*

$$W^s = s^{m_1-1} e^{sa_{\mathrm{MAX}}} W^s_{\mathrm{MAX}} + s^{m_d-1} e^{sa_{\mathrm{MIN}}} W^s_{\mathrm{MIN}} + R^s \tag{3.8}$$

where W^s_{MAX} and W^s_{MIN} are vector fields which are periodic in s with period 2π. The vector fields W^s_{MAX}, W^s_{MIN}, and R^s are defined by $\sum_{j=0}^{k(X,Y)} c_j^s \operatorname{ad}_X^j Y$ where the scalars c_j^s are the first columns of the matrices $D_{\mathrm{MAX}}(s)$, $D_{\mathrm{MIN}}(s)$, and $R(s)$ respectively.

Proof. $W^s = \sum_{i=0}^{k} b_i^s \operatorname{ad}_X^i Y$ by Lemma 3.1, where $k = k(X, Y)$, and b_i^s are the entries in the first column of $e^{sQ(X,Y)}$. Using (3.7) to decompose $e^{sQ(X,Y)}$ results in (3.8). The next theorem shows that the vector fields W^s_{MAX} and W^s_{MIN} are semi-controlled.

THEOREM 3.3. *Suppose that X is a complete, freely-controlled vector field for the system* (1.1) *and Y is a trace vector field with $k(X, Y) < \infty$. Consider the trace vector field W^s decomposed in* (3.8) *as the sum*

$$s^{m_1-1} e^{sa_{\mathrm{MAX}}} W^s_{\mathrm{MAX}} + s^{m_d-1} e^{sa_{\mathrm{MIN}}} W^s_{\mathrm{MIN}} + R^s.$$

Then,

(i) *If $a_{\mathrm{MAX}} > 0$ then $W^s_{\mathrm{MAX}} \in \mathcal{F}^+$ for all $s \in R$.*
(ii) *If $a_{\mathrm{MIN}} < 0$ then $(-1)^{m_d-1} W^s_{\mathrm{MIN}} \in \mathcal{F}^+$ for all $s \in R$.*
(iii) *If $a_{\mathrm{MAX}} = 0$ and $m_1 > 1$ then $W^s_{\mathrm{MAX}} \in \mathcal{F}^+$ for all $s \in R$.*
(iv) *If $a_{\mathrm{MIN}} = 0$ and $m_d > 1$ then $(-1)^{m_d+1} W^s_{\mathrm{MIN}} \in \mathcal{F}^+$ for all $s \in R$.*

Proof. Set $k = k(X, Y) < \infty$ and suppose $a_{\mathrm{MAX}} > 0$. Let $r_1, \cdots, r_l$ denote the distinct real parts of the eigenvalues $\lambda_1, \cdots, \lambda_d$ of $Q(X, Y)$ in decreasing order, so that

$r_1 = a_{\text{MAX}}$ and $r_l = a_{\text{MIN}}$. Using (3.8) for W^s one sees that

$$s^{1-m_1} e^{(-sa_{\text{MAX}})} W^s - W^s_{\text{MAX}} = s^{m_d-m_1} e^{s(a_{\text{MIN}}-a_{\text{MAX}})} W^s_{\text{MIN}} + s^{1-m_1} e^{-sa_{\text{MAX}}} R^s.$$

Set $\alpha(s) = s^{1-m_1} e^{-sa_{\text{MAX}}}$ and set $\hat{W}^s = \alpha(s) W^s$. Since $a_{\text{MIN}} - a_{\text{MAX}} < 0$ here and W^s_{MIN} is periodic in s, it follows that

$$\lim_{s\to\infty} s^{m_d-m_1} e^{s(a_{\text{MIN}}-a_{\text{MAX}})} W^s_{\text{MIN}} = 0$$

and similarly $\lim_{s\to\infty} \alpha(s) R^s = 0$. Thus as s tends to infinity, $\hat{W}^s(x)$ approaches $W^s_{\text{MAX}}(x)$ uniformly with respect to x. Also since $\hat{W}^s$ is a multiple of W^s by $\alpha(s)$, it follows that for $\tau > 0$, the integral curve $\hat{W}^s_\tau(x) = W^s_{\sigma(\tau,s)}(x)$ for some real analytic function $\sigma(\tau, s)$ on $R^+ \times R$. Because $\alpha(s) \to 0$ as $s \to \infty$, the function $\sigma(\tau, s) \to 0$ uniformly in τ as $s \to \infty$, and because $W^s \in \mathcal{T}$, $W^s_{\sigma(\tau,s)}(x) \in \overline{\mathcal{R}_{\sigma(\tau,s)}(x)}$. This means that $\hat{W}^s_\tau(x) \in \overline{\mathcal{R}_{\sigma(\tau,s)}(x)}$ and

$$A_{t-\sigma(\tau,s)} \circ \hat{W}^s_\tau(x) \in \overline{\mathcal{R}_t(x)} \quad \text{for } \sigma(\tau, s) < t. \tag{3.9}$$

Because $\hat{W}^s(x)$ approaches $W^s_{\text{MAX}}(x)$ uniformly in x as $s \to \infty$, the integral curves for $\hat{W}^s$ and W^s_{MAX} approach one another as $s \to \infty$. Fix $s_0 \in R$ and set $s_n = s_0 + 2\pi n$ for n a positive integer. Since W^s_{MAX} is periodic with respect to s with period 2π, it follows that $W^{s_n}_{\text{MAX}} = W^{s_0}_{\text{MAX}}$ and $\lim_{n\to\infty} \hat{W}^{s_n}_\tau(x) = (W^{s_0}_{\text{MAX}})_\tau(x)$ for $\tau > 0$. From (3.9) $A_{t-\sigma(\tau,s_n)} \circ \hat{W}^{s_n}_\tau(x) \in \overline{\mathcal{R}_t(x)}$ for $\alpha(\tau, s_n) < t$, and since $\lim_{n\to\infty} \sigma(\tau, s_n) = 0$ it follows that

$$A_t \circ \lim_{n\to\infty} \hat{W}^{s_n}_\tau(x) = A_t \circ (W^{s_0}_{\text{MAX}})_\tau(x) \in \overline{\mathcal{R}_t(x)}$$

for all $x \in M$ and for all $\tau > 0$. Lemma 2.1 implies that $W^{s_0}_{\text{MAX}} \in \mathcal{F}^+$. This completes the proof of part (i).

Assume $a_{\text{MIN}} < 0$. Using (3.8) one sees that

$$s^{1-m_d} e^{-sa_{\text{MIN}}} W^s - W^s_{\text{MIN}} = s^{m_1-m_d} e^{s(a_{\text{MAX}}-a_{\text{MIN}})} W^s_{\text{MAX}} + s^{1-m_d} e^{-sa_{\text{MIN}}} W^s.$$

Set $\alpha(s) = s^{1-m_d} e^{-sa_{\text{MAX}}}$ and set $\hat{W}^s = \alpha(s) W^s$. Here $a_{\text{MAX}} - a_{\text{MIN}}$ is positive so $\lim_{s\to-\infty} s^{m_1-m_d} e^{s(a_{\text{MAX}}-a_{\text{MIN}})} W^s_{\text{MAX}} = 0$ and the proof of part (i) can be repeated with $s \to -\infty$ instead of $+\infty$. If $m_d - 1$ is odd then for s negative $\hat{W}^s$ and W^s will have the opposite signs which accounts for the $(-1)^{m_d+1}$ term in (ii).

Suppose $a_{\text{MAX}} = 0$ and $m_1 > 0$. Since $r_1 = 0$, the numbers $r_2, \cdots, r_l$ must be negative so that $s^{m_d-m_1} e^{s(a_{\text{MIN}}-a_{\text{MAX}})}$ tends to zero as $s \to +\infty$ and the proof used in part (i) can be repeated.

Finally, suppose $a_{\text{MIN}} = 0$ and $m_d > 1$. Then

$$\lim_{s\to-\infty} s^{m_1-m_d} e^{s(a_{\text{MAX}}-a_{\text{MIN}})} = 0$$

and the proof of part (ii) can be used. This completes the proof.

Remark 1. If M is a Lie group and $X \in \mathcal{F}$ and $Y \in \mathcal{T}$ are both right (left)-invariant vector fields on M then $k(X, Y) < \infty$ because $\{X, Y\}_{LA}$ is a finite dimensional Lie algebra over R.

Remark 2. Suppose $X \in \mathcal{F}$ is complete, $Y \in \mathcal{T}$, $k(X, Y) = k < \infty$ and the eigenvalues of $Q(X, Y)$ are all identically 0 (this is a situation where the results of Kunita [8] can be used to show that $\text{ad}^k_X Y$ and $(-1)^k \text{ad}^k_X Y$ are semi-controlled). In this case Theorem 3.3 part (iii) results in $W^s_{\text{MAX}} = \text{ad}^k_X Y \in \mathcal{F}^+$ and part (iv) yields $(-1)^k \text{ad}^k_X Y \in \mathcal{F}^+$.

Remark 3. The assumption that X be complete in this theorem and in Theorem 3.5 cannot be omitted. For example the systems

$$\dot{x}_1 = u \text{ with } x_1, x_2 \in R \quad \text{and} \quad \dot{y}_1 = u \text{ with } y_1 \in (-1, 1),$$
$$\dot{x}_2 = x_1, \qquad\qquad \dot{y}_2 = y_1 \text{ and } y_2 \in R$$

have identical algebraic properties. They generate isomorphic Lie algebras but only the first is strongly controllable. This is because the vector field $B(x) = (1, 0)$ is not complete when y_1 is restricted to $(-1, 1)$.

Remark 4. If $\hat{K}$ is any positive integer greater than or equal to $K(X, Y)$, then one could replace $K(X, Y)$ by $\hat{K}$ and repeat Lemmas 3.1 and 3.2 to generate $\hat{W}^s$, $\hat{a}_{\text{MAX}}$, $\hat{a}_{\text{MIN}}$ and $\hat{W}^s_{\text{MAX}}$, $\hat{W}^s_{\text{MIN}}$. Then Theorem 3.3 is still true. This means that one does not need to use the smallest k such that (3.1) holds.

Example 3.4. Consider the system modelled by

$$\dot{x} = A(x) + uB(x), \qquad x \in M$$

where $M = (0, \infty) \times S^1 = \{(x_1, x_2, x_3) \in R^3 | x_1 > 0, x_2^2 + x_3^2 = 1\}$, $A(x_1, x_2, x_3) = (1, 0, 0)$, and $B(x_1, x_2, x_3) = (x_1, x_3, -x_2)$. Since A is a trace vector field and B is a complete freely-controlled vector field one can set $X = B$, $Y = A$ and compute $k(X, Y)$. Here $\text{ad}_X\, Y(x) = (dA)_x B(x) - (dB)_x A(x) = -A(x) = -Y(x)$. Using (3.1) $a_0 = -1$, $k(X, Y) = 0$ and $Q(X, Y) = [a_0] = [-1]$ from (3.5). The eigenvalues for $Q(X, Y)$ consist of $\lambda_1 = -1$ so $d = 1$, $a_{\text{MAX}} = a_{\text{MIN}} = -1$, $m_1 = 1$, and (3.8) becomes $W^s = e^{-s} Y = e^{-s} A = (e^{-s}, 0, 0)$ where one can consider $W^s_{\text{MAX}} = A$, $W^s_{\text{MIN}} = 0$, $R^s = 0$ or take $W^s_{\text{MAX}} = 0$, $W^s_{\text{MIN}} = A$, $R^s = 0$. In the latter case Theorem 3.3 part (ii) asserts $(-1)^2 W^s_{\text{MIN}} = A \in \mathcal{F}^+$. Thus $A + \alpha A$ is a trace vector field for all $\alpha \in [0, \infty)$, which means that $((1+\alpha)A)_t(x)$ is in the closure of $\mathcal{R}_t(x)$ for all $\alpha > 0$ or equivalently $A_s(x) \in \overline{\mathcal{R}_t(x)}$ for all $s \geqq t$. Usually one knows that $A_t(x) \in \overline{\mathcal{R}_t(x)}$, but in this case even as t approaches zero, one can travel arbitrarily far along the A-integral curve. Geometrically one sees this since the leaves of the B distribution crowd together as $s \to -\infty$ and for "small t" $A_t(x)$ passes through "many B leaves", see Fig. 1.

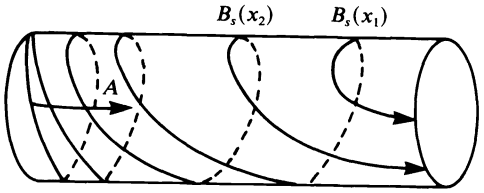


FIG. 1

Example 3.5. Consider the system (1.1) with $A(x) = (x_1 x_2, x_1, e^{x_1} x_2)$ and $B(x) = (x_2^2, 0, 0)$ where $M = R^3$. Here the triple of Lie algebras $(\mathcal{L}, \mathcal{L}_0, \mathcal{B})$ has $\mathcal{L}_0$ infinite-dimensional, $\mathcal{B}$ one-dimensional and we know that $X = B \in \mathcal{F}$ is complete and $Y = A \in \mathcal{T}$. In order to use Theorem 3.3 to find new semi-controlled vector fields, the controllability index $k(X, Y)$ must be finite. Here $\text{ad}_X\, Y = (x_2^3 - 2x_1 x_2, x_2^2, x_2^3 e^{x_1})$, $\text{ad}_X^2\, Y = (-4x_2^3, 0, x_2^5 e^{x_1})$, $\text{ad}_X^3\, Y = (0, 0, x_2^7 e^{x_1})$ and it is easy to see that there is no positive constant k such that $\text{ad}_X^{k+1}\, Y$ is a *constant* linear combination of $Y, \text{ad}_X\, Y, \cdots, \text{ad}_X^k\, Y$ so that $k(X, Y) = \infty$. It is true, however, that $\text{ad}_X^3\, Y(x) = x_2^2\, \text{ad}_X^2\, Y(x) - 4x_2 X(x)$. From Theorem 2.1 $Z = -4x_2 X \in \mathcal{F}$, and if $a_2(x) = x_2^2$ then

$$\text{ad}_X^3\, Y(x) = a_2(x)\, \text{ad}_X^2\, Y + Z$$

where $Z \in \mathcal{F}$. It is true that a_2 is not a constant, but $Xa_2(x) = (da_2)_x X(x) = 0$ for all

$x \in M$ so that a_2 is *constant along integral curves for* X. This case is considered in Theorem 3.5, and this example will be continued after Theorem 3.5.

Suppose that $X \in \mathcal{F}$ is complete, $Y \in \mathcal{T}$ and for some positive integer k

$$\operatorname{ad}_X^{k+1} Y = a_0 Y + \cdots + a_k \operatorname{ad}_X^k Y + Z \tag{3.10}$$

where $Z \in \mathcal{F}$, $a_i \in C^\omega(M)$, and $(Xa_i)(x) = (da_i)_x X(x) = 0$ for all $x \in M$ (i.e. the functions a_i are constant along the integral curves of the vector field X). The least positive integer k for which (3.10) holds is called the *local controllability index of* (X, Y), $k_*(X, Y)$. If (3.10) never holds, then $k_*(X, Y) = \infty$. Clearly $k_*(X, Y) < k(X, Y)$ and in Example 3.5 $k_*(X, Y) = 2$ while $k(X, Y) = \infty$.

Following the steps used before, let $k = k_*(X, Y)$, set $V = \operatorname{span}\{Y, \cdots, \operatorname{ad}_X^k Y\}$, and for a fixed $y \in M$ let T^y be the linear operator on V defined by

$$T^y(\operatorname{ad}_X^i Y) = \begin{cases} \operatorname{ad}_X^{i+1} Y & \text{for } 1 \leqq i < k, \\ \sum_{i=0}^{k} a_i(y) \operatorname{ad}_X^i Y & \text{for } i = k \end{cases}$$

so that $(\exp sT^y)(Y)$ is a vector field in V. This means that $x \to (\exp sT^y)(Y)(x)$ defines a vector field on M and it follows that $x \to (\exp sT^x)(Y)(x)$ defines a vector field in $V(M)$ which will be called W_*^s. Because $Xa_i \equiv 0$ one can repeat the same argument which was used earlier in this section to show that the W_*^s is a trace vector field for all real s. As in Lemma 3.1 $W_*^s(x) = \sum_{i=0}^{k_*(X,Y)} b_i^s(x) \operatorname{ad}_X Y(x)$ where $b_i^s(x)$ is the first column of the matrix exponential $e^{sQ_x(X,Y)}$ with

$$Q_x(X, Y) = \begin{bmatrix} 0 & 0 & \cdots & 0 & a_0(x) \\ 1 & 0 & \cdots & 0 & a_1(x) \\ 0 & 1 & & & \\ \vdots & & \ddots & & \vdots \\ 0 & & \cdots & 1 & a_k(x) \end{bmatrix}. \tag{3.11}$$

One can get a decomposition for $e^{sQ_x(X,Y)}$ similar to the expression (3.7) for $e^{sQ(X,Y)}$ by restricting x to an open dense submanifold of M. Let $d(x)$ denote the number of distinct eigenvalues of $Q_x(X, Y)$ and $l(x)$ the number of distinct real parts of the eigenvalues of $Q_x(X, Y)$. Define d and l to be the maximum values of $d(x)$ and $l(x)$ respectively as x varies over M and set

$$M(X, Y) = \{x \in M \mid d(x) = d \text{ and } l(x) = l\}.$$

Because the entries in $Q_x(X, Y)$ are real analytic in x, it follows that $M(X, Y)$ is an open dense submanifold of M. For each $x \in M(X, Y)$ let $a_{\mathrm{MAX}}(x)(a_{\mathrm{MIN}}(x))$ denote the maximum (minimum) real part of the eigenvalues $\lambda_1(x), \cdots, \lambda_d(x)$ of $Q_x(X, Y)$ which have multiplicities $m_1(x), \cdots, m_d(x)$. Relabel these eigenvalues so that among those with real part $a_{\mathrm{MAX}}(x)$ the ones with maximum multiplicity are $\lambda_1(x), \cdots, \lambda_{q(x)}(x)$ and among those eigenvalues with real part $a_{\mathrm{MIN}}(x)$ the ones with minimum multiplicity are $\lambda_{p(x)}, \cdots, \lambda_d(x)$. While $p(x)$, $q(x)$ and $m_i(x)$ may vary with x, it is straightforward to show that these functions are constant on the connected components of $M(X, Y)$. Thus the decomposition (3.7) for $e^{sQ(X,Y)}$ can be used on the connected components of $M(X, Y)$. Thus on $M(X, Y)$

$$e^{sQ_x(X,Y)} = s^{m_1(x)-1} e^{sa_{\mathrm{MAX}}(x)} D_{\mathrm{MAX}}(s, x) + e^{m_d(x)-1} e^{sa_{\mathrm{MIN}}(x)} D_{\mathrm{MIN}}(s, x) + R(s, x) \tag{3.12}$$

where D_{MAX} and D_{MIN} are defined as in (3.7) but are x-dependent as $Q_x(X, Y)$ varies with x.

LEMMA 3.4. *Suppose that X is a complete, freely-controlled vector field for the system* (1.1) *and Y is a trace vector field with $k_*(X, Y) < \infty$. Then on the open dense submanifold $M(X, Y)$ of M the trace vector field W^s_* can be decomposed as the sum*

$$W^s_*(x) = s^{m_1(x)-1} e^{sa_{\mathrm{MAX}}(x)} W^s_{*\mathrm{MAX}}(x) + s^{m_d(x)-1} e^{sa_{\mathrm{MIN}}(x)} W^s_{*\mathrm{MIN}}(x) + R^s(x). \tag{3.13}$$

The real analytic vector fields $W^s_{*\mathrm{MAX}}$ and $W^s_{*\mathrm{MIN}}$ are periodic with respect to s with period 2π, and $W^s_{*\mathrm{MAX}}(x)$, $W^s_{*\mathrm{MIN}}(x)$ and $R^s(x)$ are defined by $\sum_{j=0}^{k_*(X,Y)} c^s_j(x)\, \mathrm{ad}^j_X\, Y(x)$ where the scalars $c^s_j(x)$ are the first columns of the matrices $D_{\mathrm{MAX}}(s, x)$, $D_{\mathrm{MIN}}(s, x)$ and $R(x, s)$ respectively.

Proof. The proof is the same as the proof for Lemma 3.2.

THEOREM 3.5. *Suppose that X is a complete, freely-controlled vector field for the system* (1.1) *and Y is a trace vector field with $k_*(X, Y) < \infty$. Consider the trace vector field W^s_* decomposed as in* (3.13) *as the sum $s^{m_1-1} e^{sa_{\mathrm{MAX}}} W^s_{*\mathrm{MAX}} + s^{m_d-1} e^{sa_{\mathrm{MIN}}} W^s_{*\mathrm{MIN}} + R^s$ on $M(X, Y)$, an open dense submanifold of M. Then for the system* (1.1) *restricted to $M(X, Y)$*

(i) *if $a_{\mathrm{MAX}} > 0$ then $W^s_{*\mathrm{MAX}} \in \mathcal{F}^+$ for all $s \in R$,*

(ii) *if $a_{\mathrm{MIN}} < 0$ then $(-1)^{m_d+1} W^s_{*\mathrm{MIN}} \in \mathcal{F}^+$ for all $s \in R$,*

(iii) *if $a_{\mathrm{MAX}} \equiv 0$ and $m_1 > 1$ then $W^s_{*\mathrm{MAX}} \in \mathcal{F}^+$ for all $s \in R$,*

(iv) *if $a_{\mathrm{MIN}} \equiv 0$ and $m_d > 1$ then $(-1)^{m_d+1} W^s_{*\mathrm{MIN}} \in \mathcal{F}^+$ for all $s \in R$.*

Proof. It suffices to prove the theorem on a connected component M_0 of $M(X, Y)$. Set $k = k_*(X, Y)$. To prove (i) the proof of Theorem 3.3 will be modified to take account of the fact that while m_1 and m_d are constant in M_0 the eigenvalues $\lambda_1, \cdots, \lambda_d$ of $Q_x(X, Y)$ vary with x. Fix $y \in M$ and choose a compact *nbhd* U_y of y and $\varepsilon_y > 0$ so that $r_1(x) = a_{\mathrm{MAX}}(x) > \varepsilon_y$ for $x \in U_y$ and $r_1(x) - r_j(x) > \varepsilon_y$ for all $x \in U_y$ and $j = 1, \cdots, l$. Here $\{r_i\}$ are the real parts of $\{\lambda_i\}$. Now let V_y be an open *nbhd* of y contained in U_y such that the integral curve of the vector field $\hat{W}^s$, which is defined in the proof of Theorem 3.3, has the property that for some $b_y > 0$, $\hat{W}^s_\tau(x) \in U_y$ $\forall x \in V_y$ $\forall s > 0$ and $\forall \tau \in [0, b_y)$. For $x \in V_y$ the proof of Theorem 3.3 can be repeated exactly to conclude that $A_t \circ (W^s_{\mathrm{MAX}})_\tau(x) \in \overline{\mathcal{R}_t(x)}$ for all $x \in V_y$, $t > 0$ and $\tau \in [0, b_y)$. To complete the proof of part (i) let $\sigma > 0$ and choose an open covering of the curve $\{(W^s_{\mathrm{MAX}})_\tau(x) | 0 \leqq \tau \leqq \sigma\}$ by sets of the form U_y defined above. Choose a finite subcover. Use the above result on each U_y in the subcover to conclude $A_t \circ (W^s_{\mathrm{MAX}})_\tau(x) \in \overline{\mathcal{R}_t(x)}$ for all $\tau > 0$ (where the integral curve is defined). This completes the proof of (i). The same modification works for parts (ii), (iii) and (iv).

Example 3.5 (*continued*). In this example $\mathrm{ad}^3_X\, Y(x) = a_2(x)\, \mathrm{ad}^2_X\, Y + Z$ where $a_2(x) = x_2^2$ and $Xa_2(x) = (da_2)_x X(x) = [0 \;\; 2x_2 \;\; 0]X(x) = 0$ so that $k_*(X, Y) = 2$. The characteristic polynomial for $Q_x(X, Y)$ is $\lambda^3 - x_2^2\lambda^2$ so that the distinct eigenvalues are $\lambda_1(x) = x_2^2$ and $\lambda_2(x) = 0$ with multiplicities $m_1(x) = 1$ and $m_2(x) = 2$ for $x_2 \neq 0$. When $x_2 = 0$ $d(x) = 1$ and $l(x) = 1$ but when $x_2 \neq 0$ $l(x) = d(x) = 2$ so that $l = d = 2$ and $M(X, Y) = \{(x_1, x_2, x_3) \in R^3 | x_2 \neq 0\}$. In this example (3.12) becomes

$$e^{sQ_x(X,Y)} = e^{sx_2^2} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1/x_2^4 & 1/x_2^2 & 1 \end{bmatrix} + s \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ -1/x_2^2 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1/x_2^4 & -1/x_2^2 & 0 \end{bmatrix}$$

and (3.13) becomes

$$W^s_*(x) = e^{sx_2^2} W^s_{*\mathrm{MAX}}(x) + s W^s_{*\mathrm{MIN}}(x) + R^s(x)$$

where

$$W^s_{*\mathrm{MAX}}(x) = \left(\frac{-4}{x_2}, 0, x_2\, e^{x_1}\right), \qquad W^s_{*\mathrm{MIN}}(x) = (x_2^3 + 4x_2 - 2x_1x_2, x_2^2, 0)$$

and $R^s(x)=(x_1x_2-4/x_2, x_1, 0)$. Since $a_{\mathrm{MAX}}(x)=x_2^2>0$ on $M(X, Y)$ Theorem 3.5 (i) implies that $W^s_{*\mathrm{MAX}}\in\mathscr{F}^+$ on $M(X, Y)$. Since $a_{\mathrm{MIN}}(x)\equiv 0$ and $m_2(x)=m_d(x)=2>1$ Theorem 3.5 (iv) implies that $(-1)^{m_d+1}W^s_{*\mathrm{MIN}}=-W^s_{*\mathrm{MIN}}\in\mathscr{F}^+$. Since $\pm B\in\mathscr{F}^+$ one can use Lemma 2.1 to show

$$e^{-x_1}\left(\frac{4}{x_2^3}B(x)+W^s_{*\mathrm{MAX}}(x)\right)=(0, 0, x_2)\in\mathscr{F}^+;$$

$$\frac{1}{x_2^2}\left(\frac{x_2^2+4-2x_1}{x_2}B(x)-W^s_{*\mathrm{MIN}}(x)\right)=(0, -1, 0)\in\mathscr{F}^+.$$

Thus on $M(X, Y)=\{(x_1, x_2, x_3)\in R|x_2\neq 0\}$ the vector fields $(\pm 1, 0, 0)$, $(0, 0, x_2)$, $(0, -1, 0)$ and positive linear combinations of these vector fields are in $\mathscr{F}^+$. This increases our knowledge of the shape of the reachable set for this system. Without Theorem 3.5 it is only evident that $(\pm 1, 0, 0)$ are in $\mathscr{F}^+$.

Acknowledgment. The author wishes to thank the anonymous referee for helpful suggestions and in particular for Remark 3 following the proof of Theorem 3.3.

REFERENCES

[1] R. W. BROCKETT, *System theory on group manifolds and coset spaces*, this Journal, 10 (1972), pp. 265-284.
[2] ———, *Finite Dimensional Linear Systems*, John Wiley, New York, 1970.
[3] G. HAYNES AND H. HERMES, *Nonlinear controllability via Lie theory*, this Journal, 8 (1970), pp. 450-460.
[4] H. HERMES, *On local and global controllability*, this Journal, 12 (1974), pp. 252-262.
[5] R. HIRSCHORN, *Global controllability of nonlinear systems*, this Journal, 14 (1976), pp. 700-711.
[6] A. KRENER, *Nonlinear controllability and observability*, IEEE Trans. Automat. Control, AC-22 (1977), pp. 728-740.
[7] ———, *A generalization of Chow's theorem and the bang-bang theorem to nonlinear systems*, this Journal, 12 (1974), pp. 43-52.
[8] H. KUNITA, *On the controllability of nonlinear systems with applications to polynomial systems*, Appl. Math. Optim., 5 (1979), pp. 89-99.
[9] V. JURDJEVIC AND I. KUPKA, *Control systems subordinated to a group action: accessibility*, J. Differential Equations, 39 (1981), pp. 186-211.
[10] S. LEFSCHETZ, *Differential Equations: Geometric Theory*, Interscience, New York, 1957.
[11] T. NAGANO, *Linear differential systems with singularities and an application to transitive Lie algebras*, J. Math. Soc. Japan, 18 (1966).
[12] H. SUSSMANN AND V. JURDJEVIC, *Controllability of nonlinear systems*, J. Differential Equations, 12 (1972), pp. 95-116.
[13] ———, *Control systems on Lie groups*, J. Differential Equations, 12 (1972), pp. 313-329.
[14] F. WARNER, *Foundations of Differentiable Manifolds and Lie Group*, Scott, Foresman, Glenview, IL, 1970.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 2, March 1986

008

ON LIMIT CYCLES OF FEEDBACK SYSTEMS WHICH CONTAIN A HYSTERESIS NONLINEARITY*

R. K. MILLER†, A. N. MICHEL‡ AND G. S. KRENZ†

Abstract. In this paper, we concern ourselves with the stability of limit cycles in feedback systems which have hysteresis nonlinearities. Although the quasi-static analysis of limit cycles (Loeb criterion) predicts, in most cases correctly, the stability properties of limit cycles, it is well known that analyses which are based on the method of describing functions may lead to erroneous conclusions. In this paper, we show to what extent the describing function method can be given a rigorous mathematical basis. We show that for a specific example, the main result of this paper predicts correctly the stability of a limit cycle while the Loeb criterion yields an incorrect result. Also, we show that our analysis explains to a certain extent the presence of distortions in solutions of the class of feedback systems considered herein.

In arriving at the main result of this paper, use is made of several known facts for functional differential equations and of a result on integral manifolds.

Key words. functional differential equations, periodic solutions, feedback control systems, describing functions, Galerkin's method, hysteresis nonlinearity, delay differential equations, stability, orbital asymptotic stability, Loeb criterion, quasi-static stability analysis, limit cycles

1. Introduction.

The analysis of limit cycle behavior in nonlinear feedback systems is of great practical importance. Questions which arise in this connection concern the existence or nonexistence of limit cycles, estimates of amplitude and frequency of limit cycles, and the stability or instability of limit cycles. The sinusoidal describing function has been found to be a useful tool for answering questions of this type, particularly when the feedback system in question is endowed with one nonlinearity.

The purpose of the present paper is to show to what extent the method of describing functions can be given a rigorous mathematical basis. Our analysis here is an extension of earlier work [1], [30], [31] to the case where the nonlinearity in the feedback system is allowed to exhibit hysteresis. Hysteresis nonlinearities abound in applications. Some of the common types of such nonlinearities include iron core inductors, relays, gears with backlash, and the like. For an extensive discussion of the qualitative characteristics of such nonlinearities, refer, e.g., to the text by Gelb and Van der Velde [2].

It will be seen that the extension of the results in [1], [30], [31] to hysteresis nonlinearities involves considerable complications in the mathematical analysis. This is to be expected, since systems with memory are necessarily more complicated than those which are memoryless. The following example illustrates one such complication. Consider the feedback system

$$y^{(6)}+0.58y^{(5)}+11.6y^{(4)}+4.8648y^{(3)}+35.5y''+7.28y'+24.5y+5n(y)=0, \tag{1.1}$$

where $n(y)$ is the nonlinearity whose graph is given in Fig. 1. The sinusoidal-input describing function of this nonlinearity is given by

$$N(A)=2-\frac{2}{\pi}\left\{\sin^{-1}\left(\frac{2}{A}\right)+\left(\frac{2}{A}\right)\sqrt{1-4A^{-2}}\right\}-\frac{8i}{\pi A^2}.$$

* Received by the editors July 19, 1983, and in revised form August 13, 1984. This research was supported by the National Science Foundation under grant ECS-8100690 and by the Engineering Research Institute, Iowa State University.

† Mathematics Department, Iowa State University, Ames, Iowa 50011.

‡ Department of Electrical Engineering, University of Notre Dame, Notre Dame, Indiana 46556. Formerly at the Department of Electrical and Computer Engineering, Iowa State University, Ames, Iowa 50011.

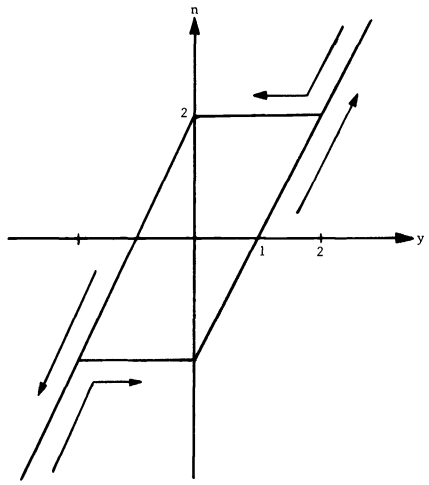


FIG. 1. *Nonlinear element for* (1.1).

A routine application of the describing function method [2] yields the prediction of a stable limit $y_0(t)$, which is approximately equal to

$$y_0(t) \cong A_0 \sin(\omega_0 t + \theta) \tag{1.2}$$

where $A_0 = 2.75$ and $\omega = 1.21$. (Without loss of generality, we take $\theta = 0$.) Numerical simulations of (1.1) verify that $y_0(t)$ does indeed exist and is approximately given by (1.2). These numerical solutions were also used to check whether or not (1.2) can be differentiated, i.e., whether or not it is true that, e.g.,

$$y_0'(t) \cong \omega_0 A_0 \cos \omega_0 t, \qquad y_0'' \cong -\omega_0^2 A_0 \sin \omega_0 t, \cdots. \tag{1.3}$$

It was found that (1.3) is not always valid. Specifically, Fig. 2 depicts the graphs of $y_0(t)$ and $y_0^{(5)}(t)$, computed numerically by simulating (1.1), as well as the predicted derivative $\omega_0^5 A_0 \cos \omega_0 t$. Clearly, (1.3) is not valid. This is curious, since (1.3) is found

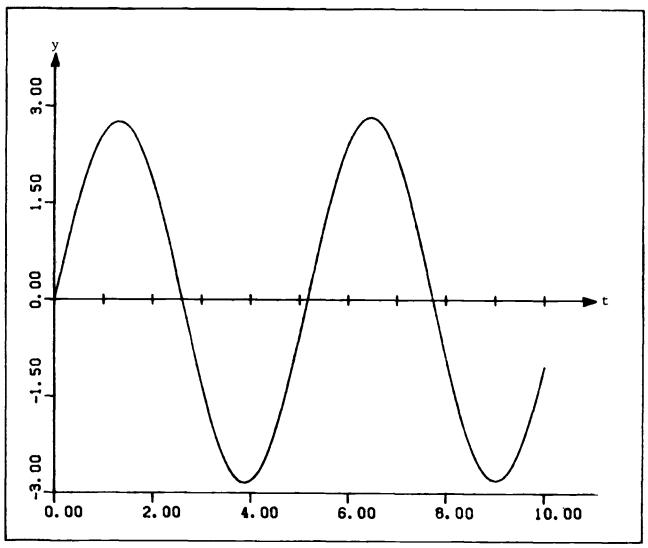


FIG. 2a. $y(t) \cong A_0 \sin \omega_0 t$.

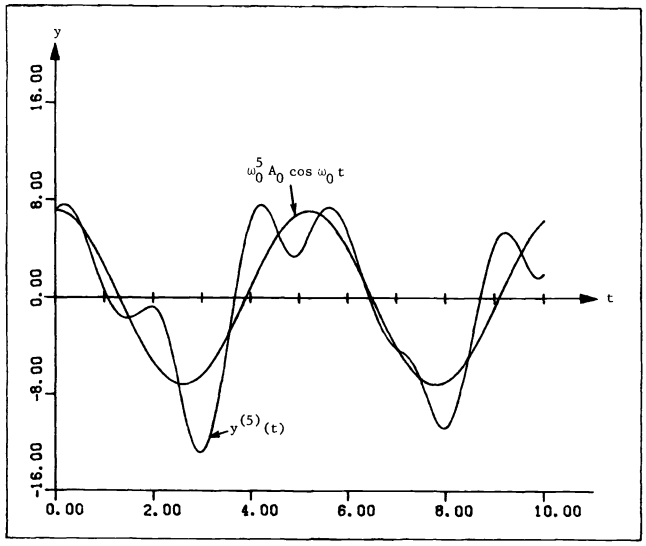


FIG. 2b. $y^{(5)}(t)$ *and* $\omega_0^5 A_0 \cos \omega_0 t$.

to be true to rather high accuracy for feedback systems where the nonlinearity $n(y)$ in (1.1) is memoryless.

Several third through sixth order equations with a variety of nonlinearities were simulated. The result was always the same: For stable limit cycles (1.3) accurately predicts the values of derivatives when $n(y)$ is memoryless and is inaccurate when $n(y)$ exhibits hysteresis. Our results will provide an explanation of this curious phenomenon. Furthermore, our results will also explain why a limit cycle can sometimes be unstable, even though the method of describing functions yields a prediction of stability.

Recently, the method of describing functions and other generalized Galerkin procedures have been studied extensively. Many of these results are concerned with *existence* of a limit cycle (e.g., Holtzman [7], Mees [8], Mees and Chua [10], Miller and Michel [11], Sandberg [12], [13], Cesari [14], Urabe [15] and Stokes [16]). *Nonexistence* results for limit cycles can be found in Mees and Bergen [9], Skar, Miller and Michel [17] and Urabe [15]. Also, *stability* results for limit cycles can be found in Mees and Chua [10], and in [1], [30], [31]. The method of analysis and results in [10], which are obtained by using Hopf bifurcation, are quite different from the results and methods of [1], [30], [31] and of this paper. Specifically, in the present paper we employ linearization about an approximate periodic solution, we make use of several appropriate transformations, and we utilize some results from the theory of integral manifolds (cf. [1], [4, Chap. 4], [5] or [6]). Our results contain a justification and correction of the popular quasi-static stability analysis (or Loeb criterion) developed by Cahen [18] (see also Loeb [19], Gibson [3], or Gelb and Van der Velde [2, pp. 120–125]). This quasi-static analysis is based on the sinusoidal-input describing function and has a well-known graphical interpretation. Furthermore, it has been established by a great deal of experience that the quasi-static stability analysis of limit cycles will usually yield correct predictions. However, we shall see that in certain situations this stability prediction can be incorrect. An example of such an incorrect prediction will be given at the end of this paper.

2. Main results. Consider the linear differential operators

$$Ly = \alpha_0 y^{(J)} + \alpha_1 y^{(J-1)} + \cdots + \alpha_{J-1} y' + \alpha_J y$$

and

$$Qy = \gamma_1 y^{(J-1)} + \gamma_2 y^{(J-2)} + \cdots + \gamma_{J-1} y' + \gamma_J y$$

where all α_j and γ_j are real numbers, $\alpha_0 = 1$ and at least one $\gamma_j \neq 0$. Let $n(y)$ denote a real valued, odd nonlinearity with hysteresis. Moreover, we assume that $n(y)$ is determined by an "upper" function, $n_U(y)$, which is used to define $n(y)$ as y decreases from $+\infty$ to $-\infty$, and a "lower" function, $n_L(y)$, which is used to define $n(y)$ as y increases from $-\infty$ to $+\infty$. For example, if $n(y)$ represents a relay with hysteresis as depicted in Fig. 3a, then $n_U(y)$ and $n_L(y)$ are the functions whose graphs are as shown in Figs. 3b and 3c, respectively.

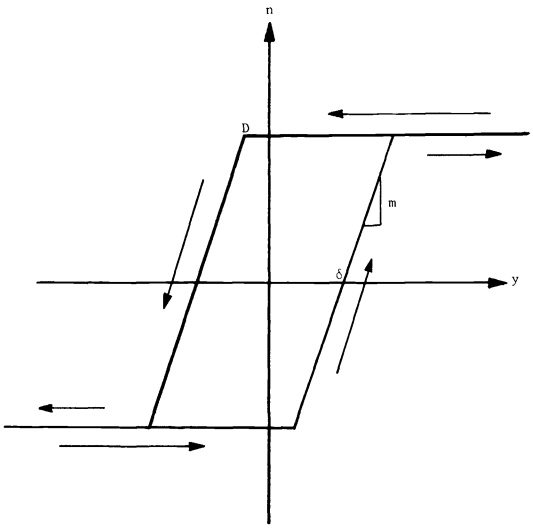


FIG. 3a. *The functional* $n(y)$.

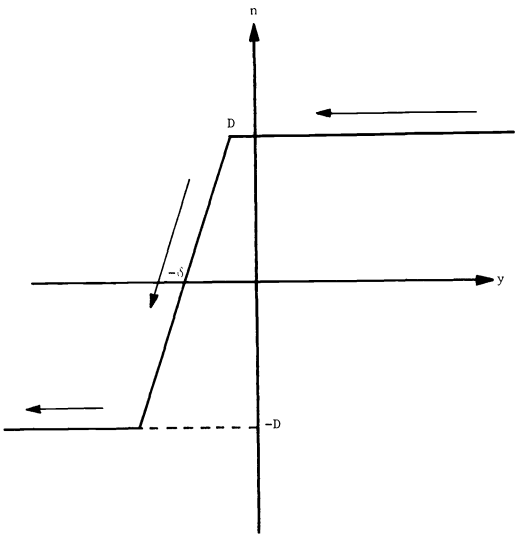


FIG. 3b. *The function* $n_U(y)$.

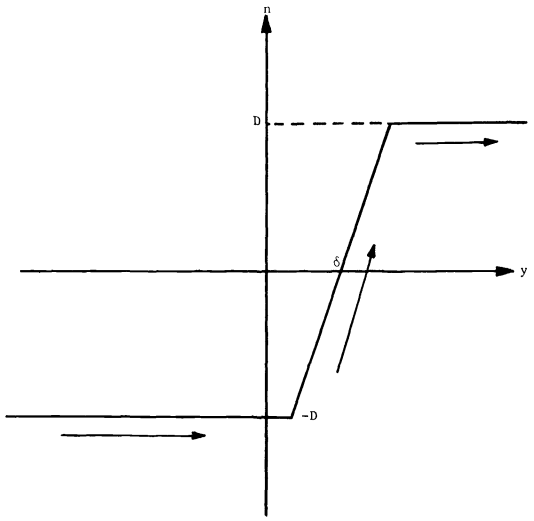


FIG. 3c. *The function* $n_L(y)$.

The *hysteresis region* of a hysteresis nonlinearity $n(y)$ is defined as the set $\{y \in R: n_U(y) \neq n_L(y)\}$. We assume that the hysteresis region is bounded. We also assume that n_U (resp., n_L) is a piecewise continuously differentiable function and that n_U'' (resp., n_L'') exists and is uniformly continuous on all intervals of the form (a, b) on which $n_U'(y)$ (resp., $n_L'(y)$) exists and is continuous.

Consider now the *feedback system*

$$Ly + n(Qy) = 0. \tag{2.1}$$

Suppose that the method of describing functions predicts a $2\pi/\omega_0$-periodic solution with amplitude A_0 (where A_0 and ω_0 are positive numbers). Thus, if

$$p(s) \triangleq \alpha_0 s^J + \alpha_1 s^{J-1} + \cdots + \alpha_{J-1} s + \alpha_J, \qquad q(s) \triangleq \gamma_1 s^{J-1} + \gamma_2 s^{J-2} + \cdots + \gamma_J$$

and if $N(A) = N_R(A) + iN_I(A)$ is the describing function for $n(y)$, then

$$p(i\omega_0) + q(i\omega_0) N(A_0 E) = 0, \qquad E = |q(i\omega_0)|. \tag{2.2}$$

We assume that $A_0 E$ is not in the hysteresis region of $n(y)$ nor is $y = A_0 E$ a value where $n'(y)$ does not exist. For example, if the graph of $n(y)$ is as shown in Fig. 4, then we assume that $A_0 E$ is not in the interval $[\delta_1, \delta_2]$ and $A_0 E$ is not equal to δ_3.

We expect that (2.1) will have a periodic solution $\phi(t)$ which is approximately equal to $A_0 \cos \omega_0 t$. In this case, ϕ will generate an integral manifold of (2.1), namely the set of points $\{(\phi(t+\theta), \phi'(t+\theta), \cdots, \phi^{(J-1)}(t+\theta)): -\infty < t, \theta < \infty\}$. We will show that a typical equation of the form (2.1) will have an associated integral manifold S whose stability type can be readily determined. The existence of S is independent of existence or nonexistence of a periodic solution ϕ. Hence, if (2.1) does not have a periodic solution $\phi(t) \cong A_0 \cos \omega_0 t$, it will still have the integral manifold S with the same type of behavior as the solutions of (2.1).

Our stability analysis depends on first finding a certain linear differential-difference equation of the form

$$\begin{aligned} &\alpha_0 y^{(J)}(t) + \alpha_1 y^{(J-1)}(t) + \cdots + \alpha_{J-1} y^{(1)}(t) + \alpha_J y(t) \\ &\quad + |N(A_0 E)|[\gamma_1 y^{(J-1)}(t-\alpha) + \gamma_2 y^{(J-2)}(t-\alpha) + \cdots + \gamma_J y(t-\alpha)] = 0. \end{aligned} \tag{2.3}$$

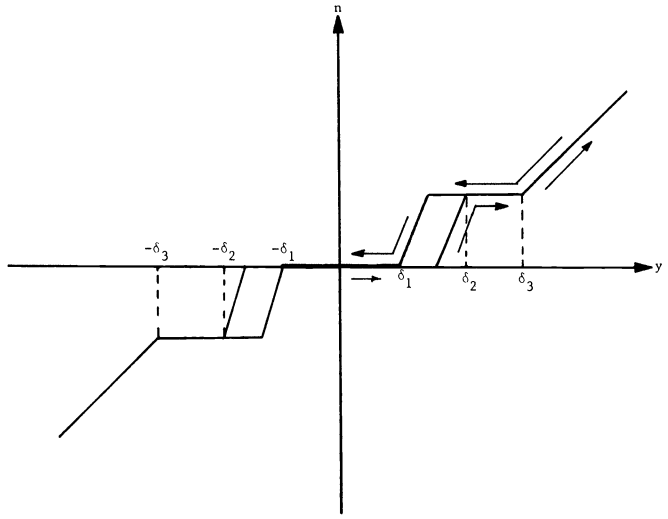


FIG. 4

The number $\alpha \geqq 0$ is computed by first writing $N(A_0E)$ in the polar form

$$N(A_0E) = |N(A_0E)|\, e^{-i\theta_0}$$

where $\theta_0 = \arg N(A_0E)$ satisfies $0 \leqq \theta_0 < 2\pi$. Define

$$\alpha \triangleq \frac{\theta_0}{\omega_0} \geqq 0.$$

The characteristic equation for (2.3) has the form $P(s) = 0$ where

$$P(s) \triangleq p(s) + |N(A_0E)|\, e^{-\alpha s} q(s) \tag{2.4}$$

and where $p(s)$ and $q(s)$ are defined as for (2.2). From (2.2) we know that $s_1 = i\omega_0$ and $s_2 = -i\omega_0$ are roots of $P(s)$. We label the remaining roots as $s_3, s_4, s_5, \cdots$.

We are now in a position to enumerate the assumptions for our main result:

(A-1) The numbers α_k and γ_k are real, $\alpha_0 = 1$, and not all γ_j are zero.

(A-2) $n(y)$ is a real functional determined by an "upper" function $n_U(y)$ and a "lower" function $n_L(y)$. The functions n_U and n_L are real valued, piecewise continuous, differentiable functions whose second derivatives are uniformly continuous on all intervals (a, b) where their first derivatives exist. Moreover, $n(y)$ is odd in the sense that

$$n_U(y) = -n_L(-y)$$

for all real numbers y. Also $n_U(y) = n_L(y)$ for $|y|$ sufficiently large.

(A-3) For some $A_0 > 0$ and $\omega_0 > 0$ equation (2.2) is true and $N(A_0E) \neq 0$. Moreover, EA_0 is not in the hysteresis region of n nor does EA_0 lie at a discontinuity of n'_U or n'_L.

(A-4) $s_1 = i\omega_0$ and $s_2 = -i\omega_0$ are simple roots of the function $P(s)$ defined by (2.4). The remaining roots $s_3, s_4, \cdots$ have nonzero real parts.

(A-5) For $k = 3, 4, \cdots$, s_k is a simple root of $P(s)$.

The main result of the present paper follows.

THEOREM 1. *Assume that* (A-1)-(A-5) *are true and define*

$$D \triangleq \operatorname{Re}\{N'(A_0E)q(i\omega_0)/P'(i\omega_0)\}.$$

If the numbers $|P'(s_k)^{-1}|$ *are sufficiently small for* $k = 1, 2, 3, \cdots$, *then* (2.1) *has, in a*

small neighborhood of the set

$$S_0 = \{(t, A_0 \cos(\omega_0 t + \theta), -A_0\omega_0 \sin(\omega_0 t + \theta), -A_0\omega_0^2 \cos(\omega_0 t + \theta), \cdots)^T: -\infty < t < \infty, 0 \leqq \theta \leqq 2\pi\} \subset R^{J+1}$$

a unique integral manifold S. This integral manifold is stable if $D > 0$ *and if* $\operatorname{Re} s_k < 0$ *for all* $k \geqq 3$. *This integral manifold is unstable if* $D < 0$ *or if* $\operatorname{Re} s_k > 0$ *for some* $k \geqq 3$.

Observe that the conclusion of the theorem is the existence of an integral manifold S which lies near the integral manifold generated by the approximate solution $A_0 \cos \omega_0 t$. The manifold S corresponds to a deformed torus fitted tightly around the predicted state-space limit cycle. While we cannot prove that solutions on S are periodic, these solutions will *appear* to the eye to be periodic and nearly sinusoidal with frequency near ω_0 and amplitude near A_0. If S is stable, then the apparent periodic solutions will appear to be orbitally asymptotically stable with asymptotic phase. If S is unstable, they will appear to be unstable in the sense of Lyapunov. This is the sense in which the theorem justifies the use of the describing function.

We remark that the hypothesis (A-5) concerning the simplicity of the roots s_k could be relaxed. In this case, the assumption that $|P'(s_k)^{-1}|$ is small must be replaced by other more complicated conditions. All of this can be accomplished only by making an already complicated analysis even more complicated. Hence, we shall retain (A-5). It will be shown later that if (A-5) is not true, then there is an $\varepsilon_0 > 0$ such that when α_J is replaced by $\alpha_J + \varepsilon$ and $0 < |\varepsilon| < \varepsilon_0$ hypothesis (A-5) is true. Hence, (A-5) does not appear to be a severe restriction.

Theorem 1 has the disadvantage that it is difficult to obtain the roots s_k for $k \geqq 3$, and hence, it is difficult to obtain the numbers $|P'(s_k)^{-1}|$. In addition, we cannot specify precisely how small the numbers $|P'(s_k)^{-1}|$ should be. However, this difficulty is partially offset by our knowledge that

$$\lim_{k \to \infty} |P'(s_k)^{-1}| = 0.$$

Moreover, for a typical equation (2.1) of degree $J \geqq 3$, these constants seem to be rather small. In all the examples which we considered, the size of the one number $|P'(i\omega_0)^{-1}|$ was a good indication of whether Theorem 1 would apply. We conjecture that in Theorem 1 only $|P'(i\omega_0)^{-1}|$ need be small while the smallness of $|P'(s_j)^{-1}|$ for $j \geqq 3$ is unnecessary. This conjecture is known to be true when $n(y)$ is a function rather than a functional which exhibits hysteresis (cf. [30], [31]).

The remainder of this paper is organized as follows. in § 3 we provide some necessary material concerning differential-delay equations; in § 4 we present some required background material on integral manifolds; in §§ 5 and 6 we prove the main result of this paper; in § 7 we discuss situations for which the quasi-static stability analysis yields incorrect results; in § 8 we present some specific examples; finally, the paper is concluded with several concluding remarks in § 9.

3. Some background material on differential-delay equations. Let R^J denote the real J-dimensional Euclidean space of column vectors with any convenient norm. Let R_J be the corresponding real J-dimensional space of row vectors. Consider a linear differential-difference equation of the form

$$z'(t) = B_0 z(t) + B_1 z(t - \alpha), \tag{3.1}$$

where $z(t)$ is an R^J-vector valued function, B_0 and B_1 are real and constant $J \times J$ matrices and α is a fixed nonnegative constant. An initial condition for (3.1) consists of specifying the value of $z(t)$ over the initial interval $-\alpha \leqq t \leqq 0$.

We shall view (3.1) as a functional differential equation with initial values in an appropriate Banach space. Our discussion follows Hale [22], [23]. Define

$$C[-\alpha, 0] = \{\phi: [-\alpha, 0] \to R^J;\ \phi \text{ is continuous}\}$$

and define a norm on $C[-\alpha, 0]$ by

$$\|\phi\| = \max\{|\phi(\theta)|, -\alpha \leqq \theta \leqq 0\}$$

where $|\cdot|$ denotes the norm on R^J. Similarly, define

$$C^*[0, \alpha] = \{\psi: [0, \alpha] \to R_J;\ \psi \text{ is continuous}\}$$

and define

$$\|\psi\| = \max\{|\psi(s)|: 0 \leqq s \leqq \alpha\}.$$

Given a continuous function $z: [-\alpha, a) \to R^J$ with $a > 0$, define an element z_t of $C[-\alpha, 0]$ by the formula

$$z_t(\theta) = z(t+\theta), \qquad -\alpha \leqq \theta \leqq 0,$$

for any $t \in [0, a)$. With the aid of this notation, (3.1) can now be written as

$$z'(t) = L_0 z_t, \tag{3.2}$$

where L_0 is the linear map from $C[-\alpha, 0]$ into R^J defined by

$$L_0\phi = B_0\phi(0) + B_1\phi(-\alpha).$$

This linear map can also be written as a Stieltjes integral, i.e.,

$$L_0\phi = \int_{-\alpha}^{0} d\eta(\theta)\phi(\theta) \tag{3.3}$$

where

$$\eta(\theta) = \begin{cases} 0 & \text{if } \theta = -\alpha, \\ B_1 & \text{if } -\alpha < \theta < 0, \\ B_0 + B_1 & \text{if } \theta = 0. \end{cases}$$

The *adjoint equation* corresponding to (3.1) is the equation

$$y'(t) = -y(t)B_0 - y(t+\alpha)B_1$$

where $y(t) \in R_J$ is a row vector valued function.

The initial value problem for (3.2) takes the form

$$z'(t) = L_0 z_t, \qquad z_0 = \phi \tag{3.4}$$

where the initial function ϕ is chosen from the set $C[-\alpha, 0]$. The initial value problem (3.4) has a unique solution defined for all $t \geqq 0$. We denote this solution by $z(t, \phi)$ if we think of z as R^J-valued and by $z_t(\phi)$ if we view z as assuming values in $C[-\alpha, 0]$.

We shall also require the notation

$$T(t)\phi \triangleq z_t(\phi)$$

for all $t \geqq 0$ and for all ϕ in $C[-\alpha, 0]$. For any fixed $t \geqq 0$, $T(t)$ is a bounded linear map from $C[-\alpha, 0]$ into itself with the usual operator norm

$$\|T(t)\| = \sup\{\|T(t)\phi\|;\ \phi \in C[-\alpha, 0] \text{ and } \|\phi\| = 1\}.$$

$T(t)$ determines a C_0-semigroup on $C[-\alpha, 0]$. The infinitesimal generator of this semigroup is the operator A_0 whose domain is

$$D(A_0) = \{\phi \in C[-\alpha, 0];\ \phi' \text{ exists},\ \phi' \in C[-\alpha, 0] \text{ and } \phi'(0) = L_0(\phi)\}.$$

For any ϕ in $D(A_0)$, A_0 is defined by the relation

$$(A_0\phi)(\theta) = \begin{cases} \phi'(\theta), & -\alpha \leqq \theta < 0, \\ L_0(\phi), & \theta = 0. \end{cases}$$

The set $D(A_0)$ is a dense subset of $C[-\alpha, 0]$. Refer, e.g., to Krein [25] or Hille and Phillips [26] for a discussion of C_0-semigroups.

The *characteristic equation* associated with (3.1) is

$$P(s) \triangleq \det (sI - B_0 - B_1 e^{-\alpha s}) = 0,$$

where I denotes the $J \times J$ identity matrix. The *eigenvalues* associated with (3.1) are the solutions of the characteristic equation. If $B_1 \neq 0$ and if $\alpha > 0$, then it is known that there is a countably infinite set

$$\sigma(L_0) = \{\lambda_n : n = 1, 2, 3, \cdots\}$$

of eigenvalues for (3.1). Each eigenvalue has finite multiplicity. These eigenvalues can be indexed so that $\operatorname{Re} \lambda_1 \geqq \operatorname{Re} \lambda_2 \geqq \operatorname{Re} \lambda_3 \geqq \cdots$ and we shall assume that this has in fact been done. It is known that $\operatorname{Re} \lambda_k$ will tend to $-\infty$ as $k \to \infty$.

Given any real number μ, define $\Lambda(\mu) = \{\lambda \in \sigma(L_0) : \operatorname{Re} \lambda \geqq -\mu\}$. There is a set $\{\phi_1, \phi_2, \cdots, \phi_V\}$ in $C[-\alpha, 0]$, called the *bases for the generalized eigenspace of* (3.1) for $\Lambda(\mu)$, and a set $\{\psi_1, \psi_2, \cdots, \psi_V\}$ in $C^*[0, \alpha]$, called the *bases for the generalized eigenspace of the adjoint of* (3.1) *for* $\Lambda(\mu)$, with the following special properties. Let Φ_μ be the $J \times V$ matrix whose columns are the vectors $\phi_1, \phi_2, \cdots, \phi_V$, i.e.,

$$\Phi_\mu = [\phi_1, \phi_2, \cdots, \phi_V].$$

Let Ψ_μ be that $V \times J$ matrix whose rows are the vectors $\psi_1, \psi_2, \cdots, \psi_V$, i.e.,

$$\Psi_\mu = \text{column}\,[\psi_1, \psi_2, \cdots, \psi_V].$$

Define a bilinear map by

$$(\psi, \phi) \triangleq \psi(0)\phi(0) - \int_{-\alpha}^{0} \left[\int_0^\theta \psi(\xi - \theta)\, d\eta(\theta) \right] \phi(\xi)\, d\xi,$$

for all $\psi \in C^*[0, \alpha]$ and for all $\phi \in C[-\alpha, 0]$. Then, Φ_μ and Ψ_μ can be chosen so that $(\psi_i, \phi_j) = \delta_{ij}$, the Kronecker delta. Define

$$P_\mu = \{\phi \in C[-\alpha, 0];\ \phi = \Phi_\mu b \text{ for some } b \in R^V\}$$

and

$$Q_\mu = \{\phi \in C[-\alpha, 0];\ (\Psi_\mu, \phi) = 0\}.$$

Then $C[-\alpha, 0]$ is the direct sum of P_μ and Q_μ. Indeed, if

$$\phi_P = \Phi_\mu(\Psi_\mu, \phi) \quad \text{and} \quad \phi_Q = \phi - \phi_P,$$

then $\phi_P \in P_\mu$, $\phi_Q \in Q_\mu$ and $\phi = \phi_P + \phi_Q$. The two subspaces P_μ and Q_μ are invariant under $T(t)$, i.e.,

$$T(t)P_\mu \subset P_\mu \quad \text{and} \quad T(t)Q_\mu \subset Q_\mu$$

for all $t \geqq 0$. There is a matrix B_μ (cf., e.g., [23, pp. 170-171] or [22]) such that

$$[T(t)\Phi_\mu b](\theta) = \Phi_\mu(0)\, e^{B_\mu(t+\theta)}b \tag{3.5}$$

for all $b \in R_V$, for all $t \geqq 0$, and for all $\theta \in [-\alpha, 0]$, while

$$\|T(t)\phi\| \leqq K_\mu\, e^{-\mu t}\|\phi\| \tag{3.6}$$

for all $\phi \in Q_\mu$ and for all $t \geqq 0$.

Consider also the nonhomogeneous initial value problem

$$z'(t) = L_0 z_t + F_0(t), \qquad z_0 = \phi \tag{3.7}$$

where $\phi \in C[-\alpha, 0]$, $F_0 : [0, \infty) \to R^J$ is continuous, and $L_0\phi$ is defined by (3.3). We again fix μ and obtain $\Lambda(\mu)$, B_μ, P_μ and Q_μ. Define

$$H_0(\theta) = \begin{cases} 0 & \text{if } -\infty \leqq \theta < 0, \\ I & \text{if } \theta = 0. \end{cases}$$

Then the solution $z_t(\phi)$ of (3.7) can be written in the form

$$z_t(\phi) = \Phi_\mu y(t) + z_{1t}$$

where

$$\begin{aligned} &y' = B_\mu y + \Psi_\mu(0)F_0(t), \qquad y(0) = (\Psi_\mu, \phi), \\ &z_{1t} = T(t)\phi_Q + \int_0^t T(t-s)[H_0 - \Phi_\mu\Psi_\mu(0)]F_0(s)\, ds. \end{aligned} \tag{3.8}$$

Here $\phi_Q = \phi - \Phi_\mu(\Psi_\mu, \phi) \in Q_\mu$ and $z_{1t} \in Q_\mu$ for all $t \geqq 0$.

All of the above results concerning (3.1) and (3.7) can be found in Hale [23, especially Chapter 7] or [22, especially pp. 94-130].

We now specialize to the equation

$$\alpha_0 y^{(J)}(t) + \alpha_1 y^{(J-1)}(t) + \cdots + \alpha_J y(t) + N[\gamma_1 y^{(J-1)}(t-\alpha) + \cdots + \gamma_J y(t-\alpha)] \tag{3.9}$$

where $\alpha_0 = 1$, $\alpha \geqq 0$, N, α_k and γ_i are real numbers, $N \neq 0$, and at least one $\gamma_i \neq 0$. As usual, if we set $z_1 = y$, $z_2 = y'$, etc., in (3.9), then that equation is equivalent to an equation of the form (3.1). Hence, the foregoing analysis concerning (3.1) will apply to (3.9). If we define

$$\Delta(\lambda) = \begin{bmatrix} \lambda & -1 & \cdots & 0 \\ 0 & \lambda & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ N\gamma_J e^{-\alpha\lambda} + \alpha_J & N\gamma_{J-1} e^{-\alpha\lambda} + \alpha_{J-1} & \cdots & N\gamma_1 e^{-\alpha\lambda} + \alpha_1 \end{bmatrix}$$

then the characteristic equation associated with (3.1) is seen to be

$$P(\lambda) = \det \Delta(\lambda) = \sum_{j=0}^{J-1} \lambda^j(\alpha_{J-j} + N\gamma_{J-j} e^{-\lambda\alpha}) + \lambda^J. \tag{3.10}$$

If A_0 is the infinitesimal generator of the corresponding semigroup and if λ is not an eigenvalue, then $(A_0 - \lambda I)\phi = \psi$ can be solved for $\phi = R(\lambda)\psi$ where $R(\lambda) = (A_0 - \lambda I)^{-1}$.

A lengthy but straightforward calculation shows that

$$\phi(\theta) = -\Delta(\lambda)^{-1}\left[\psi(0)\, e^{\lambda\theta} - e^{\lambda\theta}\int_{-\alpha}^{0} d\eta(u)\left(\int_0^u e^{\lambda(u-s)}\psi(s)\, ds\right)\right] + \int_0^\theta e^{\lambda(\theta-s)}\psi(s)\, ds \tag{3.11}$$

(see, e.g., Hale [22] or [23]).

Now fix a large positive μ. If $\phi \in Q_\mu \cap D(A_0)$, then for some $\sigma > 0$ it will be true that for $t > \alpha$,

$$T(t)\phi = -\frac{1}{2\pi i}\int_{\sigma-i\infty}^{\sigma+i\infty} e^{\lambda t} R(\lambda)\phi\, d\lambda, \tag{3.12}$$

where the integral is taken in the principal value sense (cf., e.g., [25, Thm. 13, p. 31] or [26, § 11.6]). If x^* is any linear functional on $C[-\alpha, 0]$, then by (3.12) it follows that

$$x^*(T(t)\phi) = -\frac{1}{2\pi i}\int_{\sigma-i\infty}^{\sigma+i\infty} e^{\lambda t} x^*(R(\lambda)\phi)\, d\lambda. \tag{3.13}$$

In particular, we pick x^* in (3.13) so that $x^*(\phi) = \phi(0)$. This yields

$$z(t, \phi) = [T(t)\phi](0) = -\frac{1}{2\pi i}\int_{\sigma-i\infty}^{\sigma+i\infty} e^{\lambda t}[R(\lambda)\phi](0)\, d\lambda.$$

The above relation, together with (3.11) results in the equation

$$z(t, \phi) = \frac{1}{2\pi i}\int_{\sigma-i\infty}^{\sigma+i\infty} e^{\lambda t}\Delta(\lambda)^{-1}\left[\phi(0) - \int_{-\alpha}^{0} d\eta(u)\left(\int_0^u e^{\lambda(u-s)}\phi(s)\, ds\right)\right] d\lambda. \tag{3.14}$$

The function F defined by

$$F(\lambda) \triangleq \Delta(\lambda)^{-1}\left[\phi(0) - \int_{-\alpha}^{0} d\eta(u)\left(\int_0^u e^{\lambda(u-s)}\phi(s)\, ds\right)\right]$$

is meromorphic and its poles can occur only at the eigenvalues $\lambda \in \sigma(L_0)$. Since $\phi \in Q_\mu$, it follows that $(\Psi_\mu, \phi) = 0$. Hence, the points $\lambda \in \Lambda(\mu)$ are removable singularities of F. The poles of $F(\lambda)$ must be in the set $\sigma(L_0) \cap \{\lambda : \operatorname{Re}\lambda < -\mu\}$. Thus, the contour of integration in (3.14) can be shifted to the left to the line $\operatorname{Re}\lambda = -\mu$, i.e.,

$$z(t, \phi) = \frac{1}{2\pi i}\int_{-\mu-i\infty}^{-\mu+i\infty} e^{\lambda t} F(\lambda)\, d\lambda. \tag{3.15}$$

Since (3.1) is now the Jth order equation (3.9), written in system form, the following result is true (cf. [24, § 12.8, especially p. 423, Thm. 12.15]).

THEOREM 2. (a) *The roots of* $P(\lambda) = \det\Delta(\lambda)$ *lie in a half plane* $\operatorname{Re}\lambda \leqq a_0$. *This set of roots is countable, the multiplicity of any root is at most J, and there can be at most J multiple roots. If* $\alpha > 0$, *then the (infinite) set of roots asymptotically approaches the curve*

$$|\lambda^M e^{\alpha\lambda}| = |N\gamma_M| \tag{3.16}$$

as $\operatorname{Re}\lambda \to -\infty$ *where M is the smallest integer j such that* $\gamma_j \neq 0$. *Moreover, there exists a number* $k > 0$ *such that the roots lie along the curve* (3.16) *an asymptotic distance of k apart.*

(b) *There exist* $M \geqq 1$ *and* $c_0 > 0$ *and for any* $c_1 > 0$ *there exists a number* $K > 0$ *such that if* $|\lambda| \geqq c_0$ *and if* λ *is at least a distance* c_1 *away from the set* $\sigma(L_0)$, *then* $|\Delta(\lambda)^{-1}| \leqq K|\lambda|^{-M}$.

Figure 5 shows the general shape of the curve (3.16). The x's indicate a typical set $\sigma(L_0)$. For $\alpha > 0$, Theorem 2 could also be applied to $\Delta'(\lambda)$. For $\Delta'(\lambda)$, the equation corresponding to (3.16) is $|\lambda^{M-1}\varepsilon^{\alpha\lambda}| = |\alpha N\gamma_M / M|$. Since the graph of this equation will not intersect the curve described by (3.16) for $|\lambda|$ sufficiently large, then there exists $c_1 > 0$ such that the roots of $P(\lambda)$ and $P'(\lambda)$ are separated by a distance of at least one when $|\lambda| \geqq c_1$ and such that $P(\lambda)$ has no roots on the circle $|\lambda| = c_1$. The number c_1 can be chosen independently of small changes in the coefficients $(\alpha_0, \alpha_1, \cdots, \alpha_J) \in R^{J+1}$, e.g. see [24]. Hence, any multiple roots of $P(\lambda)$ must lie in the disk $|\lambda| < c_1$. An elementary complex variable argument shows that if $P(\lambda)$ has multiple roots in the disk $|\lambda| < c_1$, then there is an $\delta_0 > 0$ such that if δ is a real number with $0 < |\delta| < \delta_0$ then $P(\lambda) + \delta$ has only simple roots. Hence, we can always arrange things so that assumption (A-5) is true by an arbitrarily small change in the parameter α_J of the differential equation (2.1).

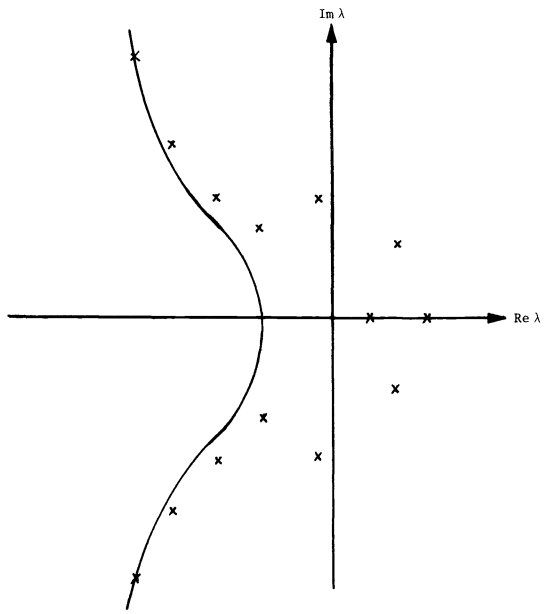


FIG. 5. *Typical curve* $|\lambda^M e^{\alpha\lambda}| = |N\gamma_M|$ *and set* $\sigma(L_0)$.

Let c_0 and k be the numbers defined in Theorem 2. Choose ε in the interval $0 < \varepsilon < k$. Since the choice of μ is at our disposal, we may pick $\mu > c_0 + k$, and so large, that all points of the set $\sigma(L_0) \cap \{\lambda : \operatorname{Re}\lambda < -\mu\}$ lie within $\varepsilon/10$ of the curve (3.16) and are located at a distance greater than $0.8k$ from each other. Define $c_2 \triangleq \inf\{\operatorname{Im}\lambda - \varepsilon/10 : \lambda \in \sigma(L_0) \text{ and } \operatorname{Re}\lambda < -\mu\}$. Define

$$S_0(\varepsilon) = \{\lambda : \lambda \text{ lies within } \varepsilon/10 \text{ of the set } \sigma(L_0)\}.$$

Let K be the number given in Theorem 2, corresponding to $c_1 = \varepsilon/10$. Define

$$\Gamma_1 = \{-\mu + i\tau;\ 0 \leqq \tau \leqq c_2\},$$

$$\Gamma_2 = \{\sigma + ic_2;\ -\mu \leqq \sigma \leqq -\mu + \varepsilon\},$$

and

$$\Gamma_3 = \{-\mu + \varepsilon + i\tau;\ c_2 \leqq \tau \leqq \infty\}.$$

Let Γ denote the upward directed curve made up of $\Gamma_1 \cup \Gamma_2 \cup \Gamma_3$ and its symmetric image below the real axis (see Fig. 6). By its construction, the contour Γ misses the set $S_0(\varepsilon)$. Hence, $|\Delta(\lambda)^{-1}| \leqq K|\lambda|^{-1}$ on Γ. The contour in (3.15) can be shifted so that

$$z(t, \phi) = \frac{1}{2\pi i} \int_{\Gamma} e^{\lambda t} F(\lambda)\, d\lambda.$$

Now

$$\begin{aligned} \frac{1}{2\pi i} \int_{\Gamma_1} e^{\lambda t} F(\lambda)\, d\lambda &= \frac{1}{2\pi} \int_0^{c_2} e^{(-\mu + i\tau)t} F(-\mu + i\tau)\, d\tau \\ &= \frac{e^{-\mu t}}{2\pi} \int_0^{c_2} e^{-i\tau t} F(-\mu + i\tau)\, d\tau. \end{aligned} \tag{3.17}$$

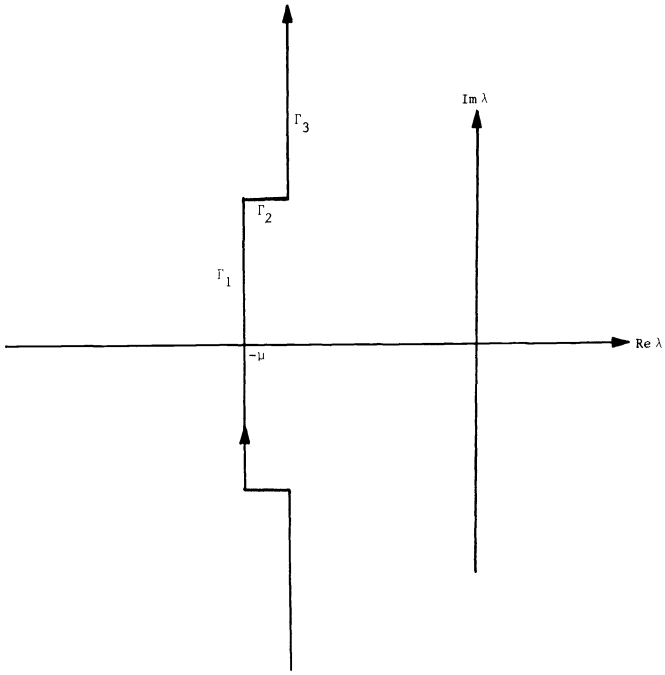


FIG. 6. *The contour* Γ.

The last integral above on the right is the Fourier transform of a function whose L^2-norm can be estimated by

$$\int_0^{c_2} |F(-\mu + i\tau)|^2\, d\tau \leqq \int_{-\infty}^{\infty} K^2 |-\mu + i\tau|^{-2}\, d\tau \triangleq K_1(\mu)^2.$$

By the Parseval relation, the last integral on the right in (3.17) defines a function of t whose L_2-norm over $(0, \infty)$ is at most $\sqrt{2\pi}\, K_1(\mu)$. Similarly, the function f defined by the integral

$$\frac{1}{2\pi i} \int_{\Gamma_3} e^{\lambda t} F(\lambda)\, d\lambda = e^{-(\mu+\varepsilon)t} f(t)$$

has L^2-norm over $(0,\infty)$ which is at most $\sqrt{2\pi}\,K_1(\mu)$. The corresponding contour integrals below the real axis satisfy identical estimates.

Now consider the integral

$$\frac{1}{2\pi i}\int_{\Gamma_2} e^{\lambda t}F(\lambda)\,d\lambda = \frac{1}{2\pi i}\int_0^{\varepsilon} e^{(-\mu+\sigma+ic_2)t}F(-\mu+\sigma+ic_2)\,d\sigma.$$

The magnitude of this integral is at most

$$\frac{e^{-\mu t}}{2\pi}\int_0^{\varepsilon} e^{\sigma t}K|-\mu+\sigma+ic_2|^{-1}\,d\sigma.$$

The corresponding integral below the real axis satisfies an identical estimate.

From these estimates we see that if $\|\phi\| \leqq 1$, then there is a constant $K_2(\mu)$ such that $K_2(\mu)\to 0$ as $\mu\to\infty$ and

$$z(t,\phi) = [g_\mu(t)+f_\mu(t)]\,e^{(-\mu+\varepsilon)t} \tag{3.18}$$

where

$$\int_0^{\infty} |f_\mu(t)|^2\,dt \leqq K_2(\mu)^2, \qquad |g_\mu(t)| \leqq K_2(\mu).$$

Since $z=(y, y', y'', \cdots, y^{(J-1)})^T$ where y solves (3.9), then we see that $y^{(J)}(t)$ satisfies a similar estimate. Hence, $z'(t,\phi)$ also satisfies a similar estimate, i.e.,

$$z'(t,\phi) = [g(t)+f(t)]\,e^{(-\mu+\varepsilon)t}, \tag{3.19}$$

with $|g(t)| \leqq K_2(\mu)$ and $\|f\|_{L^2} \leqq K_2(\mu)$. From (3.18) and (3.19) we see that

$$z(t,\phi) = -\int_t^{\infty} [g(s)+f(s)]\,e^{(-\mu+\varepsilon)s}\,ds,$$

or

$$z(t,\phi) = e^{(-\mu+\varepsilon)t}\int_t^{\infty} [g(s)+f(s)]\,e^{(-\mu+\varepsilon)(s-t)}\,ds. \tag{3.20}$$

We can obtain an estimate for the integral (3.20) as follows:

$$\begin{aligned}\int_t^{\infty} [g(s)+f(s)]\,e^{(-\mu+\varepsilon)(s-t)}\,ds &\leqq \int_t^{\infty} [K_2(\mu)+|f(s)|]\,e^{(-\mu+\varepsilon)(s-t)}\,ds\\ &= K_2(\mu)(\mu-\varepsilon)^{-1} - \int_t^{\infty} |f(s)|\,e^{(-\mu+\varepsilon)(s-t)}\,ds\\ &\leqq K_2(\mu)(\mu-\varepsilon)^{-1} + \left(\int_0^{\infty} e^{2(-\mu+\varepsilon)s}\,ds\right)^{1/2}\|f\|_{L^2}\\ &\leqq K_2(\mu)(\mu-\varepsilon)^{-1} + [2(\mu-\varepsilon)]^{-1/2}K_2(\mu) \triangleq K(\mu).\end{aligned}$$

Since $K(\mu)\to 0$ as $\mu\to\infty$, we can prove the following result.

THEOREM 3. *If y solves* (3.9), *$z=(y, y', \cdots, y^{(J-1)})^T$, and ε is any small positive constant, then for all sufficiently large μ there is a constant $K(\mu)$ such that $K(\mu)\to 0$ as $\mu\to\infty$ and*

$$\|T(t)\phi\| \leqq K(\mu)\,e^{(-\mu+\varepsilon)t}\|\phi\|$$

for all $t>\alpha$ and for all $\phi\in Q_\mu$.

Proof. We have proved the result when $\phi \in Q_\mu \cap D(A_0)$ and $\alpha > 0$. Since this set is dense in Q_μ, the theorem follows by continuity. The case $\alpha = 0$ is trivial. □

Remark 1. We note that $T(t)\phi_0$ makes sense for the piecewise continuous function $\phi_0 \triangleq H_0 - \Phi_\phi \Psi_\mu(0)$ given in (3.8). Since $T(t)\phi_0$ is bounded for $0 \leqq t \leqq \alpha$ and since $T(t)\phi_0 \in Q_\mu$ for $t \geqq \alpha$, then $T(t)\phi_0 = T(t-\alpha)T(\alpha)\phi_0$ for $t > \alpha$ and we have

$$\begin{aligned}\|T(t)\phi_0\| &= \|T(t-\alpha)T(\alpha)\phi_0\| \leqq K(\mu)\, e^{-\mu(t-\alpha)} \|T(\alpha)\phi_0\| \\ &= (K(\mu)\|T(\alpha)\phi_0\|\, e^{\mu\alpha})\, e^{-\mu t}.\end{aligned}$$

This inequality will be needed later.

4. Some background material on integral manifolds. Consider a real-valued coupled system of integrodifferential equations of the form

$$\begin{aligned}\theta' &= \varepsilon\theta(t, \theta, r, v_1, v_{2t}, \varepsilon), \\ r' &= \varepsilon\beta r + \varepsilon R(t, \theta, r, v_1, v_{2t}, \varepsilon), \\ v_1' &= Cv_1 + \varepsilon V_1(t, \theta, r, v_1, v_{2t}, \varepsilon), \\ v_{2t}'(\tau) &= F_1(t)(\tau) + \varepsilon \int_{t_0}^{t} F_2(t-s)(\tau)\, V_2(s, \theta(s), r(s), v_1(s), v_{2s}, \varepsilon)\, ds\end{aligned} \tag{4.1}$$

defined on a set

$$\begin{aligned}D_1 = \{(t, \theta, r, v_1, \phi, \varepsilon)&: t, \theta, r \in R^1, 0 < \varepsilon \leqq \varepsilon_0, v_1 \in R^m, \\ &\phi \in C_\mu[-\alpha, 0], |r| \leqq \eta_1, |v_1| \leqq \eta_2, \|\phi\| \leqq \eta_3\}.\end{aligned}$$

Here $C_\mu[-\alpha, 0]$ is a closed subspace of $C[-\alpha, 0]$. We assume $\beta \neq 0$, C is a constant, noncritical $m \times m$ matrix, θ and R are real-valued, V_2 is R^J-valued, and V_1 is R^m-valued. Assume that θ, R, V_1 and V_2 are $2\pi/\omega_0$-periodic in t for some $\omega_0 > 0$, 2π-periodic in θ, and continuous on D_1. Assume that F_1 maps $[0, \infty)$ into $C_\mu[-\alpha, 0]$, F_1 is continuous and $\|F_1(t)\| \leqq K_1\, e^{-\mu t}$ for some $K_1 > 0$ and $\mu > 0$. Assume that $F_2(t)(\theta)$ is a piecewise continuous matrix-valued map such that each column of F_2 is a continuous map from $[\alpha, \infty)$ to $C_\mu[-\alpha, 0]$ and $\|F_2(t)\| \leqq K_1\, e^{-\mu t}$ for all $t \geqq \alpha$.

We assume that V_i is a Lipschitz continuous function of (θ, r, v, ϕ) on D_1 for $i = 1, 2$. Also, there is a continuous and nonnegative function $M_1(\theta, a_1, a_2, a_3, \varepsilon)$ such that M_1 is 2π-periodic in θ, piecewise continuous in θ, and such that

$$\int_0^{2\pi} M_1(\theta, a_1, a_2, a_3, \varepsilon)\, d\theta \to 0 \tag{4.2}$$

as $\varepsilon \to 0$, $a_1 \to 0$, $a_2 \to 0$ and $a_3 \to 0$. Moreover, for all points $(t, \theta_i, r_j, v_k, \phi_m, \varepsilon)$ in D_1 with $|r_j| \leqq a_1$, $|v_k| \leqq a_2$ and $\|\phi_m\| \leqq a_3$, we have

$$|R(t, \theta_1, r_0, v_0, \phi_0, \varepsilon) - R(t, \theta_2, r_0, v_0, \phi_0, \varepsilon)| \leqq M_1(\omega_0 t - \theta_1, a_1, a_2, a_3, \varepsilon)|\theta_1 - \theta_2|,$$

$$|R(t, \theta_0, r_1, v_0, \phi_0, \varepsilon) - R(t, \theta_0, r_2, v_0, \phi_0, \varepsilon)| \leqq M_1(\omega_0 t - \theta_0, a_1, a_2, a_3, \varepsilon)|r_1 - r_2|,$$

$$|R(t, \theta_0, r_0, v_1, \phi_0, \varepsilon) - R(t, \theta_0, r_0, v_2, \phi_0, \varepsilon)| \leqq M_1(\omega_0 t - \theta_0, a_1, a_2, a_3, \varepsilon)|v_1 - v_2|,$$

and

$$|R(t, \theta_0, r_0, v_0, \phi_1, \varepsilon) - R(t, \theta_0, r_0, v_0, \phi_2, \varepsilon)| \leqq M_1(\omega_0 t - \theta_0, a_1, a_2, a_3, \varepsilon)\|\phi_1 - \phi_2\|.$$

We assume that θ satisfies these same four Lipschitz conditions using the same Lipschitz function M_1.

We shall require the notion of an integral manifold of (4.1).

DEFINITION 4. A surface S_ε in $(t, \theta, r, v, \phi, \varepsilon)$-space is an integral manifold of system (4.1) for a fixed ε if given any point $(t_0, \theta_0, r_0, v_0, \phi_0, \varepsilon)$ in $S_\varepsilon \cap D_1$, the solution of (4.1) which satisfies $\theta(t_0) = \theta_0$, $r(t_0) = r_0$, $v_1(t_0) = v_0$ and $v_{2t_0} = \phi$ is defined for all times t and $(t, \theta(t), r(t), v_1(t), v_{2t}, \varepsilon) \in D \cap S_\varepsilon$ for all $t \in R$.

For example, in (4.1) if $\theta \equiv R = V_1 \equiv V_2 \equiv F_1 \equiv 0$, then there is an integral manifold, namely

$$S = \{(t, \theta, 0, 0, 0, \varepsilon);\ 0 \leqq t < \infty,\ -\infty < \theta < \infty\}.$$

This integral manifold is stable if $\beta < 0$ and C is a stable matrix, and it is unstable if $\beta > 0$ or if C has an eigenvalue with positive real part. Intuitively, since θ, R, V_1 and V_2 are small, then one might expect that for ε small there will exist an integral manifold S_ε of (4.1) which is close to S and which has the same stability properties as S. This conjecture is correct. Indeed, the following result is true.

THEOREM 5. *Suppose that θ, R, V_1 and V_2 have the periodicity and continuity properties enumerated above, that M_1 satisfies* (4.2), *that $\beta \neq 0$, and that C is noncritical. There is an $\eta > 0$ such that if $K_1 < \eta$ then in the region*

$$D(\eta) = \{(t, \theta, r, v_1, \phi, \varepsilon) D_1 \colon |r| \leqq \eta, |v_1| \leqq \eta, \|\phi\| \leqq \eta \text{ and } 0 < \varepsilon \leqq \eta\}$$

there is a unique integral manifold S_ε of (4.1). *This integral manifold will have the same stability properties as S.*

One can show that S_ε is determined by three functions f_1, f_2 and f_3 with

$$S_\varepsilon = \{(t, \theta, r, v, \phi, \varepsilon) \in D(\eta);\ r = f_1(t, \theta, \varepsilon),\ v = f_2(t, \theta, \varepsilon) \text{ and } \phi = f_3(t, \theta, \varepsilon)\}.$$

The functions f_i, $i = 1, 2, 3$, are continuous in (t, θ, ε), $2\pi/\omega_0$-periodic in t, 2π-periodic in θ, and $f_i(t, 0, 0) \equiv 0$. The proof of Theorem 5 is long and detailed but involves standard techniques. It follows the same outline as used by Hale [4, Chap. 7].

5. Transformation of the feedback system. In this section, we shall prove our main result, Theorem 1. This will be accomplished by using several transformations to reduce (2.1) to an integral manifold problem. We assume that the hypotheses of § 2 are true and we employ the notation established in that section. Assume for now that $\alpha > 0$. The case $\alpha = 0$ will be handled later. Since the limit cycle predicted by the method of describing functions is $A_0 \cos \omega_0 t$, where $A_0 E$ is not in a hysteresis region of $n(y)$, then by restricting our attention to functions $y(t)$ for which

$$|y(t+\theta) - A_0 \cos \omega_0 t| + |y'(t+\theta) + A_0 \omega_0 \sin \omega_0 t|$$

is small over $0 \leqq t \leqq 2\pi/\omega_0$ for some phase shift θ (which may depend on y), there will be no problem in defining $n(y(t))$ given $y(t+s)$ over $-\alpha \leqq s \leqq 0$. Hence, we can assume that $n(y)$ is a functional with domain in $C[-\alpha, 0]$.

Given the values A_0, ω_0 and α such that $E = |q(i\omega_0)|$, $p(i\omega_0) + N(A_0 E) q(i\omega_0) = 0$ while $N(A_0 E) = |N(A_0 E)|\, e^{-i\alpha\omega_0}$, we define an operator L_1 and a nonlinearity $n_1(y)$ by the relations

$$L_1 y(t) \triangleq Ly(t) + NQy(t-\alpha), \qquad N = |N(A_0 E)|$$

and

$$n_1(y_t) \triangleq NQy(t-\alpha) - n(Qy).$$

Thus (2.1) can be written as

$$L_1 y(t) = n_1(y_t). \tag{5.1}$$

As usual, we let $z=(y, y', y'', \cdots, y^{(J-1)})^T$ and we represent (5.1) equivalently as

$$z'(t)=L_0 z_t+e_J n_1(y_t), \tag{5.2}$$

where

$$e_J=(0,0,\cdots,0,1)^T\in R^J,$$

$$L_0\phi \triangleq \int_{-\alpha}^{0} d\eta(\theta)\phi(\theta)\triangleq B_0\phi(0)+B_1\phi(-\alpha),$$

and

$$B_0=\begin{bmatrix} 0 & 1 & 0 & \cdots & 0\\ 0 & 0 & 1 & \cdots & 0\\ \cdots & \cdots & \cdots & \cdots & \cdots\\ 0 & 0 & 0 & \cdots & 1\\ -\alpha_J & -\alpha_{J-1} & -\alpha_{J-2} & \cdots & -\alpha_1 \end{bmatrix}$$

$$B_1=\begin{bmatrix} 0 & 0 & \cdots & 0\\ \vdots & \vdots & & \vdots\\ 0 & 0 & \cdots & 0\\ -N\gamma_J & -N\gamma_{J-1} & \cdots & -N\gamma_1 \end{bmatrix}.$$

Fix μ, a large positive number and let $\Lambda(\mu)$ be the set of eigenvalues s_k of $z'=L_0 z$ such that $\operatorname{Re} s_k \geqq -\mu$. We enumerate $\Lambda(\mu)=\{s_1, s_2, \cdots, s_{N(\mu)}\}$ in such a way that $s_1=i\omega_0$, $s_2=-i\omega_0$ and in such a way that complex conjugate pairs are ordered as s_k, s_{k+1} where $\operatorname{Im} s_k>0$ and $\operatorname{Im} s_{k+1}=\operatorname{Im}\bar{s}_k<0$. We wish to think of $F_0(t)=e_J n_1(y(t))$ and then write a decomposition of the form (3.8). Hence, we shall need to compute the generalized eigenspaces of the linear problem $z'=L_0 z_t$ and of its adjoint. A sample calculation of the type which we must perform can be found in [23, pp. 182-184] or [22, pp. 116-119].

Let $\xi(s)\in R^J$ be the vector defined by $\xi(s)=(1, s, s^2, \cdots, s^{J-1})^T$. Define $\eta(s)=(s^{J-1}+[\alpha_1+\gamma_1 N e^{-\alpha s}]s^{J-2}+\cdots+[\alpha_{J-1}+\gamma_{J-1}N e^{-\alpha s}], \cdots, s+[\alpha_1+N\gamma_1 e^{-\alpha s}], 1)$. Let s_k be a real root of $\Delta(s)$. It is easy to check that $\phi_k(\theta)\triangleq e^{s_k\theta}\xi(s_k)$ is the corresponding eigenfunction and that $\hat{\psi}_k(\theta)\triangleq e^{-s_k\theta}\eta(s_k)$ is an eigenfunction of the adjoint. Then $\hat{\psi}(0)\phi(0)=\eta(s_k)\xi(s_k)=Js^{J-1}+(J-1)s^{J-2}[\alpha_1+\gamma_1 N e^{-\alpha s}]+\cdots+[\alpha_{J-1}+N\gamma_{J-1}]e^{-\alpha s}$ and

$$\begin{aligned}\int_{-\alpha}^{0}\int_{0}^{\theta} &\eta(s_k)\, e^{-s_k(u-\theta)}\, d\eta(\theta)\xi(s_k)\, e^{s_k u}\, du\\ &=\eta(s_k)\left[\int_{-\alpha}^{0}\int_{0}^{\theta} e^{-s_k\theta}\, d\eta(\theta)\, du\right]\xi(s_k)=\eta(s_k)\int_{-\alpha}^{0}\theta\, e^{-s_k\theta}\, d\eta(\theta)\xi(s_k)\\ &=\alpha|N(A_0)|q(s_k)\, e^{-\alpha s_k}.\end{aligned}$$

Hence $(\hat{\psi}_k, \phi_k)=P'(s_k)\neq 0$. Define $\psi_k(\theta)=\hat{\psi}_k(\theta)/P'(s_k)$ so that $\psi_k(0)e_J=P'(s_k)^{-1}$. Then $(\psi_k, \phi_k)=1$ as required. Let $z_t(\phi)$ be the solution of (5.2) which satisfies the initial condition $z_0(\phi)=\phi$ and let $\psi_k^*=\psi_k$. Then $y_k(t)\triangleq(\psi_k^*, z_t(\phi))$ solves

$$y_k'=s_k y_k+\frac{P'(s_k)}{|P'(s_k)|^2}n_1(y_t), \qquad y_k(0)=(\psi_k^*, \phi). \tag{5.3}$$

Now consider an eigenvalue s_k such that $\operatorname{Im} s_k > 0$. From the ordering of the eigenvalues which we have employed, we know that $s_{k+1} = \bar{s}_k$. Then $u_k(\theta) \triangleq e^{s_k\theta}\xi(s_k)$ is an eigenfunction corresponding to the eigenvalue s_k and $v_k(\theta) \triangleq e^{-s_k\theta}\eta(s_k)$ is an eigenfunction for the adjoint problem. Hence

$$\phi_{1k}(\theta) \triangleq \operatorname{Re} u_k(\theta), \qquad \phi_{2k}(\theta) \triangleq \operatorname{Im} u_k(\theta)$$

generates the generalized eigenspace for the eigenvalues $\{s_k, s_{k+1}\}$ while

$$\hat{\psi}_{1k}(\theta) = \operatorname{Re} v_k(\theta), \qquad \hat{\psi}_{2k}(\theta) = \operatorname{Im} v_k(\theta)$$

will generate the generalized eigenspace for the adjoint. Since $s_k \neq s_{k+1}$, then from general considerations in Hale [22] or [23] we know that $(\bar{v}_k, u_k) = (v_k, \bar{u}_k) = 0$. By a computation similar to the case where s_k is real, we can compute

$$(v_k, u_k) = P'(s_k), \qquad (\bar{v}_k, \bar{u}_k) = P'(\bar{s}_k) = \overline{P'(s_k)}.$$

Since $\phi_{1k} = (u_k + \bar{u}_k)/2$ and ϕ_{2k}, $\hat{\psi}_{1k}$ and $\hat{\psi}_{2k}$ can be similarly expressed, then

$$D_k = \begin{bmatrix} (\hat{\psi}_{1k}, \phi_{1k}) & (\hat{\psi}_{1k}, \phi_{2k}) \\ (\hat{\psi}_{2k}, \phi_{1k}) & (\hat{\psi}_{2k}, \phi_{2k}) \end{bmatrix}$$

is easily evaluated. Indeed,

$$D_k = \frac{1}{2}\begin{bmatrix} \operatorname{Re} P'(s_k) & \operatorname{Im} P'(s_k) \\ \operatorname{Im} P'(s_k) & -\operatorname{Re} P'(s_k) \end{bmatrix}.$$

Hence, the pair ψ_{1k} and ψ_{2k} defined by

$$\begin{bmatrix} \psi_{1k} \\ \psi_{2k} \end{bmatrix} \triangleq D_k^{-1}\begin{bmatrix} \hat{\psi}_{1k} \\ \hat{\psi}_{2k} \end{bmatrix} = 2\begin{bmatrix} \operatorname{Re}[P'(\bar{s}_k)^{-1}v_k] \\ -\operatorname{Im}[P'(\bar{s}_k)^{-1}v_k] \end{bmatrix}$$

will generate the generalized eigenspace for the adjoint and in addition will satisfy the relations $(\psi_{nk}, \phi_{mk}) = \delta_{mn}$. Moreover,

$$\begin{bmatrix} \psi_{1k}(0) \\ \psi_{2k}(0) \end{bmatrix} e_J = \begin{bmatrix} \psi_{1k}(0)e_J \\ \psi_{2k}(0)e_J \end{bmatrix} = 2\begin{bmatrix} \operatorname{Re} P'(\bar{s}_k)^{-1} \\ -\operatorname{Im} P'(\bar{s}_k)^{-1} \end{bmatrix}.$$

Let $z_t(\phi)$ be the solution of (5.2) which satisfies the initial condition $z_0(\phi) = \phi$. If we define $w_{1k}(t) = (\psi_{1k}, z_t(\phi))$ and $w_{2k}(t) = (\psi_{2k}, z_t(\phi))$ then

$$\begin{bmatrix} w_{1k}' \\ w_{2k}' \end{bmatrix} = \begin{bmatrix} \operatorname{Re} s_k & -\operatorname{Im} s_k \\ \operatorname{Im} s_k & \operatorname{Re} s_k \end{bmatrix}\begin{bmatrix} w_{1k} \\ w_{2k} \end{bmatrix} + \begin{bmatrix} \operatorname{Re} P'(\bar{s}_k)^{-1} \\ -\operatorname{Im} P'(\bar{s}_k)^{-1} \end{bmatrix}[2n_1(y_t)]. \tag{5.4}$$

Equivalently, we can define $y_k = (w_{1k} + iw_{2k})/2$, $y_{k+1} = \bar{y}_k$, $\psi_k^* = (\psi_{1k} + i\psi_{2k})/2$ and $\psi_{k+1}^* = \bar{\psi}_k^*$. Then

$$y_k' = s_k y_k + \frac{P'(s_k)}{|P'(s_k)|^2} n_1(y_t), \qquad y_k(0) = (\psi_k^*, \phi)$$

and

$$y_{k+1}' = \bar{s}_k y_{k+1} + \frac{P'(\bar{s}_k)}{|P'(\bar{s}_k)|^2} n_1(y_t), \qquad y_{k+1}(0) = (\psi_{k+1}^*, \phi).$$

Since $s_{k+1} = \bar{s}_k$, both of these equations have the form (5.3).

Given the set $\Lambda(\mu) = \{s_1, s_2, s_3, \cdots, s_{N(\mu)}\}$ we have seen that for $k = 3, 4, \cdots, N(\mu)$, there is an eigenvector ψ_k^* of the adjoint problem and a function $y_k(t)$ which satisfies (5.3). (For $k = 1, 2$ we choose to leave the solution in the form (5.4).)

The solution $z_t(\phi)$ of (5.2) which satisfies $z_0(\phi)=\phi$ can be written in the form

$$z_t(\phi)=\phi_{11}w_{11}(t)+\phi_{21}w_{21}(t)+\operatorname{Re}\left[\sum_{j=3}^{N(\mu)}u_k y_k(t)\right]+z_{1t}(\mu,\phi).$$

Since $z_t(\phi)(0)=z(t,\phi)=(y(t),y'(t),\cdots,y^{(J-1)}(t))$, this is equivalent to

$$\begin{aligned}(y(t),\cdots,y^{(J-1)}(t))^T=\phi_{11}(0)w_{11}(t)+\phi_{21}(0)w_{21}(t)+\operatorname{Re}\left[\sum_{j=3}^{N}\xi(s_k)y_k(t)\right]\\ +z_{1t}(\mu,\phi)(0)\end{aligned}\tag{5.5}$$

where w_{11}, w_{21}, y_k and z_{1t} satisfy the equations

$$\begin{bmatrix}w_{11}\\ w_{21}\end{bmatrix}'=\begin{bmatrix}0 & -\omega_0\\ \omega_0 & 0\end{bmatrix}\begin{bmatrix}w_{11}\\ w_{21}\end{bmatrix}+2\begin{bmatrix}\operatorname{Re}P'(i\omega_0)^{-1}\\ -\operatorname{Im}P'(i\omega_0)^{-1}\end{bmatrix}n_1(y_t),\tag{5.6}$$

$$y_k'=s_k y_k+(P'(\bar{s}_k))^{-1}n_1(y_t),\qquad 3\leqq k\leqq N(\mu)\tag{5.7}$$

and

$$z_{1t}(\phi)=T(t)\phi_Q+\int_0^t T(t-s)[H_0-\Phi_\mu\Psi_\mu(0)]n_1(y_t)\,ds.\tag{5.8}$$

Moreover, by Theorem 3 we have $\|T(t)\phi_Q\|\leqq K(\mu)c^{-\mu t}\|\phi\|$ for all $\phi\in C[-\alpha,0]$ and $\|T(t)[H_0-\Phi_\mu\Psi_\mu(0)]\|\leqq K(\mu)\,e^{-\mu t}$ for some constant $K(\mu)$ such that $K(\mu)\to 0$ as $\mu\to\infty$.

Equation (5.6) will now be further transformed in the manner used in [1], [30]. Define $y_1(t)\triangleq y(t)-\phi_{11}(0)w_{11}(t)-\phi_{21}(0)w_{21}(t)$,

$$\Phi(t)\triangleq\begin{bmatrix}\cos\omega_0 t & \sin\omega_0 t\\ -\sin\omega_0 t & \cos\omega_0 t\end{bmatrix}\quad\text{and}\quad\begin{bmatrix}w_{11}\\ w_{21}\end{bmatrix}\triangleq\Phi(t)\begin{bmatrix}x_1\\ x_2\end{bmatrix}$$

so that

$$\begin{bmatrix}x_1\\ x_2\end{bmatrix}'=\Phi(-t)\begin{bmatrix}2\operatorname{Re}P'(i\omega_0)^{-1}\\ -2\operatorname{Im}P'(i\omega_0)^{-1}\end{bmatrix}n_1(y(t))\triangleq f(t,x_1,x_2,y_1).\tag{5.9}$$

Since we are assuming that $|P'(i\omega_0)|^{-1}$ is small, then averaging can be applied to (5.9). Let $f_0(x_1,x_2)$ be the mean value with respect to t of $f(t,x_1,x_2,0)$, i.e.,

$$f_0(x_1,x_2)=\frac{\omega_0}{2\pi}\int_0^{2\pi/\omega_0}f(t,x_1,x_2,0)\,dt.$$

Define

$$U(t,x_1,x_2)=\int_0^t[f(t,x_1,x_2,0)-f_0(x_1,x_2)]\,dt$$

and

$$\begin{bmatrix}x_1\\ x_2\end{bmatrix}=w+U(t,w),\qquad w=\begin{bmatrix}w_1\\ w_2\end{bmatrix}.\tag{5.10}$$

Clearly U is $(2\pi/\omega_0)$-periodic in t. We shall later show that U is continuously differentiable in x_1 and x_2 and that $\partial U/\partial x_i$ is Lipschitz continuous in (x_1,x_2) uniformly in t. In particular, the 2×2 Jacobian matrix $U_w=[\partial U_i/\partial x_j]$ exists and is continuous. Thus, the change in variables (5.10) can be used to see that

$$w' = f_0(w) + [I + U_w(t, w)]^{-1}[f(t, w + U(t, w), y_1) - f(t, w, 0) - U_w(t, w) f_0(w)]$$
$$\triangleq f_0(w) + f_1(t, w, y_1).$$

Here f_1 is the two-dimensional column vector whose components are f_{11} and f_{12}.

We shall now compute the averaged term $f_0(w)$. Recall that by the definition of the sinusoidal input describing function, if $n_0(z) = Nz(t-\alpha) - n(z)$,

$$\frac{\omega_0}{2\pi}\int_0^{2\pi/\omega_0} n_0(A\cos(\omega_0 t + \theta))\, e^{i\omega_0 t}\, dt = \frac{A}{2} e^{i\theta} N_0(A, \omega_0) \tag{5.11}$$

for any $A > 0$ and for any real number θ. Let $N_0(A, \omega_0) = N_{0R}(A) + iN_{0I}(A)$. On equating the real and imaginary parts of (5.11), we find that

$$\frac{\omega_0}{2\pi}\int_0^{2\pi/\omega_0} n_0(A\cos(\omega_0 t + \theta))\cos\omega_0 t\, dt = \frac{A}{2}[N_{0R}(A)\cos\theta - N_{0I}(A)\sin\theta)]$$

and

$$\frac{\omega_0}{2\pi}\int_0^{2\pi/\omega_0} n_0(A\cos(\omega_0 t + \theta))\sin\omega_0 t\, dt = \frac{-A}{2}[N_{0I}(A)\cos\theta + N_{0R}(A)\sin\theta].$$

Since $(y, y', \cdots, y^{(J-1)})^T = \phi_{11}(0)w_{11} + \phi_{21}(0)w_{21}$, then

$$Qy = \sum_{j=1}^{J} \gamma_{J-j} y^{(j)} = \operatorname{Re} q(i\omega_0) w_{11} + \operatorname{Im} g(i\omega_0) w_{21}.$$

Let $q(i\omega_0) = E e^{i\theta_1}$ where $E = |q(i\omega_0)|$. Let $x_1 = A\cos\theta_2$ and $x_2 = -A\sin\theta_2$. Then

$$\begin{aligned} Qy &= E\cos\theta_1(x_1\cos\omega_0 t + x_2\sin\omega_0 t) + E\sin\theta_1(-x_1\sin\omega_0 t + x_2\cos\omega_0 t) \\ &= E\cos\theta_1(A\cos(\omega_0 t + \theta_2)) + E\sin\theta_1(-A\sin(\omega_0 t + \theta_2)) \\ &= EA\cos(\omega_0 t + \theta) \end{aligned}$$

where $\theta = \theta_1 + \theta_2$. Then

$$f(t, x_1, x_2, 0) = \begin{bmatrix} \cos\omega_0 t & -\sin\omega_0 t \\ \sin\omega_0 t & \cos\omega_0 t \end{bmatrix} \begin{bmatrix} \operatorname{Re} P'(i\omega_0)^{-1} \\ -\operatorname{Im} P'(i\omega_0)^{-1} \end{bmatrix} n_0(EA\cos(\omega_0 t + \theta)).$$

Hence

$$f_0(x_1, x_2) = \begin{bmatrix} \frac{AE}{2}[N_{0R}(AE)\cos\theta - N_{0I}(AE)\sin\theta] & \frac{AE}{2}[N_{0I}(AE)\cos\theta + N_{0R}(AE)\sin\theta] \\ -\frac{AE}{2}[N_{0I}(AE)\cos\theta + N_{0R}(AE)\sin\theta] & \frac{AE}{2}[N_{0R}(AE)\cos\theta - N_{0I}(AE)\sin\theta] \end{bmatrix} \times \begin{bmatrix} \operatorname{Re} P'(i\omega_0)^{-1} \\ -\operatorname{Im} P'(i\omega_0)^{-1} \end{bmatrix}$$

or

$$f_0(x_1, x_2) = \frac{1}{2}\begin{bmatrix} N_{2R}(A)x_1 + N_{21}(A)x_2 & N_{2I}(A)x_1 - N_{2R}(A)x_2 \\ -N_{21}(A)x_1 + N_{2R}(A)x_2 & N_{2R}(A)x_1 + N_{2I}(A)x_2 \end{bmatrix} \times \begin{bmatrix} \operatorname{Re} P'(i\omega_0)^{-1} \\ -\operatorname{Im} P'(i\omega_0)^{-1} \end{bmatrix} \tag{5.12}$$

where $x_1^2 + x_2^2 = A^2$ and $N_2(A) = N_0(AE)q(i\omega_0)$.

We introduce polar coordinates $w_1 = (r+A_0)\cos\theta$, $w_2 = (r+A_0)\sin\theta$ so that

$$\begin{bmatrix} (r+A_0)\theta' \\ r' \end{bmatrix} = \frac{1}{(r+A_0)} \begin{bmatrix} w_1' w_2 - w_2' w_1 \\ w_1' w_1 + w_2' w_2 \end{bmatrix}.$$

Then (5.9) and (5.12) yield

$$\begin{aligned} r' = {} & \tfrac{1}{2}[N_{2R}(r+A_0)\ \mathrm{Re}\ P'(i\omega_0)^{-1}(r+A_0) - N_{2I}(r+A_0)\ \mathrm{Im}\ P'(i\omega_0)^{-1}(r+A_0)] \\ & + \cos\theta f_{11}(t, (r+A_0)\cos\theta, (r+A_0)\sin\theta, y) \\ & + \sin\theta f_{12}(t, (r+A_0)\cos\theta, (r+A_0)\sin\theta, y) \end{aligned}$$

and

$$\begin{aligned} \theta' = {} & \tfrac{1}{2}[N_{2R}(r+A_0)\ \mathrm{Im}\ P'(i\omega_0)^{-1} + N_{2I}(r+A_0)\ \mathrm{Re}\ P'(i\omega_0)^{-1}] \\ & + \{-\sin\theta f_{11} + \cos\theta f_{12}\}/(r+A_0). \end{aligned}$$

These two equations are more conveniently written in the form

$$\begin{aligned} r' = {} & \frac{(r+A_0)}{2}\ \mathrm{Re}\ \{N_0(E(r+A_0))q(i\omega_0)/P'(i\omega_0)\} + \cos\theta f_{11} + \sin\theta f_{12}, \\ \theta' = {} & \tfrac{1}{2}\ \mathrm{Im}\ \{N_0(E(r+A_0))q(i\omega_0)/P'(i\omega_0)\} + \{-\sin\theta f_{11} + \cos\theta f_{12}\}/(r+A_0). \end{aligned} \tag{5.13}$$

Let $K(\mu)$ be the constant associated with (5.8) and let $\varepsilon = \max\{|P'(s_k)|^{-1}: k = 1, 2, 3, \cdots, \}$. Define $\hat{y}_k = \varepsilon^{-1/2} y_k$ and $\hat{z}_{1t}(\phi) = K(\mu)^{-1/2} z_{1t}(\phi)$. Then (5.7) and (5.8) can be written in the form

$$\hat{y}_k' = s_k \hat{y}_k + (P'(\bar{s}_k)\sqrt{\varepsilon})^{-1} n_1(y(t)), \qquad 3 \leqq k \leqq N(\mu) \tag{5.14}$$

and

$$\hat{z}_{1t}(\phi) = T(t)\phi_Q K(\mu)^{-1/2} + \int_0^t T(t-s)[H_0 - \Phi_\mu \Psi_\mu(0)] K(\mu)^{-1/2} n_1(y(s))\, ds. \tag{5.15}$$

Here

$$z(t, \phi) = \phi_{11}(0) w_{11}(t) + \phi_{21}(0) w_{21}(t) + \sqrt{\varepsilon}\ \mathrm{Re}\left[\sum_{j=3}^{N(\mu)} u_k(0)\hat{y}_k(t)\right] + K(\mu)^{1/2}\hat{z}_{1t}(\phi)(0)$$

where $\phi_{11}(0) = \mathrm{Re}\ \xi(i\omega_0)$, $\phi_{21}(0) = \mathrm{Im}\ \xi(i\omega_0)$ and $u_k(0) = \xi(s_k)$. Now

$$\begin{aligned} & \phi_{11}(0) w_{11}(t) + \phi_{21}(0) w_{21}(t) \\ & \quad = \phi_{11}(0)[x_1 \cos\omega_0 t + x_2 \sin\omega_0 t] + \phi_{21}(0)[-x_1 \sin\omega_0 t + x_2 \cos\omega_0 t] \\ & \quad = \phi_{11}(0)[(w_1 + U_1(t, w))\cos\omega_0 t + (w_2 + U_2(t, \omega))\sin\omega_0 t] \\ & \qquad + \phi_{21}(0)[-(w_1 + U_1(t, \omega))\sin\omega_0 t + (w_2 + U_2(t, w))\cos\omega_0 t] \\ & \quad = \phi_{11}(0)[w_1 \cos\omega_0 t + w_2 \sin\omega_0 t] + \phi_{21}[-w_1 \sin\omega_0 t + w_2 \cos\omega_0 t] + \mathcal{O}(\varepsilon) \\ & \quad = \phi_{11}(0)[(r+A_0)\cos\theta\cos\omega_0 t - (r+A_0)\sin\theta\sin\omega_0 t] \\ & \qquad + \phi_{21}(0)[-(r+A_0)\cos\theta\sin\omega_0 t - (r+A_0)\sin\theta\cos\omega_0 t] + \mathcal{O}(\varepsilon) \\ & \quad = \phi_{11}(0)(r+A_0)\cos(\omega_0 t + \theta) - \phi_{21}(0)(r+A_0)\sin(\omega_0 t + \theta) + \mathcal{O}(\varepsilon). \end{aligned}$$

Hence, for $t \geqq 0$ we have

$$z(t, \phi) = \phi_{11}(0)(r + A_0)\cos(\omega_0 t + \theta) - \phi_{21}(0)(r + A_0)\sin(\omega_0 t + \theta)$$
$$+ \sqrt{\varepsilon}\,\mathrm{Re}\sum_{k=3}^{N(\mu)} \xi(s_k)\hat{y}_k(t) + \mathcal{O}(\varepsilon) + \mathcal{O}(K(\mu)^{1/2}).$$

In particular,

$$y(t) = (r(t) + A_0)\cos(\omega_0 t + \theta) + \sqrt{\varepsilon}\,\mathrm{Re}\left(\sum_{k=3}^{N(\mu)} \hat{y}_k(t)\right) + \mathcal{O}(\varepsilon) + \mathcal{O}(K(\mu)^{1/2}) \tag{5.16}$$

and

$$y^{(j)} = (r(t) + A_0)\frac{d^j}{dt^j}(\cos(\omega_0 t + \theta))$$
$$+ \sqrt{\varepsilon}\,\mathrm{Re}\left(\sum_{k=3}^{N(\mu)} s_k^j \hat{y}_k(t)\right) + \mathcal{O}(\varepsilon) + \mathcal{O}(K(\mu)^{1/2}). \tag{5.17}$$

The terms $\mathcal{O}(\varepsilon)$ and $\mathcal{O}(K(\mu)^{1/2})$ are small, independently of j, when ε and $K(\mu)$ are small, and these terms have small Lipschitz constants. Hence, if we take into account the smoothness results for $U(t, w)$ which are proved in the next section, we see that Theorem 5 can be applied to (5.13), (5.14), (5.15) and (5.16) with $C_\mu[-\alpha, 0] = Q_\mu$ to complete the proof of Theorem 1 when $\alpha > 0$.

We also note that by taking μ large we can make $K(\mu)$ as small as desired without affecting the constant ε. Hence, the necessary assumption for system (5.13)-(5.16) is not that $K(\mu)$ and ε be small, but only that ε be small. We conjecture that it is really only necessary to assume that $|P'(i\omega_0)|^{-1}$ is small but we have no proof as yet.

Finally, we note that the distortions from a sinusoidal curve for the high derivatives $y^{(j)}(t)$ can readily be explained by using (5.17). Since $\mathcal{O}(\varepsilon)$ and $\mathcal{O}(K(\mu)^{1/2})$ are small independently of j, then to good approximation solutions near the integral manifold S which was obtained in Theorem 1 have the form

$$y^{(j)}(t) \cong (r(t) + A_0)\frac{d^j}{dt^j}\left(\cos(\omega_0 t + \theta) + \sqrt{\varepsilon}\,\mathrm{Re}\left(\sum_{k=3}^{N(\mu)} s_k^j \hat{y}_k(t)\right)\right).$$

When $\sqrt{\varepsilon}$ is small, $r(t)$ is small and the $\hat{y}_k$'s are small. Thus $y(t) \cong A_0 \cos(\omega_0 t + \theta)$. However, the numbers s_k are complex numbers and Im s_k grows rapidly as k increases. Hence, for moderate sized integers j the quantity

$$\mathrm{Re}\left[\sqrt{\varepsilon}\sum_{k=3}^{N(\mu)} s_k^j \hat{y}_k(t)\right] \tag{5.18}$$

can be large, even though $\sqrt{\varepsilon}$ and $\hat{y}_k(t)$ are small. This is the source of the distortions discussed in § 1. We note that in the stable case $\hat{y}_k(t) \to 0$ as $t \to \infty$. Hence, the distortions are not usually too severe. On the other hand, if for some k $\hat{y}_k(t)$ is growing, then (5.18) seems to predict the possibility of very severe distortions. Several numerical simulations indicate that this does indeed happen.

When $\alpha = 0$, the proof is almost the same. Replace $C[-\alpha, 0]$ by $C[-\alpha_1, 0]$ for any $\alpha_1 > 0$. Since in this case L_1 is an ordinary differential operator (with no delay), there will be only a finite number (i.e. J) eigenvalues. This means that μ can be taken so large that the ϕ_Q and $H_0 - \Phi_\mu \Psi_\mu$ in (5.15) are zero. Hence, Theorem 5 can again be applied but now with $C_\mu[-\alpha_1, 0]$ equal to the zero subspace.

When $\alpha = 0$, it is really only necessary to assume that β_1 and β_2 are small. A different proof is needed to see this. The transformations used in [30], [31] can be used to reduce (5.1) to the form

$$\begin{aligned}
\frac{dr}{d\tau} &= \frac{r+A_0}{2}\operatorname{Re}\{N_0(EA)q(i\omega_0)(\beta_1 - i\beta_2)\} + \cos\theta f_{11} + \sin\theta f_{12}, \\
\frac{d\theta}{d\tau} &= \frac{1}{2}\operatorname{Im}\{N_0(EA)q(i\omega_0)(\beta_1 - i\beta_2)\} + \{-\sin\theta f_{11} + \cos\theta f_{12}\}/(r+A_0), \\
\frac{dx_4}{dt} &= C_1 x_4 - [\beta_1\hat{\xi}_1 + \beta_2\hat{\xi}_2]n_1(y_t).
\end{aligned} \tag{5.19}$$

Here $\tau = \omega_0 t$, $\beta_1 - i\beta_2 = 2/\omega_0 P'(i\omega_0)$, $\varepsilon = |\beta_1 - i\beta_2|$, C_1 is the companion matrix for the $(J-2)$-degree polynomial

$$\prod_{j=3}^{J}(s - s_j/\omega_0),$$

and $\hat{\xi}_1$ and $\hat{\xi}$, are the $J-2$ vectors

$$\hat{\xi}_1 = (1, 0, -1, 0, 1, 0, -1, \cdots)^T, \qquad \hat{\xi}_2 = (0, 1, 0, -1, 0, 1, \cdots)^T.$$

Moreover, $y(t) = y(\tau/\omega_0)$ has the form

$$y(\tau/\omega_0) = (r+A_0)\cos(\tau + \theta + \theta_1) + \sqrt{\varepsilon}\, B_2 x_4 + \mathcal{O}(\varepsilon)$$

where $\theta_1 = \arg q(i\omega_0)$. For ε small (i.e. for β_1 and β_2 small), Theorem 5 can be applied to (5.19) with $C_\mu[-\alpha, 0]$ equal to the zero subspace. This proves the assertion.

6. Smoothness of U. It was shown in [1, § 5] that if $n(y)$ is a function (i.e., $n(y)$ exhibits no hysteresis), then the function $U(t, w)$ defined in § 5 has continuous derivatives $\partial U/\partial w_1$ and $\partial U/\partial w_2$ which are Lipschitz continuous in w. The integral used to define $U(t, w)$ can be decomposed into pieces of the form

$$\int_0^t [\cos\omega_0 s]n_1(r\cos(\theta + \omega_0 s))\, ds$$

and

$$\int_0^t [\sin\omega_0 s]n_1(r\cos(\theta + \omega_0 s))\, ds$$

where $w_1 = r\cos\theta$ and $w_2 = -r\sin\theta$. These two integrals can be decomposed into pieces on which $\cos(\theta + \omega_0 s)$ is either strictly increasing or strictly decreasing. When $y = \cos(\theta + \omega_0 s)$ is increasing $n_1(y)$ can be replaced by $n_{1L}(y)$ (and when decreasing it can be replaced by $n_{1U}(y)$). Since n_{1L} and n_{1U} are ordinary functions, the results in [1] still apply.

7. The incorrectness of the quasi-static stability analysis. Given the feedback equation (2.1) and the solution ω_0 and A_0 of (2.2), the quasi-static stability analysis proceeds as follows (cf., e.g., [2, pp. 120–125]). Define $V(a, s)$ by the relation

$$V(a, s) = 1 + N(a)G(s), \qquad G(s) = q(s)/p(s).$$

It is required that all roots s_j of the polynomial $p(s) + N(A_0)q(s)$ with $s_j \neq \pm i\omega_0$ have

negative real parts and, in addition

$$\text{Im}\left(\overline{\frac{\partial V}{\partial a}}\frac{\partial V}{\partial s}i\right)=\text{Re}\left(\overline{\frac{\partial V}{\partial a}}\frac{\partial V}{\partial s}\right)>0 \tag{7.1}$$

at $a=A_0E$ and $s=i\omega_0$. Condition (7.1) has a well-known graphical interpretation. An equivalent statement is that for some A_1 and A_2 with $A_1<A_0<A_2$ the polynomial $p(s)+N(A)q(s)$ is stable when $A_0<A<A_2$ and is unstable when $A_1<A<A_0$.

Condition (7.1) can be written in terms of N, p and q as follows. Since $1+N(A_0E)G(i\omega_0)=0$, then at $a=A_0E$, $s=i\omega_0$ we have

$$G'=\frac{q'}{p}-\frac{qp'}{p^2}=\frac{q'}{p}+\frac{p'}{NP}=\frac{p'+Nq'}{NP}.$$

Hence, at $a=A_0E$ and $s=i\omega_0$

$$\begin{aligned}\overline{\frac{\partial V}{\partial a}}\frac{\partial V}{\partial s}&=\bar{N}'\bar{G}NG'=\left(\overline{\frac{N'}{G'}}\right)\bar{G}N|G'|^2\\&=\left(\overline{\frac{N'}{p'+Nq'}}\right)(\overline{Np})\bar{G}N|G'|^2\\&=\left(\overline{\frac{N'q}{p'+Nq'}}\right)|N|^2|G'|^2,\end{aligned}$$

so that (7.1) reduces to

$$\text{Re}\left\{\frac{N'(A_0E)q(i\omega_0)}{p'(i\omega_0)+N(A_0)q'(i\omega_0)}\right\}>0. \tag{7.2}$$

The stability condition corresponding to (7.2) which was obtained from Theorem 1 of the paper is that

$$\text{Re}\left\{\frac{N'(A_0E)q(i\omega_0)}{P'(i\omega_0)}\right\}>0, \tag{7.3}$$

or equivalently that the real part of

$$N'(A_0E)q(i\omega_0)/(p'(i\omega_0)+N(A_0E)q'(i\omega_0)-\alpha N(A_0E)q(i\omega_0))$$

is positive. It is also required that all roots of

$$P(s)=p(s)+|N(A_0E)|\,e^{-\alpha s}q(s)$$

other than $\pm i\omega_0$ have negative real parts. Our stability criterion can be interpreted in the following way. We write (2.1) in the form

$$Ly(t)+|N(A_0E)|Qy(t-\alpha)+n_2(y)=0, \tag{7.4}$$

where $n_2(y)=n(Qy)-|N(A_0E)|Qy(t-\alpha)$. We now apply the quasi-static stability analysis to (7.4). In doing so, we define

$$V_1(a,s)=1+[N(a)-|N(A_0E)|q(s)\,e^{-\alpha s}]q(s)/P(s).$$

We require that all roots of $P(s)=0$ other than $\pm i\omega_0$ have negative real parts and that at $a=A_0E$, $s=i\omega_0$

$$\text{Re}\left\{\overline{\frac{\partial V_1}{\partial a}}\frac{\partial V_1}{\partial s}\right\}>0. \tag{7.5}$$

Inequalities (7.3) and (7.5) are equivalent and can be graphically verified. The condition that all roots of $P(s)$ have negative real parts (or that one has positive real part) can also be graphically verified.

This situation is rather interesting. The quasi-static stability analysis usually seems to yield correct predictions for the stability of limit cycles. However, our stability criteria do not agree with this well-known method. It follows that one of the two methods of prediction must be incorrect. In the next section, we discuss a sixth order equation whose periodic solution is predicted to be stable according to the quasi-static analysis and is predicted to be unstable according to Theorem 1 of this paper. Numerical simulations show that this periodic solution is unstable.

8. An example. The quasi-static stability analysis will usually yield a correct stability prediction. This is because the number α is normally rather small. When α is small enough, (7.2) will imply (7.3). The condition that all roots of $P(s)$ except $\pm i\omega_0$ have negative real parts is equivalent to the condition that the graph

$$\left\{\frac{q(i\omega)}{p(i\omega)}\, e^{-i\alpha(\omega-\omega_0)}\colon -\infty<\omega<\infty\right\}$$

does not touch the point $-1/N(A_0E)$ (except at $\pm i\omega_0$) and does not encircle this point. When $|\omega|$ is large, then $|q(i\omega)/p(i\omega)|$ will be less than $|N(A_0E)|^{-1}$. Hence, for α small, this graphical condition is not very different from the same graphical condition for $\{q(i\omega)/p(i\omega)\colon -\infty<\omega<\infty\}$. However, it is possible to find examples with α so large that the quasistatic stability analysis gives an incorrect prediction while Theorem 1 of the present paper yields a correct prediction of limit cycle stability. We shall now present one such example.

Let $q(s)=1$ while

$$p(s)=(s^2+0.5s+0.09)(s^2+0.8s+9)(s^2+0.03s+31)$$

or

$$p(s)=s^6+1.33s^5+40.529s^4+45.1567s^3+295.13716s^2+141.7563s+25.11.$$

Consider the sixth order equation

$$P(D)y+250n(y)=0 \tag{8.1}$$

where $n(y)$ is the hysteresis nonlinearity depicted in Fig. 7. The describing function for this nonlinearity is known to be

$$N(A)=2-\frac{2}{\pi}\left\{\sin^{-1}\left(\frac{2}{A}\right)+\frac{2}{A}\sqrt{1-\frac{4}{A^2}}\right\}-\frac{8i}{\pi A^2}.$$

For the given $p(s)$ and $N(A)$ the equation $p(i\omega)+250N(A)=0$ has a solution ω_0 and A_0 with $A\cong 2.91$ and $\omega_0\cong 1.293$. (Refer to the polar plot in Fig. 8.) It is clear from Fig. 8 that the quasistatic stability analysis yields the prediction that the limit cycle of (8.1), which is approximately equal to $A_0\cos\omega_0 t$, is stable.

Given A_0 and ω_0 we can compute $\mathrm{Re}\,[N'(A_0)/P'(i\omega_0)]$. This number is positive as required for stability. Fig. 9 depicts a polar plot of

$$100P(i\omega)^{-1}=100[p(i\omega)+|N(A_0)|\,e^{i\alpha\omega}]^{-1}.$$

We see that the curve of $100P(i\omega)^{-1}$ encircles the point $-(2.5|N(A_0)|)^{-1}$. Hence, $P(s)=0$ has a root s_3 with positive real part, i.e., Theorem 1 predicts that the limit

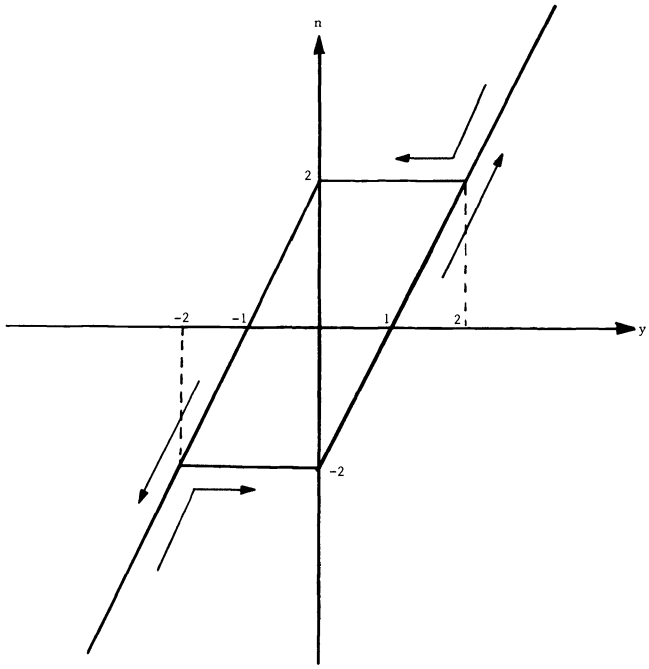


FIG. 7. *A hysteresis nonlinearity.*

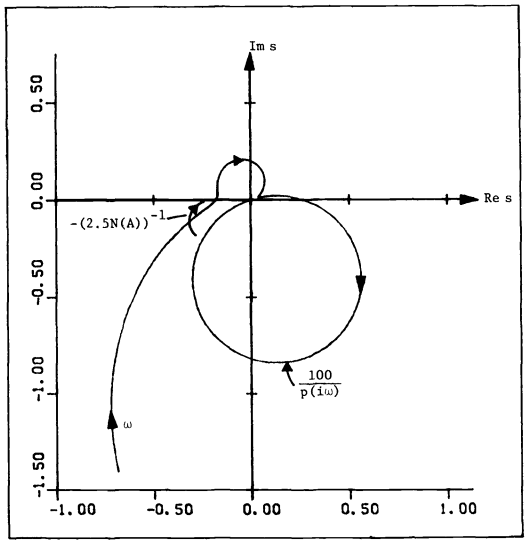


FIG. 8. *Polar plot of* $100/p(i\omega)$ *and* $-(2.5N(A))^{-1}$.

cycle is unstable. A few calculations show that the unstable root s_3 is approximately equal to $0.022+5.575i$. Moreover, since

$$\frac{1}{P'(i\omega_0)} \cong -0.0000689+0.0020285i$$

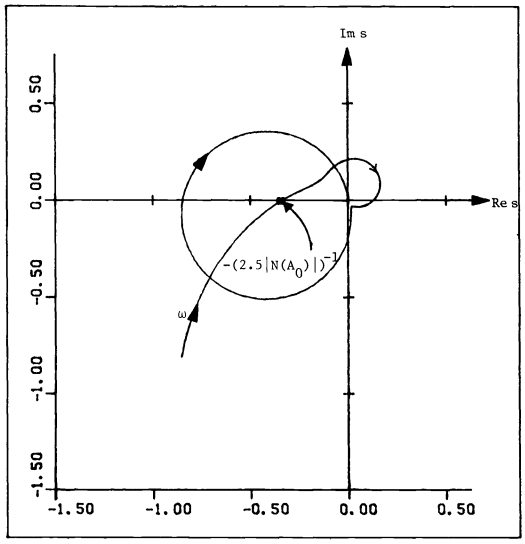


FIG. 9. *Polar plot of* $100/P(i\omega)$.

and

$$\frac{1}{P'(s_3)} \cong -0.000038515 + 0.00010724i$$

then the first four of the parameters $|P'(s_i)|^{-1}$ are certainly small. Hence, it seems likely that Theorem 1 will yield the correct stability prediction.

Solutions of (8.1) were simulated using initial conditions near the predicted periodic solution. These numerical simulations verify our theoretical predictions. Only one (apparent) periodic solution $y(t)$ was found anywhere near the predicted periodic solution. This periodic solution appears to be nearly sinusoidal with frequency and amplitude very near the predicted values $A_0 = 2.91$ and $\omega_0 = 1.293$. Some of the results of our simulations are displayed in Fig. 10 for $0 \leqq t \leqq 6$. Clearly $y(t)$ is unstable. The

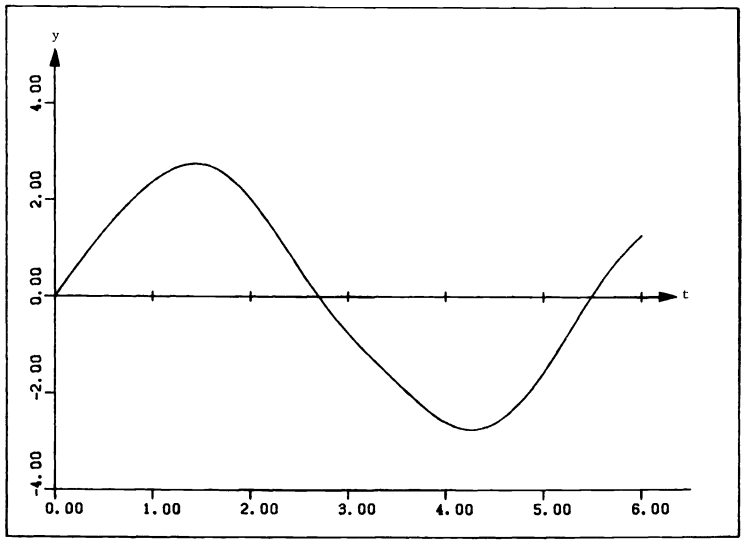


FIG. 10a. *Solution* $y(t)$.

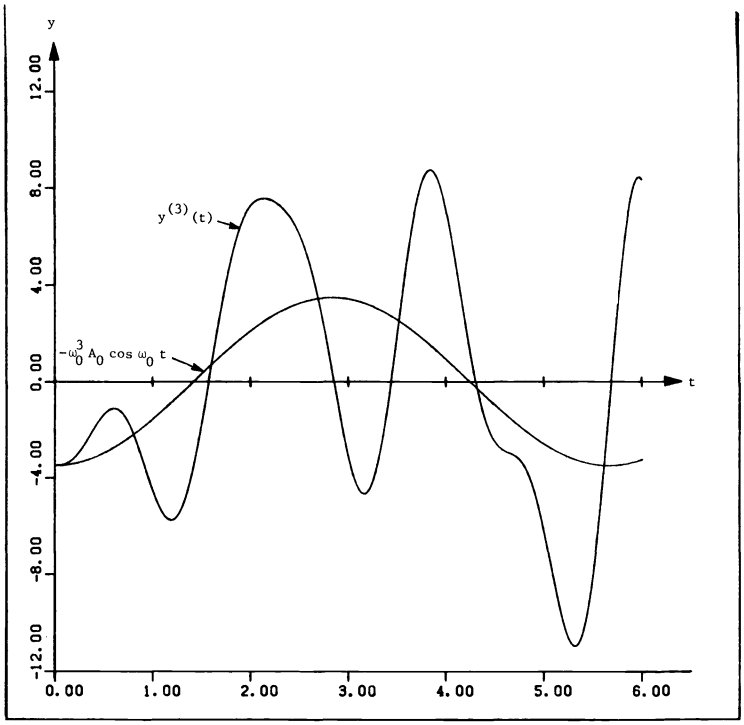


FIG. 10b. $y^{(3)}(t)$ *and* $-\omega_0^3 A_0 \cos \omega_0 t$.

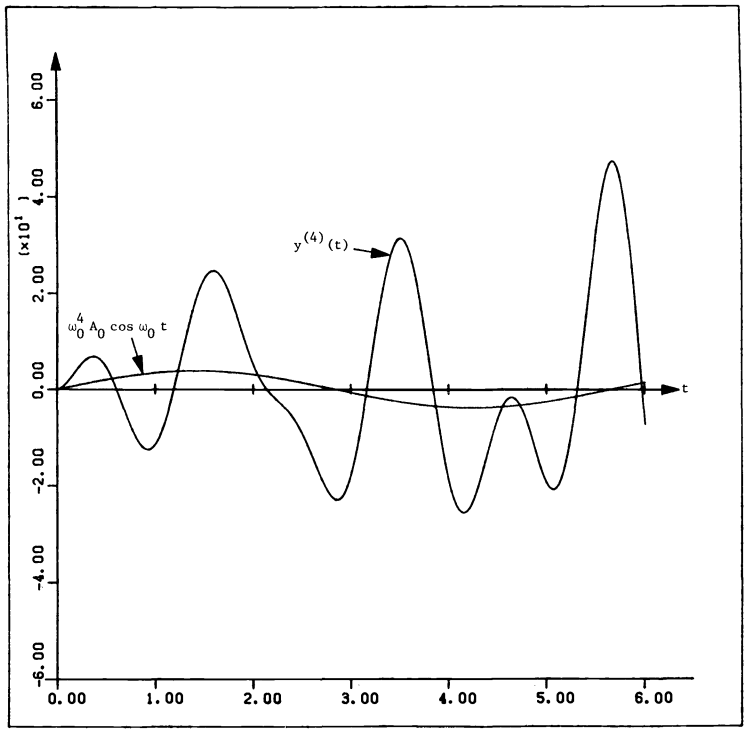


FIG. 10c. $y^{(4)}(t)$ *and* $-\omega_0^4 A_0 \cos \omega_0 t$.

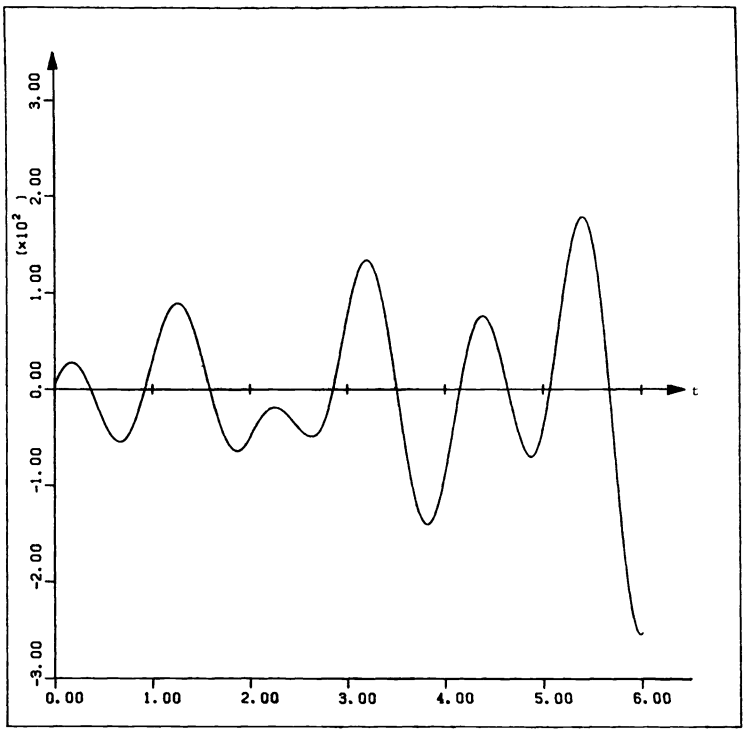


FIG. 10d. $y^{(5)}(t)$.

instability shows up first in the high derivatives of $y(t)$ and then is exhibited in the lower order derivatives. The unstable behavior becomes even more pronounced for $t > 6$.

From the present example we can conclude that the quasistatic stability analysis is not always correct. From the simulation of this example and from several other simulations of third through sixth order equations (not included here) we feel that Theorem 1 does provide a reliable stability analysis when $|P'(i\omega_0)|^{-1}$ is small.

9. Concluding remarks. We have presented an analysis (Theorem 1) for the stability of limit cycles for feedback systems with hysteresis nonlinearities. To obtain these results, we have made use of certain results for functional differential equations and of a result for integral manifolds. Theorem 1 constitutes a generalization of a result for feedback systems with nonlinearities which do not exhibit hysteresis.

A specific example was included for which Loeb's criterion does not predict correctly the stability of a limit cycle while Theorem 1 predicts correctly the stability of the limit cycle.

Like most stability criteria obtained via linearization, our result is local, that is, certain parameters must be "sufficiently small". Nevertheless, Theorem 1 will allow the analyst to proceed with much more confidence when using the method of describing functions. Moreover, our results also correctly predict in remarkable detail the behavior of solutions which are near the (apparent) sinusoidal periodic solution, e.g., the rate of approach to the periodic solution, the speed-up or slow-down of the phase as the periodic solution is approached, and the distortions in derivatives.

REFERENCES

[1] R. K. MILLER, A. N. MICHEL AND G. S. KRENZ, *On the stability of limit cycles in nonlinear feedback systems: analysis using describing functions*, IEEE Trans. Circuits and Systems, 30 (1983), pp. 684-696.

[2] A. GELB AND W. E. VAN DER VELDE, *Multiple-Input Describing Functions and Nonlinear Systems*, McGraw-Hill, New York, 1968.

[3] J. E. GIBSON, *Nonlinear Automatic Control*, McGraw-Hill, New York, 1963.

[4] J. K. HALE, *Ordinary Differential Equations*, Wiley Interscience, New York, 1969.

[5] ———, *Oscillations in Nonlinear Systems*, McGraw-Hill, New York, 1963.

[6] S. P. DILIBERTO, *Perturbation theorems for periodic surfaces*, Circ. Mat. Palermo, 2 (1960), pp. 265-2435.

[7] J. M. HOLTZMAN, *Contraction maps and equivalent linearization*, Bell System Tech. J., 46 (1967), pp. 2405-2435.

[8] A. I. MEES, *The describing function matrix*, J. Inst. Math. Appl., 10 (1972), pp. 49-67.

[9] A. I. MEES AND A. R. BERGEN, *Describing functions revisited*, IEEE Trans. Automat. Control, 20 (1975), pp. 473-478.

[10] A. I. MEES AND L. O. CHUA, *The Hopf bifurcation theorem and its application to nonlinear oscillations in circuits and systems*, IEEE Trans. Circuits and Systems, 26 (1977), pp. 235-254.

[11] R. K. MILLER AND A. N. MICHEL, *On existence of periodic motions in nonlinear control systems with periodic inputs*, this Journal, 18 (1980), pp. 585-598.

[12] I. W. SANDBERG, *On the response of nonlinear control systems to periodic input signals*, Bell System Tech. J., 43 (1964), pp. 911-926.

[13] ———, *Some results on the theory of physical systems governed by nonlinear functional equations*, Bell System Tech. J., 44 (1965), pp. 871-898.

[14] L. CESARI, *Functional analysis and periodic solutions of nonlinear differential equations*, Contributions to Differential Equations, 1 (1961), pp. 149-187.

[15] M. URABE, *Galerkin's procedure for nonlinear periodic systems*, Arch. Rational Mech. Anal., 20 (1965), pp. 120-152.

[16] ARNOLD STOKES, *On the approximation of nonlinear oscillations*, J. Differential Equations, 12 (1972), pp. 537-558.

[17] S. J. SKAR, R. K. MILLER AND A. N. MICHEL, *On the nonexistence of limit cycles in interconnected systems*, IEEE Trans. Automat. Control, 26 (1981), pp. 669-676.

[18] G. CAHEN, *Perturbations des oscillateurs filtres*, Comptes. Rend. Acad. Sci., 235 (1952), pp. 1614-1617.

[19] J. M. LOEB, *Recent advances in nonlinear servo theory*, in Frequency Responses, R. Oldenberger, ed., Macmillan, New York, 1956, pp. 260-268.

[20] E. P. POPOV, *The Dynamics of Automatic Control Systems*, Addison-Wesley, Reading, MA, 1962 (see especially p. 586).

[21] R. A. JOHNSON AND B. W. LEACH, *Stability of oscillations in low-order nonlinear systems*, IEEE Trans. Automat. Control, 17 (1972), pp. 672-675.

[22] J. K. HALE, *Functional Differential Equations*, Springer-Verlag Series in Applied Mathematical Sciences, No. 3, Springer-Verlag, New York, Heidelberg and Berlin, 1971.

[23] ———, *Theory of Functional Differential Equations*, Springer-Verlag, New York, Heidelberg and Berlin, 1977.

[24] RICHARD BELLMAN AND KENNETH L. COOKE, *Differential Difference Equations*, Academic Press, New York, 1963.

[25] S. G. KREIN, *Linear Differential Equations in Banach Space*, American Mathematical Society, Providence, RI, 1972.

[26] E. HILLE AND R. S. PHILLIPS, *Functional Analysis and Semigroups*, Amer. Mathematical Society, Providence, R.I., 1957.

[27] E. A. CODDINGTON AND N. LEVINSON, *Theory of Differential Equations*, McGraw-Hill, New York, 1955.

[28] R. K. MILLER AND A. N. MICHEL, *Ordinary Differential Equations*, Academic Press, New York, 1982.

[29] P. R. HALMOS, *Measure Theory*, Van Nostrand, New York, 1950 (Section 39).

[30] R. K. MILLER, A. N. MICHEL AND G. S. KRENZ, *On the stability of limit cycles in nonlinear feedback systems: improved results*, IEEE Trans. Circuits and Systems, CAS. 31 (1984), pp. 561-567.

[31] G. S. KRENZ AND R. K. MILLER, *Qualitative analysis of oscillations in nonlinear control systems: a describing function approach*, Proc. 1984 Midwest Ordinary Differential Equations Conference, to appear.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 2, March 1986

009

GENERALIZED QUASICONVEX MAPPINGS AND VECTOR OPTIMIZATION*

J. JAHN† AND E. SACHS‡

Abstract. Quasiconvexity for mappings is generalized in such a way that this notion gives the sufficiency of necessary optimality conditions such as multiplier rules. One can show that it is the weakest type of generalized convexity notions in the sense that this generalized quasiconvexity holds if certain multiplier rules are sufficient for optimality. It also yields the equivalence of local and global minima. The theory is applied to a multi-objective programming problem and a vector approximation problem.

Key words. multi-objective programming, necessary optimality conditions, quasiconvexity, vector approximation

AMS(MOS) subject classification. 49B27

1. Introduction. In optimization theory the notion of convexity plays an important role for the development of powerful theorems concerning existence of optimal points, their characterization and numerical computation. In this paper we focus mainly on the characterization of optimal points using convexity conditions. According to the particular class of problems treated, there are various generalizations of the definition of a convex function which are already discussed in several textbooks on optimization theory, e.g. Mangasarian [16], Avriel [1] and Krabs [12]. In addition, a number of studies on the comparison of different convexity definitions have been published, see e.g. Ponstein [23] among the earliest and for most recent surveys Schaible/Ziemba [25].

Let us recall the most widely used generalizations of a convex real-valued function in optimization theory: A function $f:\mathbb{R}^n \to \mathbb{R}$ is called *quasiconvex* if for all $\lambda \in [0,1]$, $x, \bar{x} \in \mathbb{R}^n$ the inequality $f(x) \leqq f(\bar{x})$ implies $f(\lambda x + (1-\lambda)\bar{x}) \leqq f(\bar{x})$ or if the level sets $L_\alpha = \{x \in \mathbb{R}^n | f(x) \leqq \alpha\}$ are convex sets for all $\alpha \in \mathbb{R}$.

In the context with optimization and game theory this definition of quasiconvexity is mostly credited to Nikaidô [19], but it is worthwhile to mention that twenty-six years earlier von Neumann [18] introduced the same class of functions using the definition with the level sets. He applied it to prove a saddlepoint theorem in game theory and called them "functions with the property (K)".

A differentiable function $f:\mathbb{R}^n \to \mathbb{R}$ is called *pseudoconvex* if for all $x, \bar{x} \in \mathbb{R}^n$ with $\langle \nabla f(\bar{x}), x - \bar{x} \rangle \geqq 0$ we also have $f(x) \geqq f(\bar{x})$. This definition can be found in Mangasarian [15] and Tuy [26].

These definitions and stronger or weaker versions can be used to prove for example that local minima are global ones or that necessary conditions for minima are also sufficient.

For vector optimization problems extensions of these definitions have been introduced by Hartwig [5] for finite-dimensional problems and by Craven [4], Nehse [17] and Peemöller [21] for problems in infinite-dimensional spaces. In some of these papers it is shown that under certain convexity assumptions the Karush-Kuhn-Tucker multiplier rule as a necessary optimality condition is also sufficient for optimal points.

Since the notion of a convex set or a convex function is purely geometric and does not require any topological properties, we formulate most of the results in linear

* Received by the editors January 17, 1984, and in revised form November 7, 1984.

† Fachbereich Mathematik, Technische Hochschule Darmstadt, 6100 Darmstadt, West Germany. This paper was written when this author was a visitor at the Department of Mathematics of the North Carolina State University in Raleigh.

‡ Department of Mathematics, North Carolina State University, Raleigh, North Carolina 27695-8205.

spaces. The second section deals with generalizations of quasiconvex mappings. Quasiconvexity as defined earlier is described by a convexity requirement in the domain of the mapping. We define an extension of this concept and also its differentiable version. The relation to pseudoconvexity is discussed together with examples.

The third section begins with the formulation of minimal and weakly minimal points in vector optimization. In § 3.2 we consider problems with equality and inequality constraints in infinite-dimensional spaces where it is shown that a certain type of convexity for the function involved is characteristic for the sufficiency of the generalized Karush-Kuhn-Tucker multiplier rule. In this sense, the convexity condition for the functions of the vector optimization problem cannot be relaxed. Also for this problem we show the equivalence of two types of Lagrange multiplier rules for vector optimization problems under a convex-likeness condition. These results extend the approaches by Hartwig [5] and Craven [4] for vector optimization problems because the C-quasiconvexity is characteristic for the sufficiency of multiplier rules. For scalar-valued objectives a similar condition has been given by Krabs [12]. However, when his almost-pseudoconvexity condition is applied to a finite-dimensional nonlinear programming problem with inequality constraints then all functions describing the inequalities have to be quasiconvex, not only those with active indices as we obtain from our general theory and as one can find in e.g. Mangasarian [16, p. 151].

Since for strictly quasiconvex functions local and global minima coincide, e.g. Mangasarian [16, p. 139], we treat this question in § 3.3. We show that the cor (C_Y)-quasiconvexity is necessary and sufficient for the property that local weakly minimal points are also global weakly minimal. A statement of this type holds for minimal points in vector optimization. In the scalar-valued case, similar characterizations have been obtained in Zang and Avriel [29].

In a final subsection we consider a nonlinear multi-objective programming problem and show which "classical" condition on the functions describing the optimization problem yield the C-quasiconvexity as defined in (16). A second application is a vector approximation problem. Under a certain representation condition for the family of functions which defines the approximation problem we prove that the differentiable C-quasiconvexity is satisfied. This applies to rational vector approximation with the maximum norm and extends a result by Krabs [11] for the real-valued case. Finally we present a generalized Kolmogorov condition.

In the following we list some symbols and notations which are used in this paper. Let X, Y be real linear spaces and let A, B, C be nonempty subsets of X, then:

$X' := \{l: X \to \mathbb{R} | l \text{ is linear on } X\}$,
$A + B := \{a + b | a \in A, b \in B\}$,
$[a, b] := \{\lambda a + (1-\lambda) b | \lambda \in [0, 1]\}$ for all $a, b \in X$
(and $(a, b), [a, b), (a, b]$ are defined in the same way),
$\operatorname{cor}(A) := \{a \in A | \text{for each } x \in X \text{ there exists some } \alpha > 0 \text{ with } [a, a + \alpha x] \subset A\}$
denotes the algebraic interior of A,
$\operatorname{lin}(A) := \{x \in X | \text{there is some } x \neq a \in A \text{ with } [a, x) \subset A\} \cup A$
denotes the algebraic closure of A,
$\operatorname{cone}(A) := \{\alpha x | \alpha \geqq 0, x \in A\}$,
$f(A) := \{f(a) | a \in A\}$ for a function $f: A \to Y$,
$\operatorname{span}(A) :=$ smallest linear subspace of X containing A.

The set A is called *algebraically open* (or *algebraically closed*), if $\operatorname{cor} A = A$ (or $\operatorname{lin}(A) = A$). The set C is said to be a *cone*, if $\operatorname{cone}(C) \subset C$. It is well known that $C = C + C$ for a convex cone C and $\operatorname{cor}(C) = \operatorname{cor}(C) + C$ for a convex cone C with nonempty

algebraic interior. A cone C is called pointed, if $C \cap (-C) = \{0_X\}$. A convex cone C induces a partial ordering $\geqq$ in X by $x \geqq 0_X$ being equivalent to $x \in C_X$. This implies a partial ordering also for the dual space X' by the *dual cone*

$$C_{X'} := \{l \in X' | l(c) \geqq 0 \text{ for all } c \in C\}.$$

The obvious extension of convex functions to partially ordered real linear spaces is the following.

DEFINITION 1.1. Let E, F be real linear spaces, C a convex cone in F, S a nonempty convex subset of E and $v: S \to F$ a mapping. v is called *convex* if $s_1, s_2 \in S$ and $\lambda \in [0, 1]$ imply

$$\lambda v(s_1) + (1-\lambda) v(s_2) - v(\lambda s_1 + (1-\lambda) s_2) \in C. \tag{1}$$

Recall the following generalization of the convexity notion (e.g., see Vogel [27] and Nehse [17]).

DEFINITION 1.2. Let E, F be real linear spaces, C a convex cone in F, S a nonempty subset of E and $v: S \to F$ a mapping. v is called *convex-like* if $v(S) + C$ is convex.

It can be shown that each convex mapping is also convex-like.

2. Generalized quasiconvex mappings. In this section we investigate some generalizations of the quasiconvexity notion. The definition of quasiconvex functionals on $\mathbb{R}^n$ given in the previous section can be carried over easily to the vector-valued case (e.g., see Hartwig [5, p. 304] and Peemöller [21, p. 134] among others).

DEFINITION 2.1. Let E, F be real linear spaces, C a convex cone in F, S a nonempty convex subset of E and $v: S \to F$ a mapping. v is called *quasiconvex* if

$$s_1, s_2 \in S \quad \text{with} \quad v(s_1) - v(s_2) \in C \tag{2}$$

implies that

$$v(s_1) - v(\lambda s_1 + (1-\lambda) s_2) \in C \quad \text{for all } \lambda \in [0, 1]. \tag{3}$$

Every convex mapping $v: S \to F$ is also quasiconvex, because (2) and C being a convex cone imply for any $\lambda \in [0, 1]$

$$d := (1-\lambda)(v(s_1) - v(s_2)) \in C.$$

Therefore, from (1) we obtain

$$v(s_1) - v(\lambda s_1 + (1-\lambda) s_2) \in C + \{d\} \subset C.$$

John von Neumann's definition (see [18, p. 307]) using the level sets is sufficient for this type of quasiconvexity.

LEMMA 2.2. *Let E, F be real linear spaces, C a convex cone in F, S a nonempty convex subset of E and $v: S \to F$ a mapping. Then the mapping v is quasiconvex if and only if for all $\bar{s} \in S$ the sets*

$$L_{\bar{s}} := \{s \in S | s \neq \bar{s},\ v(\bar{s}) - v(s) \in C\} \tag{4}$$

contain $[s, \bar{s})$ whenever $s \in L_{\bar{s}}$.

The proof of Lemma 2.2 follows immediately from Definition 2.1.

2.1. *C*-quasiconvexity. In this subsection we extend the class of quasiconvex mappings considerably by the following definition.

DEFINITION 2.3. Let E, F be real linear spaces, S and C nonempty subsets of E and F, respectively, $v: S \to F$ a mapping and $\bar{s} \in S$ a given element. v is called C-

quasiconvex at $\bar{s}$ if the following holds: Whenever there is some $s \in S$ with

$$s \neq \bar{s} \quad \text{and} \quad v(\bar{s}) - v(s) \in C, \tag{5}$$

then there exists some $\tilde{s} \in S$ with

$$\tilde{s} \neq \bar{s}, \quad s_\lambda := \lambda \tilde{s} + (1-\lambda)\bar{s} \in S \quad \text{and} \quad v(\bar{s}) - v(s_\lambda) \in C \quad \text{for all } \lambda \in (0, 1]. \tag{6}$$

Example 2.4.

(a) Clearly, every quasiconvex mapping $v: S \to F$ is C-quasiconvex at all $\bar{s} \in S$ where C is the ordering cone in F.

(b) Let the mapping $v: \mathbb{R} \to \mathbb{R}^2$ be given by

$$v(\alpha) = (\alpha, \sin \alpha) \quad \text{for all } \alpha \in \mathbb{R}$$

where $\mathbb{R}^2$ is partially ordered in the componentwise sense. The mapping v is $\mathbb{R}^2_+$-quasiconvex at 0 but it is not quasiconvex (at 0).

The following lemma shows that C-quasiconvexity of v at $\bar{s}$ can also be characterized by a property of the level set $L_{\bar{s}}$ in (4).

LEMMA 2.5. *Let E, F be real linear spaces, S and C nonempty subsets of E and F, respectively, $v: S \to F$ a mapping and $\bar{s} \in S$ a given element. The mapping v is C-quasiconvex at $\bar{s} \in S$ if and only if the set $L_{\bar{s}}$ defined by* (4) *is empty or it contains a half-open line segment starting at $\bar{s}$, excluding $\bar{s}$.*

Proof. Rewrite (6) as

$$[\tilde{s}, \bar{s}) \subset L_{\bar{s}}$$

and the statement of the lemma is clear. □

As seen from Lemma 2.5 the relaxation of the requirement (3) to (6) by allowing $\tilde{s} \neq s_2$ extends the class of quasiconvex mappings considerably.

2.2. Differentiable C-quasiconvexity. For several applications it is necessary to investigate C-quasiconvex mappings which are differentiable in a certain sense. For such mappings it is reasonable to introduce an appropriate framework for differentiable C-quasiconvexity. Before we present this definition we introduce the following weak differentiability notion which extends the usual notion of Gâteaux variations to real linear spaces which are not equipped with any topology along the lines of Kirsch/Warth/Werner [10, p. 33].

DEFINITION 2.6. Let E, F be real linear spaces, S and A nonempty subsets of E and F, respectively, $v: S \to F$ a mapping and $\bar{s} \in S$ a given element. A mapping $v'(\bar{s}): S - \{\bar{s}\} \to F$ is called a *Gâteaux variation of v at $\bar{s}$ with respect to A* if the following holds: Whenever there is an element $s \in S$ with

$$s \neq \bar{s} \quad \text{and} \quad v'(\bar{s})(s - \bar{s}) \in A,$$

then there exists some $\bar{\tau} > 0$ with

$$s_\tau := \bar{s} + \tau(s - \bar{s}) \in S \quad \text{and} \quad \frac{1}{\tau}(v(s_\tau) - v(\bar{s})) \in A \quad \text{for all } \tau \in (0, \bar{\tau}].$$

Example 2.7. In addition, let F be a linear topological space and let $v'(\bar{s}): S - \{\bar{s}\} \to F$ represent the directional derivative, i.e. assume that for each $s \in S$ there exists $\bar{\tau} > 0$ with $[s, s_{\bar{\tau}}] \subset S$ and

$$v'(\bar{s})(s - \bar{s}) = \lim_{\tau \downarrow 0} \frac{1}{\tau}(v(\bar{s} + \tau(s - \bar{s})) - v(\bar{s})).$$

Then $v'(\bar{s})$ is a Gâteaux variation of v at $\bar{s}$ with respect to all open sets of F. Suppose $A \subset F$ is open and for some $s \in S$

$$w = v'(\bar{s})(s - \bar{s}) \in A,$$

then $A - \{w\}$ is an open 0_F-neighborhood and there exists $\bar{\tau} > 0$ with

$$\frac{1}{\tau}(v(s_\tau) - v(\bar{s})) - w \in A - \{w\} \quad \text{for all } \tau \in (0, \bar{\tau}],$$

which yields the desired relation. If v is an affine mapping, a Gâteaux variation exists with respect to all sets A.

Next we present the notion of differentiable C-quasiconvexity.

DEFINITION 2.8. Let E, F be real linear spaces, S a nonempty subset of E, C_1 and $C_2 \subset C_3$ nonempty subsets of F, $\bar{s} \in S$ a given element and $v: S \to F$ a mapping which has a Gâteaux variation at $\bar{s}$ with respect to C_3. v is called *differentiably C_1-C_2-quasiconvex at $\bar{s}$* if the following holds: Whenever there is some $s \in S$ with

$$s \neq \bar{s} \quad \text{and} \quad v(s) - v(\bar{s}) \in C_1 \tag{7}$$

then there exists some $\tilde{s} \in S$ with

$$\tilde{s} \neq \bar{s}, \quad [\tilde{s}, \bar{s}] \subset S \quad \text{and} \quad v'(\bar{s})(\tilde{s} - \bar{s}) \in C_2. \tag{8}$$

In the case of $C_1 = C_2 =: C$ the mapping v is simply called *differentiably C-quasiconvex at $\bar{s}$.*

Example 2.9. Let E, F be normed linear spaces, S a subset of E with a nonempty topological interior $\operatorname{int}(S)$, C a convex cone in F, $\bar{s} \in \operatorname{int}(S)$ a given element and $v: S \to F$ a mapping which is Fréchet-differentiable at $\bar{s}$. Then the mapping v is called *pseudoconvex at $\bar{s}$* (for similar definitions, compare Hartwig [5, p. 305] and Craven [4, p. 665] among others), if for all $s \in S$ the following holds:

$$v'(\bar{s})(s - \bar{s}) \in C \Rightarrow v(s) - v(\bar{s}) \in C.$$

This implication is equivalent to

$$v(s) - v(\bar{s}) \notin C \Rightarrow v'(\bar{s})(s - \bar{s}) \notin C.$$

Therefore, each mapping v which is pseudoconvex at $\bar{s}$ is also differentiably $(F \setminus C)$-quasiconvex at $\bar{s}$. This shows that the class of pseudoconvex mappings is contained in the larger class of differentiably $C_1 - C_2$-quasiconvex mappings.

Example 2.10. Let E, F be Banach spaces, S a nonempty convex subset of E, C a closed convex cone in F and $v: S \to F$ a Fréchet-differentiable mapping at a point $\bar{s} \in S$. Nehse [17, p. 484] defines v to be *strong pseudoconvex at $\bar{s}$ with respect to C* if there exists a functional $p_v: S \times S \to \{\delta \in \mathbb{R} : \delta > 0\}$ such that

$$p_v(s, \bar{s})(v(s) - v(\bar{s})) - v'(\bar{s})(s - \bar{s}) \in C \quad \text{for all } s \in S.$$

It is easy to verify that these mappings are also differentiably $(-C)$-quasiconvex and differentiably $(-\operatorname{cor}(C))$-quasiconvex provided $\operatorname{cor}(C) \neq \emptyset$.

With the next theorem we investigate some relations between C-quasiconvexity and differentiable C-quasiconvexity.

THEOREM 2.11. *Let E, F be real linear spaces, S a nonempty subset of E, $C \subset \hat{C}$ nonempty subsets of F such that $C \cup \{0_F\}$ is a cone, $\bar{s} \in S$ a given element and $v: S \to F$ a mapping.*

(a) *If v is $(-C)$-quasiconvex at $\bar{s}$ and has a Gâteaux variation at $\bar{s}$ with respect to $\hat{C}$ and $F \setminus C$, then v is differentiably C-quasiconvex at $\bar{s}$.*

(b) *If v is differentiably C-quasiconvex at $\bar{s}$ with a Gâteaux variation of v at $\bar{s}$ with respect to C, then v is $(-C)$-quasiconvex at $\bar{s}$.*

Proof. (a) Let some $s \in S$ be given with (7). Since v is assumed to be $(-C)$-quasiconvex at $\bar{s}$ there exists some $\tilde{s} \in S$ such that for all $\lambda \in (0, 1]$

$$\tilde{s} \neq \bar{s}, \quad s_\lambda = \lambda\tilde{s} + (1-\lambda)\bar{s} \in S \quad \text{and} \quad v(\bar{s}) - v(s_\lambda) \in -C. \tag{9}$$

Suppose that for all Gâteaux variations of v at $\bar{s}$ with respect to $\hat{C}$ and $F \setminus C$

$$v'(\bar{s})(\tilde{s} - \bar{s}) \notin C.$$

Then, from the definition of a Gâteaux variation with respect to $F \setminus C$ there exists some $\bar{\tau} > 0$ with

$$s_\tau := \bar{s} + \tau(\bar{s} - \tilde{s}) \quad \text{and} \quad \frac{1}{\tau}(v(s_\tau) - v(\bar{s})) \notin C \quad \text{for all } \tau \in (0, \bar{\tau}].$$

By assumption $C \cup \{0_F\}$ is a cone, and therefore we conclude

$$v(\bar{s}) - v(s_\tau) \notin -C \quad \text{for all } \tau \in (0, \bar{\tau}].$$

But this contradicts (9) and shows that for some Gâteaux variation of v at $\bar{s}$ with respect to $F \setminus C$ and $\hat{C}$

$$v'(\bar{s})(\tilde{s} - \bar{s}) \in C,$$

which shows that (7) implies (8) in Definition 2.8 with $C_1 = C_2 = C$ and $C_3 = \hat{C}$.

(b) Let $s \in S$ be given with $s \neq \bar{s}$ and $v(s) - v(\bar{s}) \in C$. Then differentiable C-quasiconvexity of v at $\bar{s}$ implies that there exists some $\tilde{s} \in S$ and a Gâteaux variation of v at $\bar{s}$ with respect to C with the property

$$\tilde{s} \neq \bar{s}, [\tilde{s}, \bar{s}] \subset S \quad \text{and} \quad v'(\bar{s})(\tilde{s} - \bar{s}) \in C.$$

Then by Definition 2.6 there exists some $\bar{\tau} > 0$ with

$$s_\tau := \bar{s} + \tau(\tilde{s} - \bar{s}) \in S \quad \text{and} \quad \frac{1}{\tau}(v(s_\tau) - v(\bar{s})) \in C \quad \text{for all } \tau \in (0, \bar{\tau}].$$

Observing that $C \cup \{0_F\}$ is a cone, we arrive at (6) and the proof of the $(-C)$-quasiconvexity at $\bar{s}$ is complete. □

If one considers Gâteaux variations with respect to algebraically open sets, in the previous theorem under (a) and (b), one should assume that C is algebraically closed and that $F \setminus C$ and $\hat{C}$ are algebraically open, respectively.

3. Application to vector optimization problems. This section is aimed to show the applicability and the usefulness of the notions of generalized convexity introduced in § 2 to vector optimization problems. The generalized Karush-Kuhn-Tucker multiplier rule with a real-valued or vector-valued Lagrangian is examined and we prove that this multiplier rule is a sufficient optimality condition for a substitute problem if and only if the corresponding mappings are generalized quasiconvex. We show also that the local optima are global optima if and only if the objective mapping is in some sense generalized convex. In the last subsection we discuss some examples.

3.1. Problem formulation. In this subsection we consider the vector optimization problem

$$\text{“}\min_{x \in S}\text{”} f(x)$$

where f is a mapping whose domain is a subset of a real linear space and whose range is a subset of a partially ordered real linear space. S is a subset of the domain of f. The set S is called the *constraint set* or *feasible set* and f is called the *objective mapping.*

There are several possibilities to define minima of f on S. We restrict ourselves to the following two optimality notions.

DEFINITION 3.1. Let X, Y be real linear spaces where Y is partially ordered by a convex cone C_Y. Furthermore, let S be a nonempty subset of X and $f: S \to Y$ be a given mapping.

(a) An element $\bar{x} \in S$ is called a *minimal point of f on S,* if

$$(\{f(\bar{x})\} - C_Y) \cap f(S) = \{f(\bar{x})\}.$$

(b) In addition, let $\operatorname{cor}(C_Y) \neq \emptyset$. An element $\bar{x} \in S$ is called a *weakly minimal point of f on S,* if

$$(\{f(\bar{x})\} - \operatorname{cor}(C_Y)) \cap f(S) = \emptyset.$$

The following lemma gives a relation between the two types of optimal points.

LEMMA 3.2. *Let X, Y be real linear spaces, $C_Y \subsetneq Y$ a convex cone with* $\operatorname{cor}(C_Y) \neq \emptyset$, *$S$ a nonempty subset of X and $f: S \to Y$ a mapping. Then each minimal point of f on S is a weakly minimal point of f on S.*

For a special case, a proof of this lemma and an example which shows that the reverse implication does not hold can be found in Lin [14, pp. 49, 47]. From a theoretical point of view the notion of weak minimality is more elegant than the notion of minimality but in the applications the notion of minimality is more desirable.

3.2. Sufficiency of the generalized Karush–Kuhn–Tucker multiplier rule. In this subsection vector optimization problems are considered where the set S of feasible points is given by inequality and equality constraints. In connection with a regularity condition necessary conditions can be established, the so-called generalized Karush-Kuhn-Tucker conditions, which generalize the well-known Lagrange multiplier rule (for a discussion of these necessary conditions for vector optimization problems we refer to Kuhn/Tucker [13], Hurwicz [7], Borwein [2], Vogel [28], Kirsch/Warth/Werner [10], Sachs [24], Hartwig [5], Oettli [20], Borwein [3], Craven [4], among others). It is our aim to give a characterization of these conditions in terms of the definition of generalized convexity as previously given. The generalized Karush-Kuhn-Tucker conditions are investigated in a vector-valued and in a real-valued form.

The vector optimization problem which is under investigation is given as follows:

$$\begin{aligned} &\text{"min"}\ f(x) \\ &\text{subject to} - g(x) \in C_{Z_1}, \\ &\qquad h(x) = 0_{Z_2}, \\ &\qquad x \in S_0. \end{aligned} \tag{10}$$

For this problem we assume the following:

$$\begin{aligned} &\text{Let } X, Y, Z_1, Z_2 \text{ be real linear spaces;} \\ &\text{let } C_Y \subset Y, C_{Z_1} \subset Z_1 \text{ and } C_{Z_2} \subset Z_2 \text{ be convex cones where} \\ &\qquad \operatorname{cor}(C_Y) \neq \emptyset \text{ and } C_{Z_2} \text{ is pointed;} \\ &\text{let } S_0 \text{ be a nonempty subset of } X; \\ &\text{let } f: S_0 \to Y,\ g: S_0 \to Z_1 \text{ and } h: S_0 \to Z_2 \text{ be mappings,} \\ &\text{let the feasible set } S \text{ be defined by} \\ &\qquad S := \{x \in S_0 \mid -g(x) \in C_{Z_1},\ h(x) = 0_{Z_2}\}. \end{aligned} \tag{11}$$

THEOREM 3.3. *Let the vector optimization problem* (10) *with* (11) *be given and suppose that for some* $\bar{x} \in S$ *there exist sets* G_i, $i = 0, 1, 2$ *with* $\operatorname{cor}(C_Y) \subset G_0 \subset Y$, $C_{Z_1} + \operatorname{cone}(g(\bar{x})) \subset G_1 \subset Z_1$, $(C_{Z_2} \cup (-C_{Z_2})) \subset G_2 \subset Z_2$, *such that the mappings* f, g, h *have Gâteaux variations at* $\bar{x}$ *with respect to* G_0, G_1, *and* G_2, *respectively. Assume that there exist some*

$$t \in C_{Y'} \setminus \{0_{Y'}\}, \quad u \in C_{Z_1'}, \quad v \in Z_2' \tag{12}$$

with

$$(t \circ f'(\bar{x}) + u \circ g'(\bar{x}) + v \circ h'(\bar{x}))(x - \bar{x}) \geqq 0 \quad \text{for all } x \in S_0 \tag{13}$$

and

$$(u \circ g)(\bar{x}) = 0. \tag{14}$$

Then $\bar{x}$ *is a weakly minimal point of* f *on*

$$\bar{S} := \{x \in S_0 \mid -g(x) \in C_{Z_1} + \operatorname{cone}(g(\bar{x})),\ h(x) = 0_{Z_2}\}$$

if and only if

$$(f, g, h, h) : S_0^4 \to Y \times Z_1 \times Z_2 \times Z_2 \tag{15}$$

is differentiably $(-C)$*-quasiconvex at* $\bar{x}$ *with*

$$C := \operatorname{cor}(C_Y) \times (C_{Z_1} + \operatorname{cone}(g(\bar{x}))) \times C_{Z_2} \times (-C_{Z_2}). \tag{16}$$

Proof. Assume that the multiplier rule (12)-(14) holds at some $\bar{x} \in S$. Then we assert that

$$(f'(\bar{x})(x - \bar{x}), g'(\bar{x})(x - \bar{x}), h'(\bar{x})(x - \bar{x}), h'(\bar{x})(x - \bar{x})) \notin -C \quad \text{for all } x \in S_0. \tag{17}$$

For the proof of this assertion assume that there exists some $x \in S_0$ with

$$f'(\bar{x})(x - \bar{x}) \in -\operatorname{cor}(C_Y),$$
$$g'(\bar{x})(x - \bar{x}) \in -C_{Z_1} - \operatorname{cone}(g(\bar{x})),$$
$$h'(\bar{x})(x - \bar{x}) \in -C_{Z_2},$$
$$h'(\bar{x})(x - \bar{x}) \in C_{Z_2}.$$

Because C_{Z_2} is a pointed cone we obtain

$$h'(\bar{x})(x - \bar{x}) \in C_{Z_2} \cap (-C_{Z_2}) = \{0_{Z_2}\}$$

and together with (12) we conclude for some $\alpha \geqq 0$

$$(t \circ f'(\bar{x}) + u \circ g'(\bar{x}) + v \circ h'(\bar{x}))(x - \bar{x}) < -\alpha(u \circ g)(\bar{x}).$$

But this inequality contradicts (13) and (14). Hence (17) holds and with the generalized quasiconvexity requirement we deduce that for all $x \in S_0$

$$(f(x) - f(\bar{x}), g(x) - g(\bar{x}), h(x) - h(\bar{x}), h(x) - h(\bar{x})) \notin -C. \tag{18}$$

Condition (18) means that there is no $x \in S_0$ with

$$f(x) \in \{f(\bar{x})\} - \operatorname{cor}(C_Y),$$
$$-g(x) \in \{-g(\bar{x})\} + C_{Z_1} + \operatorname{cone}(g(\bar{x})) = C_{Z_1} + \operatorname{cone}(g(\bar{x})),$$
$$h(x) \in C_{Z_2} \cap (-C_{Z_2}) = \{0_{Z_2}\}.$$

So (18) is equivalent with the statement that

$$(\{f(\bar{x})\} - \operatorname{cor}(C_Y)) \cap f(\bar{S}) = \emptyset,$$

i.e. $\bar{x}$ is a weakly minimal point of f on $\bar{S}$.

Conversely, if in addition to (12)-(14) $\bar{x}$ is a weakly minimal point of f on $\bar{S}$, then (18) holds and the differentiable C-quasiconvexity of (15) at $\bar{x}$ is trivial. □

In the previous theorem we showed the equivalence of the generalized quasiconvexity with the sufficiency of the generalized Karush-Kuhn-Tucker conditions for optimality of a substitute problem where S is replaced by $\bar{S}$. For the original problem the following conclusion holds.

COROLLARY 3.4. *Let the assumptions of Theorem* 3.3 *be satisfied and let the mapping* (15) *be differentiably* $(-C)$*-quasiconvex at* $\bar{x} \in S$ *with* C *given by* (16); *then* $\bar{x}$ *is a weakly minimal point of* f *on* S.

Proof. By Theorem 3.3 $\bar{x} \in S$ is a weakly minimal point of f on $\bar{S}$, i.e.

$$(\{f(\bar{x})\} - \operatorname{cor}(C_Y)) \cap f(\bar{S}) = \emptyset.$$

But $S \subset \bar{S}$ implies that $\bar{x}$ is also a weakly minimal point of f on S. □

Corollary 3.4 extends results of Vogel [28, p. 100], Hartwig [5, p. 313-314] (for another optimality notion) and Craven [4, p. 666-667]. In Theorem 3.3 a real-valued Lagrangian $t \circ f + u \circ g + v \circ h$ is considered implicitly. For a vector-valued Lagrangian $f + L_1 \circ g + L_2 \circ h$ where L_1 and L_2 are appropriate linear mappings we obtain a similar result as in Theorem 3.3 without using a separate approach (as in Craven [4]). Under a convex-likeness assumption there is no difference if we use a real-valued or a vector-valued Lagrangian. This result is formulated in

THEOREM 3.5. *Let the vector optimization problem* (10) *with* (11) *be given. For some* $\bar{x} \in S$ *we assume that* $f'(\bar{x})\colon S_0 - \{\bar{x}\} \to Y$, $g'(\bar{x})\colon S_0 - \{\bar{x}\} \to Z_1$ *and* $h'(\bar{x})\colon S_0 - \{\bar{x}\} \to Z_2$ *are arbitrary mappings. Then for the two statements,*

(19) *there exist* $t \in C_{Y'} \setminus \{0_{Y'}\}$, $u \in C_{Z_1'}$, *and* $v \in Z_2'$
with the properties $(t \circ f'(\bar{x}) + u \circ g'(\bar{x}) + v \circ h'(\bar{x}))(x - \bar{x}) \geqq 0$ *for all* $x \in S_0$
and $(u \circ g)(\bar{x}) = 0$,

and

(20) *there exist a linear mapping* $L_1\colon Z_1 \to Y$ *with*
$L_1(C_{Z_1}) \subset (\operatorname{cor}(C_Y) \cup \{0_Y\})$ *and a linear mapping*
$L_2\colon Z_2 \to Y$ *with the properties*
$(f'(\bar{x}) + L_1 \circ g'(\bar{x}) + L_2 \circ h'(\bar{x}))(x - \bar{x}) \notin -\operatorname{cor}(C_Y)$ *for all* $x \in S_0$
and $(L_1 \circ g)(\bar{x}) = 0_Y$,

the following holds:

(a) *The statement* (19) *implies the statement* (20).

(b) *If the statement* (20) *holds and the mapping* $f'(\bar{x}) + L_1 \circ g'(\bar{x}) + L_2 \circ h'(\bar{x})$ *is convex-like, then the statement* (19) *is true.*

Proof.

(a) We assume that the statement (19) is true. Because of $t \in C_{Y'} \setminus \{0_{Y'}\}$ and $\operatorname{cor}(C_Y) \neq \emptyset$ there exists a $\tilde{y} \in \operatorname{cor}(C_Y)$ with $t(\tilde{y}) = 1$. Then, following an idea due to Borwein [2, p. 62], we define the mappings $L_1\colon Z_1 \to Y$ and $L_2\colon Z_2 \to Y$ by

$$L_1(z_1) = u(z_1)\tilde{y} \quad \text{for all } z_1 \in Z_1 \tag{21}$$

and

$$L_2(z_2) = v(z_2)\tilde{y} \quad \text{for all } z_2 \in Z_2.$$

Obviously, L_1 and L_2 are linear mappings, and we have

$$L_1(C_{Z_1}) \subset \operatorname{cor}(C_Y) \cup \{0_Y\}.$$

Furthermore, we obtain $t \circ L_1 = u$ and $t \circ L_2 = v$. Consequently, the inequality in the statement (19) can be rewritten as

$$(t \circ (f'(\bar{x}) + L_1 \circ g'(\bar{x}) + L_2 \circ h'(\bar{x})))(x - \bar{x}) \geqq 0 \quad \text{for all } x \in S_0.$$

Then we conclude with a scalarization result (e.g., compare [9, Cor. 2.3(c)])

$$(f'(\bar{x}) + L_1 \circ g'(\bar{x}) + L_2 \circ h'(\bar{x}))(x - \bar{x}) \notin -\operatorname{cor}(C_Y) \quad \text{for all } x \in S_0.$$

Finally, with the equality (21) we get

$$(L_1 \circ g)(\bar{x}) = (u \circ g)(\bar{x})\tilde{y} = 0_Y.$$

So the statement (20) is true.

(b) Now we assume that the statement (20) is true. Then we have

$$(f'(\bar{x}) + L_1 \circ g'(\bar{x}) + L_2 \circ h'(\bar{x}))(x - \bar{x}) \notin -\operatorname{cor}(C_Y) \quad \text{for all } x \in S_0,$$

and if we notice that $\operatorname{cor}(C_Y) = \operatorname{cor}(C_Y) + C_Y$, this implies

$$(S + C_Y) \cap (-\operatorname{cor}(C_Y)) = \varnothing$$

where S is defined as

$$S := \{(f'(\bar{x}) + L_1 \circ g'(\bar{x}) + L_2 \circ h'(\bar{x}))(x - \bar{x}) | x \in S_0\}.$$

Since the mapping $f'(\bar{x}) + L_1 \circ g'(\bar{x}) + L_2 \circ h'(\bar{x})$ is assumed to be convex-like, the set $S + C_Y$ is convex. By a known separation theorem (e.g., see Holmes [7, § 4.B]) there exists a linear functional $t \in C_{Y'} \backslash \{0_{Y'}\}$ with

$$(t \circ f'(\bar{x}) + t \circ L_1 \circ g'(\bar{x}) + t \circ L_2 \circ h'(\bar{x}))(x - \bar{x}) \geqq 0 \quad \text{for all } x \in S_0.$$

If we define $u := t \circ L_1$ and $v := t \circ L_2$ we obtain

$$(t \circ f'(\bar{x}) + u \circ g'(\bar{x}) + v \circ h'(\bar{x}))(x - \bar{x}) \geqq 0 \quad \text{for all } x \in S_0$$

and

$$(u \circ g)(\bar{x}) = (t \circ L_1 \circ g)(\bar{x}) = 0.$$

Furthermore, for any $z_1 \in C_{Z_1}$ we obtain

$$u(z_1) = (t \circ L_1)(z_1) \geqq 0$$

which implies $u \in C_{Z_1'}$. This completes the proof. □

Although we use the notation of a derivative, Theorem 3.5 holds for arbitrary mappings $f'(\bar{x})$, $g'(\bar{x})$ and $h'(\bar{x})$. The "multiplier" L_1 in (20) is (in a stronger sense) a monotone mapping.

If the generalized quasiconvexity assumption in Theorem 3.3 is strengthened, then a similar theorem holds for minimal points of f on $\bar{S}$.

THEOREM 3.6. *Let all the assumptions of Theorem* 3.3 *be satisfied. Then* $\bar{x}$ *is a minimal point of* f *on* $\bar{S}$ *if and only if the composite mapping* (f, g, h, h) *is differentiably* $(-C_1)-(-C_2)$*-quasiconvex at* $\bar{x}$ *with*

$$C_1 := (C_Y \backslash \{0_Y\}) \times (C_{Z_1} + \operatorname{cone}(g(\bar{x}))) \times C_{Z_2} \times (-C_{Z_2})$$

and

$$C_2 := \operatorname{cor}(C_Y) \times (C_{Z_1} + \operatorname{cone}(g(\bar{x}))) \times C_{Z_2} \times (-C_{Z_2}).$$

The proof of this theorem is almost identical to the one of Theorem 3.3 and it is therefore omitted. A result which is similar to that of Corollary 3.4 can also be obtained.

3.3. Local and global minima. It is well-known that a real-valued strictly quasiconvex function on a convex subset of $\mathbb{R}^n$ has the property that all its local minima are also global minima (e.g., see Mangasarian [16, p. 139]). We show in this subsection that this fact extends to hold under C-quasiconvexity and it is even characteristic for it.

DEFINITION 3.7. Let X, Y be real linear spaces, S a nonempty subset of X, C_Y a convex cone in Y and $f: S \to Y$ a mapping.

(a) An element $\bar{x} \in S$ is called a *local minimal point of f on S*, if there is a set $U \subset X$ with $\bar{x} \in \operatorname{cor}(U)$ such that $\bar{x}$ is a minimal point of f on $S \cap \operatorname{cor}(U)$.

(b) In addition, let $\operatorname{cor}(C_Y) \neq \emptyset$. An element $\bar{x} \in S$ is called a *local weakly minimal point of f on S*, if there is a set $U \subset X$ with $\bar{x} \in \operatorname{cor}(U)$ such that $\bar{x}$ is a weakly minimal point of f on $S \cap \operatorname{cor}(U)$.

The following two theorems state a necessary and sufficient condition under which local minima are even global minima.

THEOREM 3.8. *Let X, Y be real linear spaces, S a nonempty subset of X, C_Y a convex cone in Y with $C_Y \neq \{0_Y\}$ and $f: S \to Y$ a mapping. Let $\bar{x} \in S$ be a local minimal point of f on S. Then $\bar{x}$ is a (global) minimal point of f on S if and only if f is $(C_Y \setminus \{0_Y\})$-quasiconvex at $\bar{x}$.*

Proof. Suppose that $\bar{x} \in S$ is a local minimal point of f on S. If $\bar{x}$ is not a minimal point of f on S, then there exists some $x \in S$ with

$$f(\bar{x}) - f(x) \in C_Y \setminus \{0_Y\}.$$

Assume f is $(C_Y \setminus \{0_Y\})$-quasiconvex at $\bar{x}$; then there exists some $\tilde{x} \in S$ with

$$x_\lambda := \lambda \tilde{x} + (1-\lambda)\bar{x} \in S, \ \tilde{x} \neq \bar{x}$$

and

$$f(\bar{x}) - f(x_\lambda) \in C_Y \setminus \{0_Y\} \quad \text{for all } \lambda \in (0, 1]. \tag{22}$$

Since $\bar{x} \in \operatorname{cor}(U)$ there exists some $\bar{\lambda} \in (0, 1]$ with

$$x_\lambda \in \operatorname{cor}(U) \quad \text{for all } \lambda \in (0, \bar{\lambda}]. \tag{23}$$

(22) and (23) together contradict the assumption that $\bar{x}$ is a local minimal point of f on S. On the other hand if $\bar{x}$ is a minimal point of f on S, then there is no $x \in S$ with

$$f(\bar{x}) - f(x) \in C_Y \setminus \{0_Y\}$$

and the $(C_Y \setminus \{0_Y\})$-quasiconvexity of f at $\bar{x}$ holds trivially. □

The following theorem can be proven similarly.

THEOREM 3.9. *Let X, Y be real linear spaces, S a nonempty subset of X, C_Y a convex cone in Y with $\operatorname{cor}(C_Y) \neq \emptyset$ and $f: S \to Y$ a mapping. Let $\bar{x} \in S$ be a local weakly minimal point of f on S. Then $\bar{x}$ is a (global) weakly minimal point of f on S if and only if f is $\operatorname{cor}(C_Y)$-quasiconvex at $\bar{x}$.*

Zang and Avriel [29] characterized functions on $\mathbb{R}^n$ whose local minima are also global by a lower semicontinuity property of the corresponding level sets.

3.4. Examples. In this last section we consider two special problems, namely a nonlinear multi-objective programming problem and a vector approximation problem.

For both problems we present conditions under which the generalized quasiconvexity assumption in Theorem 3.3 is fulfilled.

The vector optimization problem (10) reduces to a nonlinear multi-objective programming problem, if we set

$$
\begin{aligned}
&X=\mathbb{R}^n, \quad Y=\mathbb{R}^m, \quad Z_1=\mathbb{R}^k, \quad Z_2=\mathbb{R}^l,\\
&C_Y=\{y\in\mathbb{R}^m \mid y_i\geqq 0 \text{ for all } i=1,\cdots,m\},\\
&C_{Z_1}=\{z\in\mathbb{R}^k \mid z_i\geqq 0 \text{ for all } i=1,\cdots,k\},\\
&C_{Z_2}=\{z\in\mathbb{R}^l \mid z_i\geqq 0 \text{ for all } i=1,\cdots,l\}.
\end{aligned}
\tag{24}
$$

For this special case we obtain

LEMMA 3.10. *Let the nonlinear multi-objective programming problem* (10) *with* (11) *and the special assumption* (24) *be given and assume that at some* $\bar{x}\in S$ *the vector functions* f, g, h *have partial derivatives at* $\bar{x}$. *If the functions* $f_1,\cdots,f_m$ *are pseudoconvex at* $\bar{x}$ *and the functions* $h_1,\cdots,h_l,-h_1,\cdots,-h_l$ *and* g_i *for all* $i\in I(\bar{x})$ *with*

$$I(\bar{x}):=\{i\in\{1,\cdots,k\} \mid g_i(\bar{x})=0\}$$

are quasiconvex at $\bar{x}$, *then the composite vector function* (f, g, h, h) *is differentiably* $(-C)$*-quasiconvex at* $\bar{x}$ *with* C *defined by* (16).

Proof. Let some $x=s\in S$ be given with (7), i.e. $\bar{x}\neq x$ and

$$
\begin{aligned}
&f_i(\bar{x})-f_i(x)>0 \quad \text{for all } i=1,\cdots,m,\\
&g_i(\bar{x})-g_i(x)\geqq \alpha g_i(\bar{x}) \quad \text{for all } i=1,\cdots,k \text{ and some } \alpha\geqq 0,\\
&h_i(\bar{x})-h_i(x)=0 \quad \text{for all } i=1,\cdots,l.
\end{aligned}
\tag{25}
$$

The inequality (25) implies

$$g_i(\bar{x})-g_i(x)\geqq 0 \quad \text{for all } i\in I(\bar{x}).$$

Using the characterization of differentiable quasiconvex functions and the definition of pseudoconvex functions (e.g., compare Mangasarian [16, ch. 9]) the previous inequalities imply

$$
\begin{aligned}
&f_i'(\bar{x})(\bar{x}-x)>0 \quad \text{for all } i=1,\cdots,m,\\
&g_i'(\bar{x})(\bar{x}-x)\geqq 0 \quad \text{for all } i\in I(\bar{x}),\\
&h_i'(\bar{x})(\bar{x}-x)=0 \quad \text{for all } i=1,\cdots,l.
\end{aligned}
$$

Since $g_i(\bar{x})<0$ for all $i\notin I(\bar{x})$ there is an $\alpha>0$ with

$$g_i'(\bar{x})(\bar{x}-x)\geqq \alpha g_i(\bar{x}) \quad \text{for all } i\in\{1,\cdots,m\}.$$

Hence $\tilde{s}=x$ satisfies (8). □

Lemma 3.10 shows in particular that the convexity type conditions are only imposed on the active constraints. This is for example not the case if one applies the condition of a function being "almost pseudoconcave" in Krabs [12, p. 172] to this standard problem of optimization. However, the almost pseudoconcavity was introduced in order to obtain the sufficiency of necessary optimality conditions for problems in Chebyshev approximation.

Next we consider a special Chebyshev vector approximation problem under the following assumption:

Let A be a nonempty convex subset of $\mathbb{R}^n$;
let A_0 be an open superset of A;
let Q be a compact space with at least $n+2$ elements;
(26) let $C(Q)$ denote the real linear space of real-valued continuous functions on Q equipped with the maximum norm $\|\cdot\|_\infty$;
let $F_1, \cdots, F_m: A_0 \to C(Q)$ be mappings which are Fréchet differentiable on A_0;
let $z_1, \cdots, z_m \in C(Q)$ be given functions.

Then the Chebyshev vector approximation problem is formalized as

$$\text{(27)} \qquad \text{"min"}_{a \in A} \begin{pmatrix} \|F_1(a) - z_1\|_\infty \\ \vdots \\ \|F_m(a) - z_m\|_\infty \end{pmatrix}$$

where the partial ordering in $\mathbb{R}^m$ is understood in the natural sense. For our investigations we replace this problem by the following:

$$\text{"min"} (\gamma_1, \cdots, \gamma_m)^T$$

subject to

$$\text{(28)} \qquad \left.\begin{aligned} F_i(a)(t) - z_i(t) - \gamma_i &\leqq 0 \\ -F_i(a)(t) + z_i(t) - \gamma_i &\leqq 0 \end{aligned}\right\} \text{for all } i = 1, \cdots, m \text{ and all } t \in Q$$

$$a \in A.$$

The vector optimization problem (28) is a special type of (10) without equality constraints. Then (11) reduces to:

$$X = \mathbb{R}^{n+m}, \quad Y = \mathbb{R}^m, \quad Z_1 = C(Q)^{2m}, \quad S_0 = A \times \mathbb{R}^m,$$
$$C_Y = \{y \mid y_i \geqq 0 \text{ for all } i = 1, \cdots, m\},$$
$$C_{Z_1} = C_{C(Q)}^{2m} \text{ with } C_{C(Q)} = \{u \in C(Q) \mid u(t) \geqq 0 \text{ for all } t \in Q\},$$
$$f: A \times \mathbb{R}^m \to Y \text{ with } f(x) = (\gamma_1, \cdots, \gamma_m)^T \text{ for all } x = (a, \gamma_1, \cdots, \gamma_m) \in A \times \mathbb{R}^m,$$
$$g: A \times \mathbb{R}^m \to Z_1 \text{ with}$$

$$\text{(29)} \qquad g(x) = \begin{pmatrix} F_1(a) - z_1 - \gamma_1 \mathbb{1} \\ -F_1(a) + z_1 - \gamma_1 \mathbb{1} \\ \vdots \\ F_m(a) - z_m - \gamma_m \mathbb{1} \\ -F_m(a) + z_m - \gamma_m \mathbb{1} \end{pmatrix} \text{ for all } x = (a, \gamma_1, \cdots, \gamma_m) \in A \times \mathbb{R}^m$$

where $\mathbb{1} \in C(Q)$ denotes the function which equals 1.

The equivalence of both problems (27) and (28) can be derived from the following lemma.

LEMMA 3.11. *Let X, Y be real linear spaces, C_Y a convex cone in Y with* cor $(C_Y) \neq \emptyset$, *S a nonempty subset of X and $f: S \to Y$ a mapping. In the space $X \times Y$ we define $S_1 = \{(x, y) \in S \times Y: y \in \{f(x)\} + C_Y\}$ and $\phi: S_1 \to Y$ by $\phi(x, y) = y$ for all $(x, y) \in S_1$. Then*

the problems

$$\text{"min"}_{x\in S} f(x) \quad \textit{and} \quad \text{"min"}_{(x,y)\in S_1} \phi(x, y)$$

are equivalent in the following sense:

If $\bar{x} \in S$ is a weakly minimal point of f on S, then $(\bar{x}, f(\bar{x}))$ is a weakly minimal point of ϕ on S_1. On the other hand, suppose $(\bar{x}, \bar{y}) \in S_1$ is a weakly minimal point of ϕ on S_1; then $\bar{x} \in S$ is a weakly minimal point of f on S.

The proof of this lemma is a consequence from the equality $C_Y = C_Y + \operatorname{cor}(C_Y)$ and the definition of the problems.

The same statement on the equivalence of minimal points holds if the condition $\operatorname{cor}(C_Y) \neq \emptyset$ is replaced by C_Y being pointed.

For the real-valued case (i.e., $m = 1$) Krabs [11] gave a condition on F_1 which assures the sufficiency of necessary conditions for the previous problem. It is also sufficient for the almost pseudo-concavity. We show that an extension of this condition, a representation condition, as well implies the differentiable C-quasiconvexity of the mapping (15) at each $\bar{x} \in S$ with C given by (16).

LEMMA 3.12. *Let the vector optimization problem* (28) *with the assumptions* (26) *and* (29) *be given. Let $(F_1, \cdots, F_m)$ satisfy the representation condition, i.e., for every $a_1, a_2 \in A$ there exist positive functions $\psi_1(a_1, a_2), \cdots, \psi_m(a_1, a_2) \in C(Q)$ and some $\tilde{a} \in A$ with*

$$F_i(a_1) - F_i(a_2) = \psi_i(a_1, a_2) F_i'(a_2)(\tilde{a} - a_2) \quad \textit{for all } i = 1, \cdots, m. \tag{30}$$

Then the composite mapping (f, g) is differentiably $(-C)$-quasiconvex at each $\bar{x} \in S$ with $C := \operatorname{cor}(C_Y) \times (C_{Z_1} + \operatorname{cone}(g(\bar{x})))$.

Proof. Let some $(\bar{a}, \bar{\gamma}_1, \cdots, \bar{\gamma}_m), (a, \gamma_1, \cdots, \gamma_m) \in S$ be given with

$$\bar{\gamma}_i - \gamma_i > 0 \quad \text{for all } i = 1, \cdots, m,$$

$$F_i(\bar{a}) - \bar{\gamma}_i - F_i(a) + \gamma_i \mathbb{1} \geqq \alpha(F_i(\bar{a}) - z_i - \bar{\gamma}_i \mathbb{1}) \tag{31}$$

$$-F_i(\bar{a}) - \bar{\gamma}_i + F_i(a) + \gamma_i \mathbb{1} \geqq \alpha(-F_i(\bar{a}) + z_i - \bar{\gamma}_i \mathbb{1}) \tag{32}$$

for all $i = 1, \cdots, m$ and some $\alpha \geqq 0$,

Then there exist positive functions $\psi_1(a, \bar{a}), \cdots, \psi_m(a, \bar{a}) \in C(Q)$ and some $\bar{a} \in A$ with (30). Furthermore there exist positive real numbers $\alpha_1, \cdots, \alpha_m, \beta_1, \cdots, \beta_m$ with

$$0 < \alpha_i \leqq \psi_i(a, \bar{a})(t) \leqq \beta_i \quad \text{for all } i = 1, \cdots, m \text{ and all } t \in Q,$$

and we define

$$\tilde{\alpha} := \frac{\alpha}{\min\{\alpha_1, \cdots, \alpha_m\}}$$

and

$$\tilde{\gamma}_i := \bar{\gamma}_i - \frac{1}{\beta_i}(\bar{\gamma}_i - \gamma_i) < \bar{\gamma}_i \quad \text{for all } i = 1, \cdots, m.$$

Then (31) implies with (30) and the feasibility of $(\bar{a}, \bar{\gamma}_1, \cdots, \bar{\gamma}_m)$

$$\begin{aligned} -F_i'(\bar{a})(\tilde{a} - \bar{a}) &= F_i'(\bar{a})(\bar{a} - \tilde{a}) \\ &= \frac{1}{\psi_i(a, \bar{a})}(F_i(\bar{a}) - F_i(a)) \\ &\geqq \frac{\alpha}{\alpha_i}(F_i(\bar{a}) - z_i - \bar{\gamma}_i \mathbb{1}) + \frac{\bar{\gamma}_i - \gamma_i}{\beta_i} \\ &\geqq \tilde{\alpha}(F_i(\bar{a}) - z_i - \bar{\gamma}_i \mathbb{1}) + (\bar{\gamma}_i - \tilde{\gamma}_i) \quad \text{for all } i = 1, \cdots, m. \end{aligned}$$

Similarly (32) implies

$$F_i'(\bar{a})(\tilde{a}-\bar{a}) = \frac{1}{\psi_i(a,\bar{a})}(F_i(a)-F_i(\bar{a}))$$

$$\geqq \frac{\alpha}{\alpha_i}(-F_i(\bar{a})+z_i-\gamma_i\mathbb{1})+\frac{\bar{\gamma}_i-\gamma_i}{\beta_i}$$

$$\geqq \tilde{\alpha}(-F_i(\bar{a})+z_i-\gamma_i\mathbb{1})+(\bar{\gamma}_i-\tilde{\gamma}_i) \quad \text{for all } i=1,\cdots,m.$$

Hence (8) holds with $\tilde{s}=(\tilde{a},\tilde{\gamma})$. □

The representation condition (30) is satisfied for rational approximating families: Let functions $p_j^i \in C(Q)$, $i=1,\cdots,m$, $j=1,\cdots,n$ be given and define for some $n_i \in \{1,2,\cdots,n-1\}$, $i=1,\cdots,m$

$$F_i(a) = \frac{\sum_{j=1}^{n_i}\alpha_j p_j^i}{\sum_{j=n_i+1}^{n}\alpha_j p_j^i} \quad \text{for } a=(\alpha_1,\cdots,\alpha_n)\in\mathbb{R}^n,$$

$$A = \left\{ a\in\mathbb{R}^n : \sum_{j=n_i+1}^{n}\alpha_j p_j^i(t)>0 \quad \text{for all } t\in Q \right\}.$$

An easy computation shows that (30) holds with

$$\psi_i(a_1,a_2) = \sum_{j=n_i+1}^{n}\alpha_j^{(1)}p_j^i \Big/ \sum_{j=n_i+1}^{n}\alpha_j^{(2)}p_j^i,$$

$$a_k = (\alpha_1^{(k)},\cdots,\alpha_n^{(k)}), \qquad k=1,2.$$

For further discussion of these types of condition for the case $m=1$ see Krabs [11].

For real-valued Chebyshev approximation problems the multiplier rule (12)-(14) can be formulated more precisely taking into account the special structure of approximation problems. This is extended to the Chebyshev vector approximation problem in the following lemma.

LEMMA 3.13. *Let the assumptions* (26) *and* (29) *be satisfied, and let a vector* $(\bar{a},\bar{\gamma})\in A\times\mathbb{R}^m$ *with* $\bar{\gamma}_i = \|F_i(\bar{a})-z_i\|_\infty$, $i=1,\cdots,m$, *be given. For each* $i=1,\cdots,m$ *let the Fréchet-derivative of* F_i *at* $\bar{a}$ *be given by*

$$F_i'(\bar{a})(a) = \sum_{k=1}^{n} a_k v_{ik} \quad \textit{for all } a\in A$$

with certain functions $v_{ik}\in C(Q)$. *The vector* $(\bar{a},\bar{\gamma})$ *satisfies the multiplier rule* (12)-(14) *for problem* (28) *if and only if there exist nonnegative numbers* $\tau_1,\cdots,\tau_m$ *where at least one* τ_i *is nonzero with the following property: For each* $i=1,\cdots,m$ *with* $\tau_i>0$ *there exist* p_i *points* $t_{ij}\in E_i(\bar{a})$ *with*

$$1\leqq p_i \leqq \dim \operatorname{span}\{v_{i1},\cdots,v_{in},\mathbb{1},F_i(\bar{a})-z_i\}\leqq n+2,$$

$$E_i(\bar{a}) := \{t\in Q \,|\, |(F_i(\bar{a})-z_i)(t)| = \|F_i(\bar{a})-z_i\|_\infty\}$$

and there are $\lambda_{ij}\in\mathbb{R}$, $j=1,\cdots,p_i$, *such that*

$$\sum_{j=1}^{p_i}|\lambda_{ij}| = 1, \tag{33}$$

$$\sum_{k=1}^{n}(a_k-\bar{a}_k)\sum_{\substack{i=1\\ \tau_i>0}}^{m}\tau_i\sum_{j=1}^{p_i}\lambda_{ij}v_{ik}(t_{ij})\geqq 0 \quad \textit{for all } a\in A, \tag{34}$$

and

$$(\operatorname{sgn} \lambda_{ij})(F_i(\bar{a}) - z_i)(t_{ij}) = \|F_i(a) - z_i\|_\infty \quad \text{if } \lambda_{ij} \neq 0. \tag{35}$$

Proof. A short calculation shows that (12)-(14) are equivalent to the existence of $\tau_1, \cdots, \tau_m \geqq 0$ where at least one τ_i is nonzero and $u_i, w_i \in C_{C(Q)'}$, $i = 1, \cdots, m$, with

$$\tau_i = u_i(\mathbb{1}) + w_i(\mathbb{1}), \tag{36}$$

$$\sum_{i=1}^{m} (u_i - w_i)(F_i'(\bar{a})(a - \bar{a})) \geqq 0 \quad \text{for all } a \in A, \tag{37}$$

and

$$u_i(F_i(\bar{a}) - z_i - \bar{\gamma}_i \mathbb{1}) = 0 \quad \text{and} \quad w_i(-F_i(\bar{a}) + z_i - \bar{\gamma}_i \mathbb{1}) = 0 \quad \text{for } i = 1, \cdots, m. \tag{38}$$

$\tau_i = 0$ implies $u_i = w_i = 0_{C(Q)'}$ and nothing needs to be shown. Otherwise define $\bar{u}_i = (1/\tau_i)u_i$, $\bar{w}_i = (1/\tau_i)w_i$ and a representation theorem for linear functionals (see Krabs [12, IV 2.3-4]) on finite-dimensional subspaces of $C(Q)$ gives the existence of q_i points $t_{ij}^+ \in Q$ and numbers $\bar{\lambda}_{ij}^+ \geqq 0$, $i = 1, \cdots, m$, $j = 1, \cdots, q_i$, with

$$\bar{u}_i(y) = \sum_{j=1}^{q_i} \bar{\lambda}_{ij}^+ y(t_{ij}^+).$$

In a similar way there exist r_i points $t_{ij}^- \in Q$ and numbers $\bar{\lambda}_{ij}^- \geqq 0$, $i = 1, \cdots, m$, $j = 1, \cdots, r_i$, with

$$\bar{w}_i(y) = \sum_{j=1}^{r_i} \bar{\lambda}_{ij}^- y(t_{ij}^-).$$

If we define for each $i = 1, \cdots, m$

$$\lambda_{ij} := \bar{\lambda}_{ij}^+ \quad \text{for } j = 1, \cdots, q_i$$

and

$$\lambda_{i\,j+q_i} := -\bar{\lambda}_{ij}^- \quad \text{for } j = 1, \cdots, r_i,$$

and if we set $p_i := q_i + r_i$, then (36) is equivalent to (33), and (37) is equivalent to (34). The analogous application of a known result from optimization (e.g., compare Krabs [13, Thm. I.5.2]) leads to $p_i \leqq \dim \operatorname{span}\{v_{i1}, \cdots, v_{in}, \mathbb{1}, F_i(\bar{a}) - z_i\}$. For $i = 1, \cdots, m$ the equations (38) can be written as

$$\sum_{j=1}^{q_i} \lambda_{ij}[(F_i(\bar{a}) - z_i)(t_{ij}^+) - \|F_i(\bar{a}) - z_i\|_\infty] = 0$$

and

$$\sum_{j=1}^{r_i} \lambda_{ij}[(F_i(\bar{a}) - z_i)(t_{ij}^-) + \|F_i(\bar{a}) - z_i\|_\infty] = 0,$$

which is equivalent to (35). □

With the aid of Lemma 3.13 the multiplier result in Theorem 3.3 leads to an alternation theorem which generalizes the well-known Kolmogorov condition to vector approximation problems. This result extends also an alternation theorem for linear Chebyshev vector approximation problems given in [8, p. 585].

4. Conclusion. In this paper, generalized quasiconvex mappings such as C-quasiconvex and differentiable C-quasiconvex mappings are presented. C-quasiconvex

mappings subsume the classes of quasiconvex, strictly quasiconvex and pseudoconvex functions. It turns out that these notions are very useful in vector optimization and can be applied to optimality conditions and the investigation of local minima. It is one of the main results that the differentiable C-quasiconvexity is characteristic for the sufficiency of a certain multiplier rule for optimal points.

REFERENCES

[1] M. AVRIEL, *Nonlinear Programming: Analysis and Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1976.

[2] J. BORWEIN, *Proper efficient points for maximizations with respect to cones*, this Journal, 15 (1977), pp. 57-63.

[3] ———, *Convex relations in analysis and optimization*, in [25], pp. 335-377.

[4] B. D. CRAVEN, *Vector-valued optimization*, in [25], pp. 661-687.

[5] H. HARTWIG, *Verallgemeinert konvexe Vektorfunktionen und ihre Anwendung in der Vektoroptimierung*, Math. Oper. Stat. Ser. Optim., 10 (1979), pp. 303-316.

[6] R. B. HOLMES, *Geometric Functional Analysis and its Applications*, Springer, New York, 1975.

[7] L. HURWICZ, *Programming in linear spaces*, in Studies in Linear and Nonlinear Programming, K. J. Arrow, L. Hurwicz and H. Uzawa, eds., Stanford Univ. Press, Stanford, CA, 1958.

[8] J. JAHN, *Zur vektoriellen linearen Tschebyscheff-Approximation*, Math. Oper. Stat. Ser. Optim., 14 (1983), pp. 577-591.

[9] ———, *Scalarization in vector optimization*, Math. Programming, 29 (1984), pp. 203-218.

[10] A. KIRSCH, W. WARTH AND J. WERNER, *Notwendige Optimalitätsbedingungen und ihre Anwendung*, Lecture Notes in Economics and Mathematical Systems 152, Springer-Verlag, Berlin, 1978.

[11] W. KRABS, *Über differenzierbare asymptotisch konvexe Funktionenfamilien bei der nicht-linearen gleichmäßigen Approximation*, Arch. Rat. Mech. Anal., 27 (1967), pp. 275-288.

[12] ———, *Optimization and Approximation*, John Wiley, New York, 1979.

[13] H. W. KUHN AND A. W. TUCKER, *Nonlinear programming*, in Proceedings of the Second Berkeley Symposium on Mathematics, J. Neyman, ed., Statistics and Probability, Vol. 5, Univ. California Press, Berkeley, 1951, pp. 481-492.

[14] J. G. LIN, *Maximal vectors and multi-objective optimization*, J. Optim. Theory Appl., 18 (1976), pp. 41-64.

[15] O. L. MANGASARIAN, *Pseudo-convex functions*, SIAM J. Control, 3 (1965), pp. 281-290.

[16] ———, *Nonlinear Programming*, McGraw-Hill, New York, 1969.

[17] R. NEHSE, *Strong pseudo-convex mappings in dual problems*, Math. Oper. Stat., Ser. Optim., 12 (1981), pp. 483-491.

[18] J. VON NEUMANN, *Zur Theorie der Gesellschaftsspiele*, Math. Annal., 100 (1928), pp. 295-320.

[19] H. NIKAIDÔ, *On von Neumann's minimax theorem*, Pacific J. Math., 4 (1954), pp. 65-72.

[20] W. OETTLI, *Optimality conditions for programming problems involving multivalued mappings*, in Modern Applied Mathematics: Optimization and Operations Research, B. Korte, ed., North-Holland, Amsterdam, 1980.

[21] J. PEEMÖLLER, *Verallgemeinerte Quasikonvexitätsbegriffe*, Meth. Oper. Res., 40 (1981), pp. 133-136.

[22] J. P. PENOT, *L'optimisation à la Pareto: Deux ou trois choses que je sais d'elle*, Publications Mathématiques de Pau, 1978.

[23] J. PONSTEIN, *Seven types of convexity*, SIAM Rev., 9 (1967), pp. 115-119.

[24] E. SACHS, *Differentiability in optimization theory*, Math. Oper. Stat., Ser. Optim., 9 (1978), pp. 497-513.

[25] S. SCHAIBLE AND W. T. ZIEMBA, eds., *Generalized Concavity in Optimization and Economics*, Academic Press, New York, 1981.

[26] H. TUY, *Sur les inégalités linéaires*, Colloquium Mathematicum, 13 (1964), pp. 107-123.

[27] W. VOGEL, *Ein Maximum-Prinzip für Vektoroptimierungs-Aufgaben*, Oper. Res. Verfahren, XIX (1974), pp. 161-184.

[28] ———, *Vektoroptimierung in Produkträumen*, Verlag Anton Hain, Mathematical Systems in Economics No. 35, Meisenheim am Glan, 1977.

[29] I. ZANG AND M. AVRIEL, *On functions, whose local minima are global*, J. Optim. Theory Appl., 16 (1975), pp. 183-190.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 2, March 1986

010

SINGULAR OPTIMAL CONTROL: A GEOMETRIC APPROACH*

J. C. WILLEMS†, A. KİTAPÇI† AND L. M. SILVERMAN‡

Abstract. Linear quadratic singular optimal control problem is solved for nonminimum phase and noninvertible systems. A state space decomposition is obtained and a reduced order nonsingular subproblem is solved. The optimal stabilizing input of the singular problem has been found when there are no transmission zeros on the imaginary axis.

Key words. optimal control, singular optimal control, geometric control, distributions

1. Introduction. This paper is concerned with linear quadratic problems in which the cost functional is not positive definite in the control. These are called singular problems. In [1], the finite horizon problem and the infinite horizon problem have been solved when the system is minimum phase. It was also shown that the regular part of the optimal input is feedback implementable.

The geometric theory of linear systems added a great deal of insight into the structure of the solution of such singular problems. In fact, it could be claimed that the theory of (almost) controlled invariant and controllable subspaces are the generic tools for studying this class of problems as demonstrated in [1].

In the present paper, we will investigate the problem further and obtain algorithms for actually computing the optimal control. The nonminimum phase case is also considered and results are found by solving reduced order algebraic Riccati equations. As is well known, the optimal control may not exist in the class of regular control functions and indeed, our optimal trajectory and the ensuing state trajectory lies in the class of distributions. In addition, for positive times, the optimal trajectory is smooth and lies on a predetermined linear subspace of the state space.

We will be using standard notation: $\mathbb{R}^m$ for m-dimensional Euclidean space, $\mathbb{R}^+ := [0, \infty)$, $\mathscr{D}'$ for the distributions with support on $\mathbb{R}^+$, $A \backslash B$ for $A \cap B^{\text{complement}}$, $\langle A | \mathscr{L} \rangle$ for the largest A-invariant subspace containing the subspace $\mathscr{L}$, and $\langle \mathscr{L} | A \rangle$ for the smallest A-invariant subspace contained in $\mathscr{L}$. Of course, for the familiar $\dot{x} = Ax + Bu$, $y = Cx$, $\langle A | \text{im } B \rangle$ is the reachable subspace, while $\langle \ker C | A \rangle$ is the nonobservable subspace.

2. Problem statement. In this paper we will study the full linear quadratic problem with nonnegative cost functional. The usual formulation is to consider, for the system $\dot{x} = Ax + Bu$, the cost functional $\int q(x, u)\, dt$ with q a quadratic form (in x and u jointly). However, since we will only be concerned with the situation in which $q \geqq 0$, we can always introduce the output $y = Cx + Du$ such that $\|y\|^2 = q(x, u)$. As in [1] we are thus led to consider the linear system

$$\Sigma \quad \dot{x} = Ax + Bu, \qquad y = Cx + Du \tag{1}$$

* Received by the editors July 17, 1984. This research was partially supported by the National Science Foundation under grant ECS-53-4510-2569.

† Department of Mathematics, University of Groningen, G.P. 0800, Groningen, The Netherlands. This paper was completed while this author was visiting at the Department of Electrical Engineering-Systems, University of Southern California, Los Angeles, California 90089-0781.

‡ Department of Electrical Engineering-Systems, University of Southern California, Los Angeles, California 90089-0781.

with state space $\mathscr{X}=\mathbb{R}^n$, input space $\mathscr{U}=\mathbb{R}^m$, and output space $\mathscr{Y}=\mathbb{R}^p$, and with cost $\int \|\underline{y}\|^2\,dt$.

Consider the following spaces of inputs:

(i) *Regular inputs*:

$$\mathscr{U}^{\text{reg}}=\mathscr{L}_2^{\text{loc}}(\mathbb{R}^+;\mathbb{R}^m)$$
$$=\left\{\underline{u}:\mathbb{R}^+\to\mathbb{R}^m \;\middle|\; \underline{u} \text{ is measurable and } \int_0^T \|\underline{u}\|^2\,dt<\infty \text{ for all } T\in\mathbb{R}^+\right\}.$$

(ii) *Distributional inputs*: Even though we could consider general distributions on $\mathbb{R}^+$, we will, as in [1], restrict our attention to Bohl type distributions (those whose Laplace transform is rational). Thus

$$\mathscr{U}^{\text{dist}}:=\{\underline{u}\in\mathscr{D}'_+ \mid \underline{u}=\underline{u}^{\text{imp}}+\underline{u}^{\text{reg}} \text{ with } \underline{u}^{\text{imp}} \text{ an impulsive distribution, and } \underline{u}^{\text{reg}}\in\mathscr{U}^{\text{reg}}\}.$$

An impulsive distribution is one with support in 0, i.e., a distribution of the form $\sum_{i=0}^{N} a_i\underline{\delta}^{(i)}$ with $a_i\in\mathbb{R}^m$, $\underline{\delta}$ the Dirac delta, and (i) the i-th derivative.

Let $\mathscr{X}^{\text{reg}}$, $\mathscr{X}^{\text{dist}}$, $\mathscr{Y}^{\text{reg}}$ and $\mathscr{Y}^{\text{dist}}$ be similarly defined. Obviously $\mathscr{U}^{\text{dist}}\supset\mathscr{U}^{\text{reg}}$. Now for any given initial condition $\underline{x}(0)=x_0$ and any $\underline{u}\in\mathscr{U}^{\text{dist}}$, Σ generated in the standard way unique solutions $\underline{x}\in\mathscr{X}^{\text{dist}}$ and $\underline{y}\in\mathscr{Y}^{\text{dist}}$ (for details, see [1, § 3]). In order to display the dependence on x_0 and $\underline{u}$ we will denote these unique solutions by $\underline{x}(x_0,\underline{u})$ and $\underline{y}(x_0,\underline{u})$. Of course, if $\underline{u}\in\mathscr{U}^{\text{reg}}$ then also $\underline{x}(x_0,\underline{u})\in\mathscr{X}^{\text{reg}}$ and $\underline{y}(x_0,\underline{u})\in\mathscr{Y}^{\text{reg}}$. However, it is important to observe that some $\underline{u}\in\mathscr{U}^{\text{dist}}\backslash\mathscr{U}^{\text{reg}}$ may lead to solutions $\underline{y}(x_0,\underline{u})\in\mathscr{Y}^{\text{reg}}$.

Now consider the cost function $\int_0^\infty \|\underline{y}\|^2\,dt$. Formally, define

$$\mathscr{J}:\mathscr{X}\times\mathscr{U}^{\text{dist}}\to\mathbb{R}^e$$

by

$$\mathscr{J}(x_0,\underline{u}):=\int_0^\infty \|\underline{y}(x_0,\underline{u})\|^2\,dt \tag{2}$$

where we will agree to set $\mathscr{J}(x_0,\underline{u})=\infty$ when

$$\underline{y}(x_0,\underline{u})\in\mathscr{Y}^{\text{dist}}\backslash\mathscr{Y}^{\text{reg}} \quad \text{or when} \quad \underline{y}(x_0,\underline{u})\in\mathscr{L}_2^{\text{loc}}\backslash\mathscr{L}_2.$$

We will be interested in minimizing $\mathscr{J}$ with or without stability conditions on the state. Let

$$\mathscr{U}^{\text{dist}}_{\text{stab}}(x_0):=\{\underline{u}\in\mathscr{U}^{\text{dist}} \mid \lim_{t\to\infty}\underline{x}(x_0,\underline{u})(t)=0\}$$

and let $\mathscr{U}^{\text{reg}}_{\text{stab}}(x_0)$ be similarly defined. Now define

$$\mathscr{J}^*(x_0):=\operatorname*{infimum}_{\underline{u}\in\mathscr{U}^{\text{dist}}}\mathscr{J}(x_0,\underline{u})$$

and

$$\mathscr{J}^*_{\text{stab}}(x_0):=\operatorname*{infimum}_{\underline{u}\in\mathscr{U}^{\text{dist}}_{\text{stab}}(x_0)}\mathscr{J}(x_0,\underline{u}).$$

We will study a number of aspects of the cost minimization problem introduced above. In particular we shall answer the following questions:

(i) How can $\mathscr{J}^*$ and $\mathscr{J}^*_{\text{stab}}$ be evaluated? When are $\mathscr{J}^*(x_0)$ and $\mathscr{J}^*_{\text{stab}}(x_0)$ finite? When are they zero?

(ii) Find, if it exists, $\underline{u}^*\in\mathscr{U}^{\text{dist}}$ such that $\mathscr{J}(x_0,\underline{u}^*)=\mathscr{J}^*(x_0)$. Is $\underline{u}^*$ unique? When is $\underline{u}^*\in\mathscr{U}^{\text{reg}}$?

(iii) Same questions for $\underline{u}^*\in\mathscr{U}^{\text{dist}}_{\text{stab}}(x_0)$ and $\mathscr{J}^*_{\text{stab}}(x_0)$.

Example. Before jumping into the details of the analysis, let us consider the special case in which we consider the controllable system Σ: $\dot{x} = Ax + Bu$ and are asked to minimize $\int_0^\infty \|\underline{x}\|^2\, dt$, i.e., $y = x$. Now set $\mathscr{X} = \mathscr{X}_1 \oplus \mathscr{X}_2$ with $\mathscr{X}_1 := \text{im } B$ and $\mathscr{X}_2 := (\text{im } B)^\perp$ In this basis, Σ becomes:

$$\dot{x}_1 = A_{11}x_1 + A_{12}x_2 + u,$$

$$\dot{x}_2 = A_{21}x_1 + A_{22}x_2,$$

and

$$\mathscr{J}(x_0) = \int_0^\infty (\|\underline{x}_1\|^2 + \|\underline{x}_2\|^2)\, dt,$$

where we have chosen also the basis in $\mathscr{U}$ suitably and we have assumed that B is injective.

Note that (A, B) controllable implies (A_{22}, A_{21}) controllable. Let $x_0 = (x_{1,0}, x_{2,0})$ be given. Now solve the classical linear quadratic problem which asks to minimize

$$\int_0^\infty (\|\underline{v}\|^2 + \|\underline{x}_2\|^2)\, dt$$

with $\underline{v}$ as control, for $\dot{x}_2 = A_{22}x_2 + A_{21}v$, $\underline{x}_2(0) = x_{2,0}$. This yields $v^* = Fx_2$ as the optimal control law and $x_{2,0}^T K x_{2,0}$ as the minimal cost. There K is the unique positive definite symmetric solution of the appropriate algebraic Riccati equation and $F = -A_{21}^T K$. Now it is easy to see that $\mathscr{J}(x_{1,0}, x_{2,0}) \geqq x_{2,0}^T K x_{2,0}$, and that $\mathscr{J}^*(x_{1,0}, x_{2,0}) = \mathscr{J}^*_{\text{stab}}(x_{1,0}, x_{2,0}) = x_{2,0}^T K x_{2,0}$ provided $x_{1,0} = Fx_{2,0}$. If, however, $x_{1,0} \neq Fx_{2,0}$ then we can use the impulsive control $\underline{u} = (Fx_{2,0} - x_{1,0})\underline{\delta}$ in order to obtain $\underline{x}_1(0^+) = F\underline{x}_2(0^+) = Fx_{2,0}$. This impulse derives the state to the desired subspace.

The optimal control law then looks like

$$\underline{u} = (Fx_{2,0} - x_{1,0})\underline{\delta} \quad \text{for } t = 0,$$

$$\underline{u} = F(A_{21}x_1 + A_{22}x_2) - A_{11}x_1 - A_{12}x_2 \quad \text{for } t > 0.$$

Our purpose is to generalize this picture: the optimal control consists of an impulse part at $t = 0$. This brings us to a subspace where the rest of the motion takes place and where a classical LQ problem needs to be solved. This surface (a linear subspace) in $\mathscr{X}$ is the *regular subspace.* The computation of this subspace and the control law to be used on it can be carried out by solving a classical algebraic Riccati equation. The computation of the impulses which bring us on this surface involves linear equations only.

3. Some notions from geometric control. The analysis of the singular LQ-problem defined by Σ via (1) and (2) needs the full power of the geometric theory of linear systems as exposed in [2], generalized to "almost" versions and distributional inputs in [3], and further generalized and made relevant to linear quadratic problems in [1]. In this section we will introduce these notions in a self-contained way and recall some relevant facts regarding them.

Consider for the system

$$\Sigma: \dot{x} = Ax + Bu, \qquad y = Cx + Du.$$

The following line-up of subspaces:

(i) $\mathscr{V}^*$, the output nulling subspace, defined as

$$\mathscr{V}^* := \{x_0 \in \mathscr{X} | \exists \underline{u} \in \mathscr{U}^{\text{reg}} \text{ such that } \underline{y}(x_0, \underline{u}) = \underline{0}\};$$

(ii) $\mathscr{R}^*$, the controllable output nulling subspace, defined as

$$\mathscr{R}^* := \{x_0 \in \mathscr{X} | \exists \underline{u} \mathscr{U}^{\text{reg}} \text{ such that } \underline{y}(x_0, \underline{u}) = \underline{0} \text{ and such that } \underline{u}(x_0, \underline{u}) \text{ has compact support}\};$$

(iii) $\mathscr{V}^*_{\mathscr{D}}$, the distributional output nulling subspace, defined as

$$\mathscr{V}^*_{\mathscr{D}} := \{x_0 \in \mathscr{X} | \exists \underline{u} \in \mathscr{U}^{\text{dist}} \text{ such that } \underline{y}(x_0, \bar{u}) = \underline{0} \text{ distribution}\};$$

(iv) $\mathscr{R}^*_{\mathscr{D}}$, the controllable distributional output nulling subspace, defined as

$$\mathscr{R}^*_{\mathscr{D}} := \{x_0 \in \mathscr{X} | \exists \underline{u} \in \mathscr{U}^{\text{dist}} \text{ such that } \underline{y}(x_0, \underline{u}) = \underline{0} \text{ and } \underline{x}(x_0, \underline{u}) \text{ has compact support}\};$$

(v) $\mathscr{V}^*_a$, the L_∞-almost output nulling subspace, defined as

$$\mathscr{V}^*_a := \{x_0 \in \mathscr{X} | \forall \varepsilon > 0, \exists \underline{u} \in \mathscr{U}^{\text{reg}} \text{ such that } \|\underline{y}(x_0, \underline{u})\|_{L_\infty} \leqq \varepsilon\};$$

(vi) $\mathscr{R}^*_a$, the controllable L_∞-almost output nulling subspace, defined as

$$\mathscr{R}^*_a := \{x_0 \in \mathscr{X} | \exists T > 0 \text{ such that } \forall \varepsilon > 0, \exists \underline{u} \in \mathscr{U}^{\text{reg}}, \text{ such that } \|\underline{y}(x_0, u)\|_{L_\infty} \leqq \varepsilon \text{ and support } \underline{x}(x_0, \underline{u}) \subset [0, T]\};$$

(vii) $\mathscr{V}^*_b$, the L_2-almost output nulling subspace, defined as

$$\mathscr{V}^*_b := \{x_0 \in \mathscr{X} | \forall \varepsilon > 0, \exists \underline{u} \in \mathscr{U}^{\text{reg}}, \text{ such that } \|\underline{y}(x_0, \underline{u})\|_{L_2} \leqq \varepsilon\};$$

(viii) $\mathscr{R}^*_b$, the controllable L_2-almost output nulling subspace, defined as

$$\mathscr{R}^*_b := \{x_0 \in \mathscr{X} | \exists T > 0 \text{ such that } \forall \varepsilon > 0, \exists \underline{u} \in \mathscr{U}^{\text{reg}}, \text{ such that } \|\underline{y}(x_0, \underline{u})\|_{L_2} \leqq \varepsilon \text{ and support } \underline{x}(x_0, \underline{u}) \subset [0, T]\}.$$

These subspaces have been studied in [3] for the case $D = 0$, and much of it has been generalized to the case $D \neq 0$ in [1]. Actually the case $D \neq 0$ is easily reduced to the case $D = 0$. Indeed, by choosing the bases in $\mathscr{U}$ and $\mathscr{Y}$ properly, we may always write Σ as

$$\dot{x} = Ax + B_1 u_1 + B_2 u_2, \qquad y_1 = C_1 x + u_1, \qquad y_2 = C_2 x$$

with $\mathscr{U} = \mathscr{U}_1 \oplus \mathscr{U}_2$, $\mathscr{U}_2 = \ker D$, $\mathscr{Y} = \mathscr{Y}_1 \oplus \mathscr{Y}_2$, $\mathscr{Y}_1 = \text{im } D$. Now define

$$\Sigma' : \dot{x} = A'x + B_2 u_2, \qquad y_2 = C_2 x$$

with $A' := A - B_1 C_1$. It is easy to see that the subspaces (i)-(viii) are identical for Σ (with input u and output y) and for Σ' (with input u_2 and output y_2). The properties desired below are easily obtained from this observation and the results of [3]. However, it is convenient to express the relations in terms of Σ directly.

PROPOSITION 1. *There holds*

1. $\mathscr{V}^*_b = \mathscr{V}^*$, $\mathscr{R}^*_b = \mathscr{R}^*_{\mathscr{D}}$;
2. $\mathscr{V}^*_a = \mathscr{V}^* + \mathscr{R}^*_a$, $\mathscr{V}^*_b = \mathscr{V}^* + \mathscr{R}^*_b$;
3. $\mathscr{R}^* = \mathscr{V}^* \cap \mathscr{R}^*_a = \mathscr{V}^* \cap \mathscr{R}^*_b$;
4. $\mathscr{R}^*_a = \mathscr{R}^*_b \cap C^{-1} \text{ im } D$, $\mathscr{R}^*_b = [A|_1 B]((\mathscr{R}^*_a \oplus \mathscr{U}) \cap \ker [C|_1 D])$.

We particularly draw attention to property 4 which yields a simple way of deriving $\mathscr{R}^*_a$ from $\mathscr{R}^*_b$ and vice versa.

In [1], [3] simple algorithms have been derived for the computation of the subspaces (i)-(viii). For the situation at hand, these are

$$\mathscr{V}_0 := \mathscr{K}, \qquad \mathscr{V}_{i+1} = \left[\begin{matrix} A \\ \hline C \end{matrix}\right]^{-1} \left((\mathscr{V}_i \oplus \{0\}) + \operatorname{im} \left[\begin{matrix} B \\ \hline D \end{matrix}\right] \right);$$

$$\mathscr{R}_0 := \{0\}, \mathscr{R}_{i+1} = (C^{-1} \operatorname{im} D) \cap [A_{\mid}^{\mid} B]((\mathscr{R}_i \oplus \mathscr{U}) \cap \ker [C_{\mid}^{\mid} D]),$$

$$\mathscr{S}_0 := \{0\}, \mathscr{S}_{i+1} = [A_{\mid}^{\mid} B]((\mathscr{S}_i \oplus \mathscr{U}) \cap \ker [C_{\mid}^{\mid} D]).$$

These recursive algorithms compute the desired subspaces. In fact,

$$\mathscr{V}_i \downarrow \mathscr{V}_n = \mathscr{V}^*,$$

$$\mathscr{R}_i \uparrow \mathscr{R}_n = \mathscr{R}_a^*,$$

$$\mathscr{S}_i \uparrow \mathscr{S}_n = \mathscr{R}_b^*,$$

$$(\mathscr{V}_i \cap \mathscr{R}_i) \uparrow (\mathscr{V}_n \cap \mathscr{R}_n) = \mathscr{R}^*,$$

$$(\mathscr{V}_i \cap \mathscr{J}_i) \uparrow (\mathscr{V}_n \cap \mathscr{J}_n) = \mathscr{R}^*.$$

These algorithms immediately yield the following.

PROPOSITION 2. *There hold*

1. $\mathscr{R}_a^* = (C^{-1} \operatorname{im} D) \cap [A_{\mid}^{\mid} B]((\mathscr{R}_a^* \oplus \mathscr{U}) \cap \ker [C_{\mid}^{\mid} D])$;
2. $\mathscr{R}_b^* = [A_{\mid}^{\mid} B]((\mathscr{R}_b^* \oplus \mathscr{U}) \cap \ker [C_{\mid}^{\mid} D])$.

The subspaces introduced allow to decide invertibility of Σ. We quote some results to this effect from [1]. We will say that Σ is *right invertible* if for every $\underline{y} \in \mathscr{Y}^{\text{dist}}$ there exists an $\underline{u} \in \mathscr{U}^{\text{dist}}$ such that $\underline{y}(0, \underline{u}) = \underline{y}$. (In [1] it is actually assumed in the definition that $\underline{y} \in \mathscr{Y}^{\text{reg}}$, but the above definition defines an equivalent and perhaps a more natural property.)

PROPOSITION 3. *The following statements are equivalent*:

(i) Σ *is right invertible.*

(ii) $\mathscr{V}_b^* = \mathscr{K}$ *and* $\operatorname{im} [C_{\mid}^{\mid} D] = \mathscr{Y}$.

(iii) *The transfer function* $T(s) = D + C(Is - A)^{-1} B$ *is right invertible over the field of rational functions.*

Also, left invertibility is readily desired from the notions introduced above. The system Σ is called *left invertible* if $\{0 \neq u \in \mathscr{U}^{\text{dist}}\} \Rightarrow \{\underline{y}(0, \underline{u}) \neq \underline{0}\}$. (In [1] it is actually assumed in the definition that $\underline{y}(0, \underline{u}) \in \mathscr{Y}^{\text{reg}}$ but the above definition defines an equivalent and perhaps a more natural property.)

PROPOSITION 4. *The following statements are equivalent*:

(i) Σ *is left invertible.*

(ii) $\mathscr{R}^* = \{0\}$ *and* $\ker \left[\begin{matrix} B \\ \hline D \end{matrix}\right] = \{0\}$.

(iii) *The transfer function* $T(s) = D + C(Is - A)^{-1} B$ *is left invertible over the field of rational functions.*

Note that left invertibility immediately implies that $\{\underline{y}(0, \underline{u}_1) = \underline{y}(0, \underline{u}_2)\} \Rightarrow \{\underline{u}_1 = \underline{u}_2\}$.

Actually, using the results of [4] we can also classify the transfer functions with a polynomial inverse.

PROPOSITION 5. *The following statements are equivalent*:

(i) $\mathscr{R}_b^* = \mathscr{K}$.

(ii) $T(s) = D + C(Is - A)^{-1} B$ *has a right inverse which is a polynomial matrix.*

Finally, the equivalence of the open loop definitions of the spaces (i)-(viii) and their feedback counterparts lies at the basis of many control theoretic applications of these notions. We will only need the following here.

PROPOSITION 6. *There exist a feedback matrix* $F: \mathscr{X} \to \mathscr{U}$ *and a chain* $B_i \subset B$ *such that*

(i) $(A+BF)\mathscr{V}^* \subset \mathscr{V}^*$ *and* $(C+DF)\mathscr{V}^* = \{0\}$;

(ii) $(A+BF)\mathscr{R}^* \subset \mathscr{R}^*$;

(iii) $\mathscr{R}_b^* = B_1 \oplus A_F B_2 \oplus \cdots \oplus A_F^{n-1} B_n$.

Proof. The proof follows from [6, Thm. 1] where the subspaces $\mathscr{R}_b^*$ and $\mathscr{V}^*$ are called strongly reachable and weakly unobservable subspaces and denoted as $\mathscr{W}$ and $\mathscr{V}$.

4. A suitable basis choice and preliminary feedback. By means of an appropriate choice of the basis in the input, state, and output space, and by applying a preliminary feedback, it is possible to simplify the analysis considerably.

Decompose $\mathscr{Y} = \mathscr{Y}_1 \oplus \mathscr{Y}_2$ with $\mathscr{Y}_1 = \operatorname{im} D$ and $\mathscr{Y}_2 = \mathscr{Y}_1^{\perp}$. Now choose $\mathscr{U}_2 = \ker D$ and $\mathscr{U}_1$ such that $\mathscr{U}_1 = \mathscr{U}_1 \oplus \mathscr{U}_2$. By suitably choosing the basis we obtain $D = \left[\begin{smallmatrix} I & 0 \\ 0 & 0 \end{smallmatrix}\right]$. This yields for Σ:

$$\dot{x} = Ax + B_1 u_1 + B_2 u_2, \quad y_1 = C_1 x + u_1, \quad y_2 = C_2 x,$$

with $\|y\|^2 = \|y_1\|^2 + \|y_2\|^2$. It is easy to see that the spaces (i)-(viii) introduced in § 3 are identical for the system Σ as for

$$\dot{x} = A'x + B_2 u_2, \qquad y_2 = C_2 x \text{ with } A' := A - B_1 C_1$$

where we consider u_2 as input and y_2 as output.

Now decompose the state space as follows:

$$\mathscr{X} = \mathscr{X}_1 \oplus \mathscr{X}_2 \oplus \mathscr{X}_3 \oplus \mathscr{X}_4 \oplus \mathscr{X}_5$$

with

$$\mathscr{X}_3 = \mathscr{R}^*, \quad \mathscr{X}_2 \oplus \mathscr{X}_3 = \mathscr{V}^*, \quad \mathscr{X}_3 \oplus \mathscr{X}_4 = \mathscr{R}_a^*,$$

and

$$\mathscr{X}_3 \oplus \mathscr{X}_4 \oplus \mathscr{X}_5 = \mathscr{R}_b^*.$$

Now choose feedback F such that $(A'+B_2F)\mathscr{V}^* \subset \mathscr{V}^*$, $C_2\mathscr{V}^* = \{0\}$, $(A'+B_2F)\mathscr{R}^* \subset \mathscr{R}^*$ and $\mathscr{R}_b^* = B_{20} \oplus (A'+B_2F)B_{21} \oplus (A'+B_2F)^2 B_{22} \oplus \cdots \oplus (A'+B_2F)^{n-1} B_{2n}$ where $B_{20} = B_2$ and B_{2i} is a chain in B_2 (see Proposition 5). This yields $\mathscr{V}^* \cap \operatorname{im} B_2 \subset \mathscr{R}^*$. Also from Proposition 1.4 we know that $\mathscr{R}_a^* = \mathscr{R}_b^* \cap \ker C_2$ and $\mathscr{R}_b^* = (A'+B_2F)\mathscr{R}_1^* + \operatorname{im} B_2$.

The choice of basis indicated and the feedback

$$u_1 = u_1' - C_1 x, \tag{3a}$$

$$u_2 = u_2' + Fx, \tag{3b}$$

reduces our system to $\bar{\Sigma}$

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \\ \dot{x}_5 \end{bmatrix} = \begin{bmatrix} A_{11} & 0 & 0 & 0 & A_{15} \\ A_{21} & A_{22} & 0 & 0 & A_{25} \\ A_{31} & A_{32} & A_{33} & A_{34} & A_{35} \\ A_{41} & 0 & 0 & A_{44} & A_{45} \\ A_{51} & 0 & 0 & A_{54} & A_{55} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} + \begin{bmatrix} B_{11} \\ B_{12} \\ B_{13} \\ B_{14} \\ B_{15} \end{bmatrix} u_1' + \begin{bmatrix} 0 \\ 0 \\ B_{23} \\ B_{24} \\ B_{25} \end{bmatrix} u_2',$$

$$y_1 = u_1', \qquad y_2 = [C_{21}\ 0\ 0\ 0\ C_{25}] \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix}. \tag{4}$$

The problem is to

$$\text{minimize} \int_0^\infty (\|u_1'\|^2 + \|y_2\|^2)\, dt. \tag{5}$$

We have the following.

PROPOSITION 7.

1. $\ker C_{25} = \{0\}$.
2. *The transfer function associated with*

$$\left\{ \begin{bmatrix} A_{33} & A_{34} & A_{35} \\ 0 & A_{44} & A_{45} \\ 0 & A_{54} & A_{55} \end{bmatrix}, \begin{bmatrix} B_{23} \\ B_{24} \\ B_{25} \end{bmatrix}, [0\ 0\ I] \right\}$$

has a right inverse which is a polynomial matrix.

3. *The system*

$$\left\{ \begin{bmatrix} A_{44} & A_{45} \\ A_{54} & A_{55} \end{bmatrix}, \begin{bmatrix} B_{24} \\ B_{25} \end{bmatrix}, [0\ I] \right\}$$

is left invertible.

4. (A_{33}, B_{23}) *is a controllable pair.*

Proof.

1. It follows from $\mathscr{R}_b^* \cap \ker C_2 = \mathscr{R}_a^*$.

2. It follows from Proposition 4 that this transfer function with the output matrix replaced by $[0\ 0\ C_{25}]$ has a polynomial right inverse. The result then follows using 1.

3. By Proposition 4 we need to show that the controllability output nulling subspace associated with this system is zero. Assume that this is not the case and add this subspace to $\mathscr{X}_3$. Clearly the subspace obtained in this way will be in the controllability output nulling subspace for the original system Σ, proving the claim.

4. Follows from $\langle A' + B_2F | \operatorname{im} B_2 \cap \mathscr{R}^* \rangle = \mathscr{R}^*$. □

We will use the feedback F and the chain B_{2i} to simplify the system representation given in (4). We obtain

PROPOSITION 8. *There exists a coordinate transformation such that* $\bar{\Sigma}_T = \Sigma^*$ *where*

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \\ \dot{x}_5 \end{bmatrix} = \begin{bmatrix} A_{11}^* & 0 & 0 & 0 & A_{15} \\ A_{21}^* & A_{22} & 0 & 0 & A_{25} \\ 0 & A_{32} & A_{33} & A_{34} & A_{35} \\ 0 & 0 & 0 & A_{44} & A_{45} \\ 0 & 0 & 0 & A_{54} & A_{55} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} + \begin{bmatrix} B_{11} \\ B_{12} \\ B_{13} \\ B_{14} \\ B_{15} \end{bmatrix} u_1' + \begin{bmatrix} 0 \\ 0 \\ B_{23} \\ B_{24} \\ B_{25} \end{bmatrix} u_2',$$

$$y_2 = [C_{21}^*\ 0\ 0\ 0\ C_{25}]\underline{x}, \qquad y_1 = u_1'$$

where (C_{21}^*, A_{11}^*) *is an observable pair.*

Proof. Follows from Lemma 1 and the Column Elimination Algorithm given in [6]. □

5. Regular *LQ* problems. Our approach in studying singular LQ-problems will be to reduce them to regular LQ-problems. Regular LQ-problems are those for which distributional inputs are not candidates for optimal controls since they always lead to infinite cost:

DEFINITION. The LQ problem Σ (as defined by (1), (2)) will be called *regular* if $\{x_0 \in \mathscr{X},\ \underline{u} \in \mathscr{U}^{\text{dist}} \setminus \mathscr{U}^{\text{reg}}\} \Rightarrow \{\mathscr{J}(x_0, \underline{u}) = \infty\}$.

It is easy to decide regularity as follows.

THEOREM 1. *The following conditions are equivalent*:

(i) Σ *defines a regular LQ-problem.*

(ii) $\{x_0 \in \mathscr{X},\ \underline{u} \in \mathscr{U}^{\text{dist}} \setminus \mathscr{U}^{\text{reg}}\} \Rightarrow \{\underline{y}(x_0, \underline{u}) \in \mathscr{Y}^{\text{dist}} \setminus \mathscr{Y}^{\text{reg}}\}$.

(iii) $\ker D = \{0\}$.

Proof. (i)$\Rightarrow$(ii): Since $\underline{y}(x_0, \underline{u}) = \underline{y}(x_0, 0) + \underline{y}(0, \underline{u})$, (ii) is equivalent to

$$\{\underline{u} \in \mathscr{U}^{\text{dist}} \setminus \mathscr{U}^{\text{reg}}\} \Rightarrow \{\underline{y}(0, \underline{u}) \in \mathscr{Y}^{\text{dist}} \setminus \mathscr{Y}^{\text{reg}}\}.$$

Now, if this were not the case, $\exists \underline{u} \in \mathscr{U}^{\text{dist}} \setminus \mathscr{U}^{\text{reg}}$ such that $\underline{y}(0, \underline{u}) \in \mathscr{Y}^{\text{reg}}$. The corresponding $\underline{x}(0, \underline{u})$ will satisfy $\underline{x}(0, \underline{u})(t) \in \langle A | \text{im } B \rangle$ for all $t > 0$. In particular, $\underline{x}(0, \underline{u})(1) \in \langle A | \text{im } B \rangle$. Hence since $\underline{x}(0, u)(1)$ belongs to the controllable subspace of Σ, we can modify, if needed, $\underline{u}(t)$ to $\underline{u}^{\text{new}}(t)$ for $t \geqq 1$ such that $y(0, \underline{u}^{\text{new}}) \in \mathscr{L}(0, \infty)$. This $\underline{u}^{\text{new}}$ is still in $\mathscr{U}^{\text{dist}} \setminus \mathscr{U}^{\text{new}}$, but $\mathscr{J}(x_0, \underline{u}^{\text{new}}) < \infty$. Hence {not (ii)}$\Rightarrow${not (i)}. The implication (ii)$\Rightarrow$(i) is obvious.

To show the equivalence of (ii) and (iii), observe that by a suitable choice of the basis in $\mathscr{U}$ and $\mathscr{Y}$, Σ may be represented as

$$\dot{x} = Ax + B_1 u_1 + B_2 u_2, \quad y_1 = C_1 x + u_1, \quad y_2 = C_2 x,$$

$$u = (x_1, u_2), \qquad y = (y_1, y_2)$$

with $u_1 \in \mathbb{R}^{m_1}$, $m_1 = \text{codim ker } D = \dim \text{im } D$ and $u_2 \in \mathbb{R}^{m_2}$, $m_2 = m - m_1$. Now (ii)$\Leftrightarrow$ $\{m_2 = 0\} \Leftrightarrow$ (iii) is obvious. □

Regular LQ problems may thus be reduced by a simple basis change and a feedback transformation to the standard LQ problems.

Recall the LQ problem standard if it is regular and if the associated $\mathscr{V}^* = \{0\}$, i.e., if $\ker D = \{0\}$ and if $\langle C^{-1} \text{ im } D | A - (D^T D)^{-1} D^T C \rangle = \{0\}$.

Let Σ define a regular LQ-problem. By Theorem 1 this is equivalent to $\ker D = \{0\}$. By choosing the basis in $\mathscr{U}$ appropriately and making an orthogonal basis change in $\mathscr{Y}$ we can then bring D into the form $\begin{bmatrix} I \\ 0 \end{bmatrix}$. Σ becomes

$$\dot{x} = Ax + Bu, \quad y_1 = C_1 x_2 + u, \quad y_2 = C_2 x, \quad y = (y_1, y_2).$$

Now use the preliminary feedback $u = v - C_1 x_2$. This yields the system

$$\dot{x} = A' x + Bv, \qquad y_2 = C_2 x$$

with $A' := A - B_1 C_1$ and $\mathscr{J} = \int_0^\infty (\|\underline{y}\|^2 + \|\underline{y}_2\|^2)\, dt$.

We obtain the familiar standard LQ problem

$$\text{minimize} \int_0^\infty (\|\tilde{u}\|^2 + \|\tilde{y}\|^2)\, dt$$

for $\dot{\tilde{x}} = \tilde{A}\tilde{x} + \tilde{B}\tilde{u}$; $\tilde{y} = \tilde{C}\tilde{x}$ with the simple basis change such that $\tilde{C} = C/L$ where $L = (\mathscr{V}^*)^\perp$.

6. The singular *LQ*-problem without stability constraints. At this point it is convenient to study the LQ-problem introduced in § 2 with and without stability separately.

We will reduce the general singular problem to regular ones, and regular problems to standard ones. In addition, we will assume throughout that (A, B) is asymptotically stabilizable. We have the well-known proposition as follows.

PROPOSITION 9. *Let (A, B) be asymptotically stabilizable and assume that Σ defines a standard LQ-problem. Then the control law $u^* = F_0 x$ generates $\mathcal{J}(x_0, \underline{u}^*) = \mathcal{J}^*(x_0)$. Here*

$$F_0 = -(D^T D)^{-1}(B^T P_0 + D^T C) \tag{6}$$

and P_0 is the unique positive semi-definite solution of the algebraic Riccati equation

$$A^T P + PA - (PB + C^T D)^T (D^T D)^{-1}(PB + C^T D) + C^T C = 0. \tag{7}$$

In fact, $P_0 > 0$ and $\mathcal{J}^*(x_0) = x_0^T P_0 x_0$. Moreover, the closed loop system $\dot{x} = (A + BF_0)x$ is asymptotically stable.

It is easy to extend Proposition 9 to the regular case.

PROPOSITION 10. *Let (A, B) be asymptotically stabilizable and assume that Σ defines a regular LQ-problem. Then the control law (6) with P_0 the infimal positive semidefinite solution of the algebraic Riccati equation (7) generates $\mathcal{J}(x_0, u^*) = \mathcal{J}^*(x_0)$, and $\mathcal{J}^*(x_0) = x_0^T P_0 x_0$. Further $\{\mathcal{J}^*(x_0) = 0\} \Leftrightarrow \{x_0 \in \ker P_0\} \Leftrightarrow \{x_0 \in \mathcal{V}^* = \langle C^{-1} \operatorname{im} D | (D^T D)^{-1} D^T C \rangle\}$. Finally, the closed loop system $\dot{x} = (A + BF_0)x$ will be asymptotically stable if and only if $\mathcal{V}^* \subset \mathcal{L}^-(A - B(D^T D)^{-1} D^T C)$ (i.e., detectability).*

Proof. See [1] with the sign change on (6.2) at page 23, [7].

Note that in order to solve for P_0 and F_0 in Proposition 10 it suffices to solve a standard algebraic Riccati equation of dimension = the codimension of $\mathcal{V}^*$, since $P_0 = P_0^T \geqq 0$ and $\ker P_0 = \mathcal{V}^*$.

We now have all the preliminary results which go into the solution of the general singular *LQ* problem *without* stability. We will assume that the problem is already in the form (4)-(5).

THEOREM 2. *Assume that (A, B) is asymptotically stabilizable and consider the singular LQ-problem (4)-(6). Then*

(i) $\mathcal{J}^*(x_0) < \infty$. *In fact,*

$$\mathcal{J}^*((x_{1,0}, x_{2,0}, x_{3,0}, x_{4,0}, x_{5,0})) = x_{1,0}^T P_0 x_{1,0}$$

with P_0 the unique positive semidefinite solution of the algebraic Riccati equation

$$\begin{aligned} A_{11}^T P + PA_{11} - (PA_{15} + C_{21}^T C_{25})^T (C_{25}^T C_{25})^{-1}(PA_{15} + C_{21}^T C_{25}) \\ - PB_{11} B_{11}^T P + C_{21}^T C_{21} = 0. \end{aligned} \tag{8}$$

Moreover, $P_0 > 0$.

(ii) *$\forall x_0 \in \mathcal{X}$, there exists an $u^* \in \mathcal{U}^{\text{dist}}$ such that $\mathcal{J}(x_0, u^*) = \mathcal{J}^*(x_0)$. This optimal control is generated as follows*

$$u_1'^* = -B_{11}^T P_0 x_1 \tag{9}$$

and $\underline{u}_2'^*$ such that $\underline{x}_5^*$ is regular and satisfies

$$x_5^* = -(C_{25}^T C_{25})^{-1}(A_{15}^T P_0 + C_{25}^T C_{21}) x_1. \tag{10}$$

There always exists a distribution $\underline{u}_2'^*$ such that (10) will be satisfied as a distribution.

(iii) *The optimal trajectory lies on the linear subspace*

$$x_5 = -(C_{25}^T C_{25})^{-1}(A_{15}^T P_0 + C_{25}^T C_{21}) x_1$$

for $t > 0$.

(iv) *The optimal trajectory*

$$\underline{x}_1^*, \underline{x}_2^*, \underline{x}_3^*, \underline{x}_4^*, \underline{x}_5^*$$

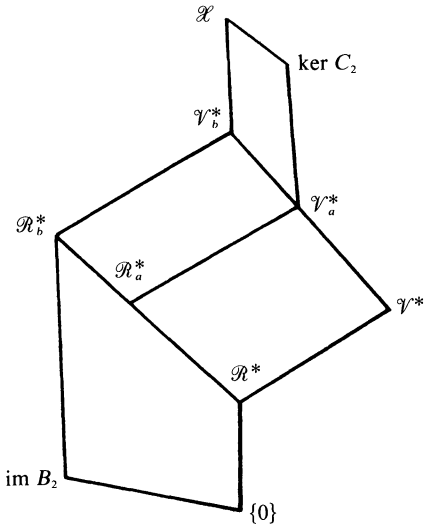
is such that

$$\underline{x}_1^* \text{ and } \underline{x}_2^* \text{ are regular}$$

and $\underline{x}_3^*$, $\underline{x}_4^*$, *and* $\underline{x}_5^*$ *may be distributions. Moreover,* $\underline{x}_1^*$, $\underline{x}_2^*$, $\underline{x}_4^*$ *and* $\underline{x}_5^*$ *are uniquely defined, while* $\underline{x}_3^*$ *is not.*

The proof of this theorem is given in Appendix A.

Theorem 1 allows us to recognize several interesting special cases of the singular LQ-problem. Recall the following lattice diagram (B_2 and C_2 are as defined in § 4):



The problem is *standard*:
$\Leftrightarrow$ {the optimal control is a regular function and $\mathscr{J}^*(x_0) > 0$ for $x_0 \neq 0$}
$\Leftrightarrow \{\mathscr{V}_b^* = \{0\}\}$.

The problem is *regular*:$\Leftrightarrow$ {the optimal control is a regular function}
$\Leftrightarrow \{\mathscr{R}_b^* = \{0\}\}$
$\Leftrightarrow$ {Ker $D = \{0\}$}.

The problem is *cheap*:$\Leftrightarrow \{\mathscr{J}^*(x_0) = 0$ for all $x_0\}$
$\Leftrightarrow \{\mathscr{V}_b^* = \mathscr{X}\}$.

The problem is *totally singular*:$\Leftrightarrow$ {the optimal control has zero regular part}
$\Leftrightarrow \{\mathscr{R}_b^* = \mathscr{X}$ and $\mathscr{R}^* = \{0\}\}$.

The problem is *potentially singular*:
$\Leftrightarrow$ {there always is an optimal control with regular part zero}
$\Leftrightarrow \{\mathscr{R}_b^* = \mathscr{X}\}$.

7. The singular LQ-problem with stability. In this section we will generalize the ideas of § 6 in order to study the singular LQ-problem with the stability constraint $\lim_{t\to\infty} \underline{x}(t) = 0$ as a side condition. We start by analyzing the geometric structure of Σ as given in § 2 in still a bit more detail and derive a refinement of the decomposition (4).

7.1. A further decomposition of $\mathscr{V}^*$. Our approach to solve the singular LQ-problem with stability needs a further decomposition of the output nulling subspace

$\mathcal{V}^*$. Consider $\mathcal{V}^-$, $\mathcal{V}^0$, and $\mathcal{V}^+$ the output nulling subspaces with respectively asymptotic stability, neutral stability, and exponential instability. These are defined and computed as follows.

Take any F such that $(A+BF)\mathcal{V}^* \subset \mathcal{V}^*$ (solutions converge neither for $t \to \pm\infty$) and $(C+DF)\mathcal{V}^* = \{0\}$. Then $(A+BF)\mathcal{R}^* \subset \mathcal{R}^*$. Now there exists such an F with the property that the characteristic polynomial of $(A+BF)|\mathcal{R}^*$ is equal to any given monic polynomial of suitable degree. However, the eigenvalues of $(A+BF)$ (mod $\mathcal{R}^*$) are independent of the F which we choose. Now choose an F such that the spectra of $(A+BF)|\mathcal{R}^*$ and $(A+BF)$ (mod $\mathcal{R}^*$) are disjoint. This yields a decomposition of $\mathcal{V}^*$ into $\mathcal{V}^* = \mathcal{V}_1 \oplus \mathcal{V}_2 \oplus \mathcal{V}_3 \oplus \mathcal{R}^*$ with $\mathcal{V}_1$, $\mathcal{V}_2$ and $\mathcal{V}_3$ $(A+BF)$-invariant and such that $(A+BF)|\mathcal{V}_1$, $(A+BF)|\mathcal{V}_2$ and $(A+BF)|\mathcal{V}_3$ have their spectra in the open right half plane, on the imaginary axis, and in the open left half plane, respectively. In terms of these, set $\mathcal{V}^+ = \mathcal{R}^* \oplus \mathcal{V}_1$, $\mathcal{V}^0 = \mathcal{R}^* \oplus \mathcal{V}_2$, and $\mathcal{V}^- = \mathcal{R}^* \oplus \mathcal{V}_3$.

Using such an F and the above decomposition of $\mathcal{X}_2 \oplus \mathcal{X}_3$ in (4) yields a decomposition of $\mathcal{X}_2$ into $\mathcal{X}_2 = \mathcal{X}_{21} \oplus \mathcal{X}_{22} \oplus \mathcal{X}_{23}$ with an associated partitioning of A_{22}, A_{25}, and B_{12} into

$$(11) \qquad A_{22} = \begin{bmatrix} A_{22,1} & 0 & 0 \\ 0 & A_{22,2} & 0 \\ 0 & 0 & A_{22,3} \end{bmatrix} \quad A_{25} = \begin{bmatrix} A_{25,1} \\ A_{25,2} \\ A_{25,3} \end{bmatrix} \quad B_{12} = \begin{bmatrix} B_{12,1} \\ B_{12,2} \\ B_{12,3} \end{bmatrix}$$

with $\sigma(A_{22,1})$, $\sigma(A_{22,2})$, and $\sigma(A_{22,3})$ in the open right half plane, on the imaginary axis, and in the open left half plane, respectively. Furthermore, $A_{32} = 0$.

7.2. The regular *LQ*-problem with stability. In the *standard* problem we obtain asymptotic stability of the closed loop system as a consequence of the minimization of $\mathcal{J}$. This shows that the standard problem has the same solution with or without the side constraint $\lim_{t\to\infty} \underline{x}(t) = 0$. The difference starts when we consider the regular problem.

Consider now the regular *LQ* problem: Σ with $\ker D = \{0\}$. By Theorem 1 we may restrict attention to $\mathcal{U}^{\text{reg}}$. Now consider the subspaces $\mathcal{V}^+$, $\mathcal{V}^0$, and $\mathcal{V}^-$ introduced in § 7.1. Because of regularity, these may be computed in more detail:

$$\mathcal{V}^* = \langle C^{-1} \operatorname{im} D | A' \rangle \quad \text{with } A' := A - (D^T D)^{-1} D^T C$$

and, also because of regularity, $\mathcal{R}^* = \{0\}$. Now make a spectral decomposition of $\mathcal{V}^*$ corresponding to the decomposition of the spectrum of $A'|\mathcal{V}^*$ into its open right half plane, its imaginary axis, and its open left half plane parts. This yields $\mathcal{V}^+$, $\mathcal{V}^0$, and $\mathcal{V}^-$ respectively.

We obtain the following proposition.

PROPOSITION 11. *Consider the regular LQ-problems*: Σ *with* $\ker D = \{0\}$, *with the stability condition* $\lim_{t\to\infty} \underline{x}(t) = 0$. *Then*

(i) *For all* $x_0 \in \mathcal{X}$, *there exists a control* $\underline{u} \in \mathcal{U}^{\text{reg}}_{\text{stab}}(x_0)$ *such that* $\mathcal{J}(x_0, \underline{u}) < \infty$ *if and only if* (A, B) *is asymptotically stabilizable. Assume this to be the case.*

(ii) *There exists a supremal nonnegative definite symmetric solution,* P_+, *to the algebraic Riccati equation* (7). *We have* $\inf_{u\in\mathcal{U}^{\text{reg}}_{\text{stab}}} \mathcal{J}(x_0, \underline{u}) = \inf_{u\in\mathcal{U}^{\text{dist}}_{\text{stab}}} \mathcal{J}(x_0, \underline{u}) = x_0^T P_+ x_0$.

(iii) *For all* $x_0 \in \mathcal{X}$ *there exists an optimal control* $u^* \in \mathcal{U}^{\text{reg}}_{\text{stab}}(x_0)$ (*hence* $\mathcal{J}(\underline{u}^*, x_0) = \underline{x}_0^T P_+ x_0$) *if and only if* $\mathcal{V}^0 = \{0\}$.

(iv) *Assuming this to be the case, then* $u^* = F_+ x$ *with* $F_+ := -(D^T D)^{-1}(B^T P_+ + D^T C)$ *generates the optimal control.*

(v) $\{\inf_{u\in\mathcal{U}^{\text{reg}}_{\text{stab}}} \mathcal{J}(x_0, \underline{u}) = 0\} \Leftrightarrow \{x_0 \in \ker P_+\} \Leftrightarrow \{x_0 \in \mathcal{V}^0 + \mathcal{V}^-\}$.

(vi) $P_+ = P_0$ (*and consequently the nonnegative definite solution of* (7) *is unique*) *if and only if* $\mathcal{V}^+ = \{0\}$ (*i.e., exponential detectability*).

(vii) $u^* = F_0 x$ *will yield also asymptotic stability if and only if* $\mathcal{V}^+ + \mathcal{V}^0 = \{0\}$. *In this case, the LQ problem* Σ *with and without stability give identical answers.*

Proof. Of course, part (i) is obvious. Assume thus that (A, B) is asymptotically stabilizable using the representation derived at the end of § 5. It follows that we should prove this proposition for the LQ-problem in which we are asked to minimize $\int_0^\infty (\|u\|^2 + \|y\|^2)\, dt$ for the system $\dot{x} = Ax + Bu$; $y = Cx$. Then $\mathcal{V}^* = \langle \ker C | A \rangle$, while $\mathcal{V}^+$, $\mathcal{V}^0$, $\mathcal{V}^-$ correspond to the decomposition of $\mathcal{V}^*$ associated with the partition of the spectrum of $A|\mathcal{V}^*$ into its components in the open right half plane, on the imaginary axis, and in the open left half plane, respectively. The associated algebraic Riccati equation is

$$A^T P + PA - PBB^T P + C^T C = 0.$$

Let P_0 be the infimal nonnegative definite symmetric solution of this algebraic Riccati equation. Since (A, B) is asymptotically stabilizable, such a solution exists. Using standard calculations, it is easy to see that, whenever $\lim_{t\to\infty} \underline{x}(t) = 0$, then $\mathcal{J}(x_0, \underline{u}) = x_0^T P_0 x_0 + \int_0^\infty \|\underline{u} + B^T P_0 \underline{x}(x_0, \underline{u})\|^2\, dt$. Now use the preliminary feedback $u = v - B^T P_0 x$. The problem then requires the minimization of $\int_0^\infty \|\underline{v}\|^2\, dt$ for $\dot{x} = A_0 x + Bv$ with $A_0 := A - BB^T P_0$, under the stability constraint $\lim_{t\to\infty} \underline{x}(x_0, \underline{v})(t) = 0$. Let $\mathcal{L}^+$, $\mathcal{L}^0$, $\mathcal{L}^-$ be the decomposition of $\mathcal{X}$ corresponding to the partition of the spectrum of A_0 into its components in the open right half plane, on the imaginary axis, and in the open left half plane. By Proposition 9, we know that $\ker P_0 = \mathcal{V}^* = \langle \ker C | A \rangle$. Further $\mathcal{V}^*$ is A_0-invariant and $A_0 \pmod{\mathcal{V}^*}$ has its eigenvalues in the open left half plane.

Now minimize $\mathcal{J}(\underline{x}_0, \underline{u}) = \int_0^\infty \|\underline{v}\|^2\, dt$ subject to $\dot{x} = A_0 x + Bv$, $\underline{x}(0) = x_0$, and $\lim_{t\to\infty} \underline{x}(x_0, \underline{v})(t) = 0$. Clearly if $x_0 \in \mathcal{L}^-$, the optimal control $\underline{v}^* = 0$, and $\min \mathcal{J}'(x_0, \underline{v}^*) = 0$. Next, if $0 \neq x_0 \in \mathcal{L}^0$, $\inf \mathcal{J}'(x_0, \underline{v}^*) = 0$ (see [5, Lemma 3.2]) but no optimal control exists since $\underline{v}^* = 0$ does not meet the condition $\lim_{t\to\infty} \underline{x}(x_0, \underline{v}^*) = 0$. Consider now the situation $x_0 \in \mathcal{L}^+$.

Define $A_+ := A_0|\mathcal{L}^+$ and $B_+ := QB$ with Q the projection of π onto $\mathcal{L}^+$ along $\mathcal{L}^0 \oplus \mathcal{L}^-$. Note that the stabilizability of (A, B) implies that (A_+, B_+) is controllable. Further the eigenvalues of A_+ are in the open right half plane. Now solve the minimization of $\int_0^\infty \|\underline{v}\|^2\, dt$ for $\dot{x}_+ = A_+ x_+ + B_+ v$ with $\underline{x}_+(0) = x_{+,0}$ and $\lim_{t\to\infty} \underline{x}_+(t, \underline{v}) = 0$. The optimal control for this problem is $v^* = -B^T W_+^{-1} x_+$ with W_+, related to the controllability Grammian, defined as the unique solution of $A_+ W_+ + W_+ A_+^T = B_+ B_+^T$. Clearly $W_+ = W_+^T > 0$, and hence $\pi_+ = W_+^{-1}$ is the supremal (alternatively, the unique positive definite) symmetric solution of $\pi_+ A_+ + A_+^T \pi_+ - \pi_+ B_+ B_+^T \pi_+ = 0$. Now combine the solution which we obtained for $\mathcal{L}^-$, $\mathcal{L}^0$, and $\mathcal{L}^+$. Define

$$\pi = \begin{bmatrix} \pi_+ & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

to conform with the partition of $\mathcal{X}$ into $\mathcal{X} = \mathcal{L}^+ \oplus \mathcal{L}^0 \oplus \mathcal{L}^-$. This yields $x_0^T \pi x_0$ as $\inf \mathcal{J}'(x_0, \underline{v})$. If $\mathcal{X} = \mathcal{L}^+ \oplus \mathcal{L}^-$ then $v^* = -B^T \pi x$ is the optimal control law.

Combining this solution with the preliminary feedback yields $x_0^T (P_0 + \pi) x_0$ for $\inf \mathcal{J}(x_0, \underline{u})$ and $u^* = -B^T (P_0 + \pi) x$. Define now $P_+ = P_0 + \pi$ and unify all the statements of Proposition 11.

7.3. The singular case with stability. We are now in a position to state the solution of the general singular LQ problems with stability. We will assume that the problem

is already in the form (4)-(5) with the refinement of $\mathscr{X}_1$ leading to the partition of A_{22} as given in (11).

THEOREM 3. *Consider the singular LQ-problem* (2) *with the stability constraint* $\lim_{t\to\infty} \underline{x}(t) = 0$. *Assume that by a preliminary feedback and a proper choice of the bases,* Σ *is already in the form* (4)-(5), *with* $\mathscr{X}_2$ *further decomposed so as to induce the form* (11). *Then*

(i) *For all* $x_0 \in \mathscr{X}$, *there exists a control* $\underline{u} \in \mathscr{U}_{\text{stab}}^{\text{dist}}(x_0)$ *with* $\mathscr{J}(x_0, \underline{u}) < \infty$ *if and only if* (A, B) *is asymptotically stabilizable. Assume this to be the case.*

(ii) *Let* P_0 *be as defined in Theorem* 2 (i). *Now let* W_+ *be the solution of*

$$A_{22,1} W_+ + W_+ A_{22,1}^T = B_{12,1} B_{12,1}^T + A_{25,1} A_{25,1}^T. \tag{12}$$

Then $W_+ = W_+^T > 0$, *and* $\mathscr{J}_{\text{stab}}^*(x_0) = x_{1,0}^T P_0 x_{1,0} + x_{21,0}^T W_+^{-1} x_{21,0}$.

(iii) *For all* $x_0 \in \mathscr{X}$, *there exists an optimal control* $\underline{u}^* \in \mathscr{U}_{\text{stab}}^{\text{dist}}(x_0)$ *if and only if* $\mathscr{X}_{22} = 0$ (*in the notation of* § 7.1 *this means* $\mathscr{V}^* = \mathscr{V}^+ + \mathscr{V}^-$). *This optimal control is generated as follows*

$$u_1^* = -B_{11}^T P_0 x_1 - B_{21,1}^T W_+^{-1} x_{21} \tag{13}$$

and $u_2'^*$ *such that* x_5^* *is regular and satisfies*

$$x_5^* = -(C_{25}^T C_{25})^{-1} (A_{15}^T P_0 + C_{25}^T C_{21}) x_1 - A_{21,1}^T W_+^{-1} x_{21}. \tag{14}$$

7.4. Computation of optimal input. In this part we will discuss the computation of the optimal input. Using Theorem 3 (iii), we will obtain $u_1'^*$, x_5^*, x_1^*, x_{21}^* for initial conditions x_{10} and $x_{21,0}$. To compute $u_2'^*$ we consider the differential equation

$$\begin{bmatrix} \dot{x}_{2,3} \\ \dot{x}_3 \\ \dot{x}_4 \\ \dot{x}_5 \end{bmatrix} = \begin{bmatrix} A_{22,3} & 0 & 0 & A_{25} \\ 0 & A_{33} & A_{34} & A_{35} \\ 0 & 0 & A_{44} & A_{45} \\ 0 & 0 & A_{54} & A_{55} \end{bmatrix} \begin{bmatrix} x_{2,3} \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} + \begin{bmatrix} B_{12,3} \\ B_{13} \\ B_{14} \\ B_{15} \end{bmatrix} u_1' + \begin{bmatrix} 0 \\ B_{23} \\ B_{24} \\ B_{25} \end{bmatrix} u_2'.$$

We will first compute $\bar{x}_{2,3}$, $\bar{x}_3$, $\bar{x}_4$, $\bar{x}_5$ by taking $u_1' = u_1^*$ and $u_2' = 0$. Since from Theorem 3 part (iii) $\mathscr{X}_{22} = 0$ we can conclude

$$\left\langle \begin{bmatrix} A_{22,3} & 0 & 0 & A_{25} \\ 0 & A_{33} & A_{34} & A_{35} \\ 0 & 0 & A_{44} & A_{45} \\ 0 & 0 & A_{54} & A_{55} \end{bmatrix}, \begin{bmatrix} 0 \\ B_{23} \\ B_{24} \\ B_{25} \end{bmatrix} \right\rangle$$

$$\begin{bmatrix} 0 & 0 & 0 & I \end{bmatrix}$$

has $\mathscr{V}^* + \mathscr{R}_b^* = \mathscr{X}$ (see Proposition 6.2). Therefore, we will compute $u_2'^*$ such that $\Delta x_5(t) = \bar{x}_5(t) - x_5^*(t)$ will be zero. If $\Delta x_5(0) \neq 0$, $u_2'^*$ contains impulses. Using the results of [6] one can directly compute the impulsive and regular parts of $u_2'^*$. We will first find the feedback F and chain B_i which are defined by Proposition 5 by using [6, Theorem 1] for $A' = A - B_1 C_1$, B_2 and C_2. By applying the nested version of the left structure algorithm one can find output transformation Q, input transformation G and feedback F such that $A_F = A' + B_2 F$, $C_2^T = [(C_1^*)^T \cdots (C_\alpha^*)^T]$ and $B_2 = [B_{21} \cdots B_{2\alpha+1}]$ where $(C_i^*)^T$, $B_{2i} \in \mathscr{R}^{nx(q_i - q_{i-1})}$. If $Q \neq I$ one has to introduce Q into *ARE* in Theorem 3. We will choose $[T_3 \ T_4] = \text{Im}\,[B_{22} \ B_{23} \ A_F B_{23} \cdots B_{\alpha+1} \cdots A_F^{\alpha-2} B_{\alpha+1}]$ and $T_5 = [B_{21} \ A_F B_{22} \cdots A_F^{\alpha-1} B_{2\alpha}]$ where columns of $[T_3 \ T_4]$ and T_5 are basis vectors for x_3, x_4 and x_5 respectively. With this

special basis selection we will obtain the matrix

$$N = \begin{bmatrix} A_{33} & A_{34} & A_{35} & B_{23} \\ 0 & A_{44} & A_{45} & B_{24} \\ 0 & A_{54} & A_{55} & B_{25} \end{bmatrix}$$

where

$$N = \begin{bmatrix} N_{11} & \cdots & N_{1\theta_1} \\ \vdots & & \vdots \\ N_{\theta_2 1} & \cdots & N_{\theta_2\theta_1} \end{bmatrix}$$

and for each $i \exists aj^*(i)$ such that $N_{i,j^*(i)} = I$ where $j^*(i) < i$. Then we will apply the column elimination algorithm [6] to eliminate nonzero elements of each row. Let $J = \{j | \exists i \ni j^*(i) = j\}$. For each $j \in J$, $\dot{x}_{j^*(i)} + A_{55}\Delta x_5(t)$. We start the procedure with $i = \theta_2$, at each step we know x_i and compute $x_{j^*(i)}$ since $j^*(i) < i$. Recursively one can find $x_i(t)$ $\forall i$ and $u_2'(t)$. From the special selection of the basis vectors in $\mathscr{X}_5$ it is not hard to prove that the impulsive part of u_2' has the following property: $u_{2i}' = [\zeta_1^T\delta(t)\ \zeta_2^T\dot{\delta}(t) \cdots \zeta_{\alpha-1}^T\delta^{\alpha-1}(t)]$. The numerical aspects of the computations are investigated in [6].

Appendix A. Proof of Theorem 2. We start with the system in the form (4)-(5). The idea is now to consider the subsystem

$$\mathscr{J}' = \int_0^\infty (\|\underline{u}_1'\|^2 + \|C_{21}\underline{x}_1 + C_{25}\underline{x}_5\|^2)\, dt$$

with x_5 considered as a control (i.e., as being unconstrained). Obviously

$$\inf \mathscr{J}'(x_{1,0}) \leqq \inf \mathscr{J}(x_{1,0}, x_{2,0}, x_{3,0}, x_{4,0}, x_{5,0}).$$

Since the LQ-problem thus obtained is regular, we can apply Proposition 10. Observe that asymptotic stabilizability of (A, B) implies asymptotic stabilizability of $(A_{11}, (A_{15}^{\dagger}B_{11}))$. The resulting optimal control $(u_1'^*, x_5^*)$ is then given by (9)-(10) and consequently in order to prove statements (i), (ii), and (iii) of Theorem 2 it suffices to show that there always exists a distribution $\underline{u}_2^*$ such that (10) will be satisfied. More explicitly, define $L := -(C_{25}^TC_{25})^{-1}(A_{15}^TP_0 + C_{25}^TC_{21})$. Then we should generate $\underline{x}_5^* = L\underline{x}_1^*$ with $\underline{x}_1^*$ defined by

$$\dot{x}_1^* = (A_{11} + A_{15}L - B_{11}B_{11}^TP_0)\underline{x}_1^*, \qquad \underline{x}_1^*(0) = x_{1,0}.$$

The fact that the desired $\underline{u}_2'^*$ exists is an immediate consequence of Proposition 7.2. Note that since in particular $x_{5,0}$ may be unequal to $-Lx_{1,0}$, we obtain in general distributions for $u_2'^*$ and hence for $\underline{x}_3^*$, $\underline{x}_4^*$, and $\underline{x}_5^*$. The uniqueness claim (iv) of Theorem 2 may be shown as follows. From the original construction of $\underline{x}_5^*$ and $\underline{u}_1'^*$ it follows that they are unique. From Proposition 4 it follows that $\underline{x}_3^*$ is not unique, while from Proposition 7.5 it follows that $\underline{x}_4^*$ and $\underline{x}_5^*$ are unique.

Appendix B. Proof of Theorem 3. We start with the problem in the form (4)-(5)-(11) and consider first the subsystem

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} A_{11} & 0 \\ A_{12} & A_{22} \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} A_{15} \\ A_{25} \end{bmatrix}x_5 + \begin{bmatrix} B_{11} \\ B_{12} \end{bmatrix}x_1'$$

for which we will minimize

$$\mathcal{J}'_{\text{stab}} = \int_0^\infty (\|\underline{u}'_1\|^2 + \|C_{21}\underline{x}_1 + C_{25}\underline{x}_5\|^2)\, dt$$

with x_5 considered as a control and with the constraint $\lim_{t\to\infty}(\underline{x}_1(t), \underline{x}_2(t)) = a$. Since this is a regular problem, we can apply Proposition 9. This yields the optimal trajectory $\underline{x}_1^*, \underline{x}_2^*, \underline{x}_5^*$ which converges to zero at $t \to \infty$. Using the ideas of Appendix A, this will yield Theorem 3 provided we can show that there exists a $\underline{u}_2'^*$ which generates $\underline{x}_3^*, \underline{x}_4^*$ which also converge to zero. This, however, immediately follows from the fact that in Proposition 7.2 a right inverse with a polynomial transfer function is obtained.

REFERENCES

[1] M. L. J. HAUTUS AND L. M. SILVERMAN, *System structure and singular control*, Linear Algebra Appl., 50 (1983), pp. 369-402.

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

[3] J. C. WILLEMS, *Almost invariant subspaces: an approach to high gain feedback design—Part* I: *almost controlled invariant subspaces*, IEEE Trans. Automat. Control., AC-26, 1 (1981), pp. 235-252.

[4] ———, *Almost invariant subspaces: an approach to high gain feedback design—Part* II: *almost conditionally invariant subspaces*, IEEE Trans. Automat. Control, AC-27, 5 (1982), pp. 1071-1084.

[5] J. L. WILLEMS AND J. C. WILLEMS, *Robust stabilization of uncertain systems*, this Journal, 21 (1983), pp. 352-374.

[6] A. KİTAPÇI AND L. M. SILVERMAN, *Determination of Morse's canonical form using the structure algorithm*, presented at 23rd CDC, Las Vegas, NV, 1984.

[7] F. M. CALLIER AND J. L. WILLEMS, *Criterion for the convergence of the solution of the Riccati differential equation*, IEEE Trans. Automat. Control., AC-26, 6 (1981), pp. 1232-1242.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 3, May 1986

ON THE HESSIAN OF LAGRANGIAN AND SECOND ORDER OPTIMALITY CONDITIONS*

S.-P. HAN†

Abstract. For a constrained minimization problem, the restriction of a Hessian of Lagrangian to a tangent space of the feasible set can be used to detect whether a Karush-Kuhn-Tucker point is a local minimum, maximum or saddle point of the problem. It is shown in this paper that the restriction of the Hessian to a normal space with respect to the indefinite inner product induced by the Hessian can be used to characterize a Karush-Kuhn-Tucker point for the Wolfe dual. From this result and by an inertia theorem in [S.-P. Han and O. Fujiwara, *An inertia theorem and its application to nonlinear programs*, manuscript March 1984], we deduce that, under a regularity condition, the Hessian is positive semidefinite if and only if the considered Karush-Kuhn-Tucker point satisfies the second order necessary condition for a local minimum point of the primal problem and, at the same time, satisfies the second order necessary condition for a local maximum point of the dual. Similar results on the positive definiteness of the Hessian are also discussed, which strengthen some results given in [O. Fujiwara, S.-P. Han and O. L. Mangasarian, *Local duality of nonlinear programs*, SIAM J. Control Optim., 22 (1984) pp. 162-169], [S.-P. Han and O. L. Mangasarian, *Characterization of positive definite and semidefinite matrices via quadratic programming duality*, SIAM J. Alg. Disc. Meth., 5 (1984) pp. 26-32].

Key words. nonlinear programming, Wolfe's dual, second order conditions

AMS(MOS) subject classifications. 90C20, 15A63

1. Introduction. The inertia of a symmetric $n \times n$ real matrix H, denoted here by $\pi(H)$, is the triple (ρ, η, θ) where ρ, η and θ are the numbers of positive, negative and zero eigenvalues of H, respectively, with multiplicities counted. When H is the Hessian of a real-valued function at a stationary point, the inertia $\pi(H)$ provides useful information about the stationary point being a local maximum, minimum or saddle point of the function. In the case of the constrained optimization problem:

$$\text{(P)} \qquad \begin{aligned} &\text{Minimize } f(x) \\ &\text{subject to } h(x) = 0 \end{aligned}$$

where $f: R^n \to R$ and $h: R^n \to R^k$ are twice continuously differentiable at a Karush-Kuhn-Tucker point x^*, such characterization of the Karush-Kuhn-Tucker point x^* and its Lagrange multiplier v^* can be done through the Hessian of the Lagrangian $L(x, v) := f(x) + v^T h(x)$. To be more specific, let H be the $n \times n$ Hessian matrix of the Lagrangian L with respect to x at (x^*, v^*) and let $S := \ker(J)$ where J is the $k \times n$ Jacobian matrix of h at x^*. Notice that S is just the tangent space of the constraint surface $h(x) = 0$ at x^* when the Jacobian J has full row rank. The relative inertia $\pi(H/S)$ of H with respect to S is defined to be the triple:

$$\pi(H/S) := (\rho(H/S), \eta(H/S), \theta(H/S))$$

where $\rho(H/S)$, $\eta(H/S)$ and $\theta(H/S)$ are the numbers of positive, negative and zero eigenvalues of the matrix $B^T H B$, respectively, and B is any matrix with its columns forming a basis of the subspace S. The notation is justified because it follows from Sylvester's law of inertia that the triple $\pi(H/S)$ is independent of the choice of a basis for the subspace S. Recall that the Karush-Kuhn-Tucker point x^* and its Lagrange

* Received by the editors April 26, 1984, and in revised form December 15, 1984. This material is based on work supported in part by the National Science Foundation under grant DMS 8203603.

† Department of Mathematics, University of Illinois, Urbana, Illinois 61801.

multiplier v^* satisfy the second order sufficient condition for a local minimum point of the problem (P) if H is positive definite on S [5]; that is,

$$x^T H x > 0 \quad \text{for any nonzero } x \text{ in } S. \tag{1.1}$$

This condition can be paraphrased in terms of the inertia $\pi(H/S)$ as follows:

$$\pi(H/S) = (\dim(S), 0, 0). \tag{1.2}$$

Similarly, under a constraint qualification, the standard second order necessary condition in [5] is that H is positive semidefinite on S; that is,

$$x^T H x \geqq 0 \quad \text{for any } x \in S. \tag{1.3}$$

This can also be written simply as:

$$\eta(H/S) = 0. \tag{1.4}$$

Consequently, when $\rho(H/S)$ and $\eta(H/S)$ are positive and a constraint qualification is satisfied, x^* is only a saddle point of (P). A more complete characterization of the Karush-Kuhn-Tucker point x^* by the relative inertia $\pi(H/S)$ is possible by further imposing a nondegeneracy condition such as the one given in [3].

For the Wolfe dual [11], [12] of (P):

$$\text{(D)} \qquad \begin{aligned} &\text{Maximize } L(x, v) \\ &\text{subject to } \nabla_x L(x, v) = 0, \end{aligned}$$

it will be shown that the Karush-Kuhn-Tucker pair (x^*, v^*) can be characterized also as a stationary point of the dual problem (D) by the inertia $\pi(H/(HS)^\perp)$. The subspace $(HS)^\perp$ is given by:

$$(HS)^\perp = \{x : x^T H y = 0 \text{ for any } y \in S\}$$

and is the orthogonal complement of S with respect to the indefinite inner product defined by $\langle x, y\rangle_H := x^T H y$ (see, for example, [1], [9]). When S is the tangent space of the constraint surface, the set $(HS)^\perp$ can be viewed as its normal space with respect to the inner product at x^*. In this sense, the Karush-Kuhn-Tucker pair (x^*, v^*) can be analyzed as a solution to both the primal and the dual problems by restricting the Hessian H to the tangent space S and the normal space $(HS)^\perp$, respectively.

We further use an inertia theorem in [6] to show that, under a regularity condition, the Hessian matrix H is positive semidefinite if and only if the Karush-Kuhn-Tucker pair (x^*, v^*) satisfies the second order necessary conditions for a local minimum point of the primal problem (P) and, at the same time, satisfies the second order necessary condition for a local maximum point of the dual problem (D). Similar results on the positive definiteness of the Hessian are also given, which strengthen some results in [4], [8].

For the simplicity of our presentation, we only consider the equality constraints here. We note that, since all the results are local, we can extend them to an inequality constrained problem by treating active inequality constraints as equalities and under the conditions of the strict complementarity and the linear independence of the gradients of active constraints.

In § 2, we characterize a Karush-Kuhn-Tucker point for Problem (D) by the inertia $\pi(H/(HS)^\perp)$. In § 3, we analyze the definiteness of the Hessian matrix H through the second order optimality conditions of (P) and (D).

2. Inertia of Hessian. For the Wolfe dual (D), the Lagrangian is given by:

$$\Psi(x, v, w) := L(x, v) + w^T \nabla_x L(x, v).$$

It is known [11], [12] that a Karush-Kuhn-Tucker pair (x^*, v^*) of Problem (P) is also a Karush-Kuhn-Tucker point of Problem (D) with its associated Lagrange multiplier $w^* = 0$. For this reason, we will consider only the Karush-Kuhn-Tucker point (x^*, v^*) of (D) that has a Lagrange multiplier $w^* = 0$. To characterize such a vector $(x^*, v^*, 0)$ as a solution of (D), we need the $(n+k)\times(n+k)$ Hessian matrix of Ψ with respect to (x, v) at $(x^*, v^*, 0)$; that is,

$$M = \begin{pmatrix} H & J^T \\ J & 0 \end{pmatrix} \tag{2.1}$$

where the matrices H and J are defined as in § 1. In this case the counterpart of the set S for (D) at (x^*, v^*) is given by:

$$T := \{(x, y): Hx + J^T y = 0\} \tag{2.2}$$

which is the tangent space of the feasible set of (D) at (x^*, v^*) when the rows of the matrix $[H \quad J^T]$ are linearly independent. Therefore, as in the primal case, the Karush-Kuhn-Tucker point (x^*, v^*) of (D) and its Lagrange multiplier $w^* = 0$ can be analyzed by the inertia $\pi(M/T)$. Similar to conditions (1.2) and (1.4), the second order sufficient condition for (x^*, v^*) to be a local maximum point of the dual problem (D) is:

$$\pi(M/T) = (0, \dim(T), 0), \tag{2.3}$$

and, under a constraint qualification such as the matrix $[H \quad J^T]$ having full rank, the second order necessary condtion is:

$$\rho(M/T) = 0. \tag{2.4}$$

It will be shown below that conditions (2.3) and (2.4) can, in turn, be rewritten in terms of the inertia $\pi(H/(HS)^\perp)$. For establishing this result, we first deduce a relation between the two subspaces S and T.

LEMMA 2.1. *Let T be the subspace in R^{n+k} defined by* (2.2) *and S be the kernel of the matrix J; then (x, y) is in T if and only if*

$$x \in (HS)^\perp \quad \textit{and} \quad y \in -(J^T)^\dagger Hx + \ker(J^T)$$

where $(J^T)^\dagger$ is the Moore-Penrose inverse of J^T.

Proof. For any x in R^n, let $y^* := -(J^T)^\dagger Hx$. We first note that $S^\perp = \text{image}(J^T)$ and if Hx is in $S^\perp$ then $Hx + J^T y^* = 0$.

If $(x, y) \in T$ then $Hx = -J^T y \in S^\perp$. Hence, we have that $x \in (HS)^\perp$ and $Hx = -J^T y = -J^T y^*$. Therefore, it follows that $x \in (HS)^\perp$ and $y \in y^* + \ker(J^T)$.

Conversely, if $x \in (HS)^\perp$ and $y \in y^* + \ker(J^T)$, then $Hx \in S^\perp = \text{image}(J^T)$ and $Hx + J^T y^* = 0$, which implies that $Hx + J^T y = 0$. □

We can now relate the inertia $\pi(H/(HS)^\perp)$ to the inertia $\pi(M/T)$ as follows.

THEOREM 2.2. *Let M be a $(n+k)\times(n+k)$ matrix of the form* (2.1), *S be the kernel of the matrix J, and T be the subspace determined by* (2.2); *then*

$$\pi(M/T) = \pi(-H/(HS)^\perp) + (0, 0, m)$$

where $m = \dim(\ker(J^T))$.

Proof. Let X be a matrix with its columns forming a basis of $(HS)^\perp$ and let $Y := -(J^T)^\dagger HX$. Hence, we have the $HX + J^T Y = 0$. We also let Z be a matrix with

its columns being a basis of $\ker(J^T)$, and let

$$P := \begin{pmatrix} X & 0 \\ Y & Z \end{pmatrix}.$$

Then by Lemma 2.1 the columns of P form a basis of the subspace T. Therefore, the inertia $\pi(M/T) = \pi(P^T M P)$. Now, by using the equations $HX + J^T Y = 0$ and $J^T Z = 0$ we have that

$$P^T M P = \begin{pmatrix} Y^T J X & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} -X^T H X & 0 \\ 0 & 0 \end{pmatrix}.$$

Thus, we have that

$$\pi(M/T) = \pi(-X^T H X) + (0, 0, m) = \pi(-H/(HS)^{\perp}) + (0, 0, m). \qquad \square$$

COROLLARY 2.3. *Let M, T, S and m be defined as in Theorem* 2.2. *If H is nonsingular then* $\pi(M/T) = \pi(-H^{-1}/S^{\perp}) + (0, 0, m)$.

Proof. Let X be a matrix with its columns forming a basis of $(HS)^{\perp}$; then by the nonsingularity of H we have that the columns of HX form a basis of the subspace $S^{\perp}$. The corollary then follows directly from $X^T H X = (HX)^T H^{-1} (HX)$. $\square$

With Theorem 2.2 we now can characterize a Karush-Kuhn-Tucker point for the dual problem (D) as follows.

THEOREM 2.4. *If (x^*, v^*) is a Karush-Kuhn-Tucker point of the dual problem* (D) *with Lagrange multiplier $w^* = 0$, then*

(1). *the vector $(x^*, v^*, 0)$ satisfies the second order sufficient condition* (2.3) *for a local maximum point of* (D) *if and only if the Jacobian J of h at x^* has full row rank and*

$$\pi(H/(HS)^{\perp}) = (\dim((HS)^{\perp}), 0, 0);$$

(2). *the vector $(x^*, v^*, 0)$ satisfies the second order necessary condition* (2.4) *for a local maximum point of* (D) *if and only if*

$$\eta(H/(HS)^{\perp}) = 0.$$

Proof. We notice that it follows from Theorem 2.2 that

$$\begin{aligned} \pi(M/T) &= \pi(-H/(HS)^{\perp}) + (0, 0, m) \\ &= (\eta(H/(HS)^{\perp}), \rho(H/(HS)^{\perp}), \theta(H/(HS)^{\perp}) + m), \end{aligned}$$

where $m = \dim(\ker(J^T))$. The results follow immediately from (2.3) and (2.4). $\square$

The following corollary is a direct consequence of Theorem 2.4 and Corollary 2.3.

COROLLARY 2.5. *Let x^* and v^* be defined as in Theorem* 2.4. *The vector $(x^*, v^*, 0)$ satisfies the second order necessary condition* (2.4) *for a local maximum point of the dual problem* (D) *if and only if*

$$x^T H x \geqq 0 \quad \textit{for any } x \in (HS)^{\perp}. \tag{2.5}$$

If, in addition, H is nonsingular then condition (2.5) *is equivalent to*

$$x^T H^{-1} x \geqq 0 \quad \textit{for any } x \in S^{\perp}. \tag{2.6}$$

Because $\ker(H) \subset (HS)^{\perp}$, the positive definiteness of H on the subspace $(HS)^{\perp}$ implies the nonsingularity of H. Therefore, from Corollary 2.3, we can rewrite the dual second order sufficient condition in terms of the inverse of the Hessian H as in the following corollary, which is a result also given in [4].

COROLLARY 2.6. *Let x^* and v^* be defined as in Theorem* 2.4. *The vector $(x^*, v^*, 0)$ satisfies the second order sufficient condition for a local maximum point of the dual*

problem (D) *if and only if the Jacobian J has full row rank, the Hessian H is nonsingular and*

$$x^T H^{-1} x > 0 \quad \textit{for any nonzero } x \in S^{\perp}. \tag{2.7}$$

We can also characterize a saddle point of (D) as in the following theorem.

THEOREM 2.7. *Let* (x^*, v^*) *be defined as in Theorem* 2.4. *A sufficient condition for* (x^*, v^*) *to be a saddle point of* (D) *is that the matrix* $[H\ J^T]$ *has full rank and*

$$\rho(H/(HS)^{\perp}) > 0 \quad \textit{and} \quad \eta(H/(HS)^{\perp}) > 0. \tag{2.8}$$

Furthermore, condition (2.8) *is also a necessary condition if H is nonsingular on* $(HS)^{\perp}$ *and J has full rank.*

Proof. Under the constraint qualification that the $n \times (n+k)$ matrix $[H\ J^T]$ has full row rank, the vector (x^*, v^*) being a local minimum point implies $\rho(H/(HS)^{\perp}) = 0$; and it being a local maximum point implies $\eta(H/(HS)^{\perp}) = 0$. Therefore, the vector (x^*, v^*) can only be a local saddle point of (D) if (2.8) holds.

Conversely, it follows from Theorem 2.2 that if H is nonsingular on $(HS)^{\perp}$ and J has full rank then

$$\pi(M/T) = \pi(-H/(HS)^{\perp}) \quad \text{and} \quad \theta(M/T) = 0.$$

Therefore, $\eta(H/(HS)^{\perp}) = 0$ implies that (x^*, v^*) is a maximum point; while $\rho(H/(HS)^{\perp}) = 0$ implies that (x^*, v^*) is a minimum point. Hence, condition (2.8) is necessary for (x^*, v^*) to be a saddle point of (D). □

3. Definiteness of Hessian. In § 2 it was shown that the triples $\pi(H/S)$ and $\pi(H/(HS)^{\perp})$ can be used to characterize a solution of the primal problem (P) and the dual problem (D), respectively. Interestingly, by an inertia theorem in [6], $\pi(H/S)$ and $\pi(H/(HS)^{\perp})$ can be closely related to the inertia $\pi(H)$ of the Hessian H itself. We give the theorem below; its proof can be found in [6].

THEOREM 3.1. *Let H be a symmetric* $n \times n$ *matrix and S be a subspace in* R^n *that satisfy the condition*:

$$S \cap (HS)^{\perp} \subset \ker(H); \tag{3.1}$$

then

$$\pi(H) = \pi(H/S) + \pi(H/(HS)^{\perp}) - (0, 0, \dim(S \cap (HS)^{\perp})).$$

As before, let the vector x^* be a Karush-Kuhn-Tucker point of (P) and v^* be its Lagrange multiplier. Consequently, the pair (x^*, v^*) is also a Karush-Kuhn-Tucker point of the dual problem (D) with it Lagrange multiplier $w^* = 0$. Let the matrices H and J also be defined as in § 1. For simplicity, we use (PN), (PS), (DN) and (DS) to denote the primal second order necessary condition (1.3), the primal second order sufficient condition (1.1), the dual second order necessary condition (2.4) and the dual second order sufficient condition (2.3), respectively. Therefore, from the results given in §§ 1 and 2, we have

(PN) $\eta(H/S) = 0,$

(PS) $\pi(H/S) = (\dim(S), 0, 0),$

(DN) $\eta(H/(HS)^{\perp}) = 0,$

(DS) $\pi(H/(HS)^{\perp}) = (\dim(HS)^{\perp}, 0, 0) \quad \text{and} \quad \operatorname{rank}(J) = k.$

We can now use Theorem 3.1 to deduce a result on the positive semidefiniteness of the Hessian H as follows.

THEOREM 3.2. *If the Hessian matrix H satisfies condition* (3.1) *then*

$$\text{(PN) and (DN)} \Leftrightarrow H \text{ is positive semidefinite.}$$

Proof.

$$\begin{aligned}\text{(PN) and (DN)} &\Leftrightarrow \eta(H/S)=0 \text{ and } \eta(H/(HS)^{\perp})=0\\ &\Leftrightarrow \eta(H)=\eta(H/S)+\eta(H/(HS)^{\perp})=0\\ &\Leftrightarrow H \text{ is positive semidefinite.}\end{aligned}$$

□

We note that condition (3.1) is essential in Theorem 3.2. Consider the example given in [4]:

$$\text{Minimize } -x_1x_2$$

$$\text{subject to } x_1=0.$$

The vector $x^*=(0, 1)$ together with $v^*=1$ constitute a Karush-Kuhn-Tucker point. Both the primal and dual second order necessary conditions are satisfied at (x^*, v^*); but the Hessian H, which is given by

$$H=\begin{pmatrix}0 & -1\\ -1 & 0\end{pmatrix},$$

is not positive semidefinite. In this case, condition (3.1) does not hold and the theorem cannot apply.

We also note here that condition (3.1) is closely related to the following often used but more restrictive ccondition in mathematical programming [2], [7], [10]:

$$x^THx=0 \quad \text{and} \quad x\in S \Rightarrow Hx=0. \tag{3.2}$$

It is clear that (3.2) implies (3.1); however, the converse is not true. This can be seen from the following example:

$$H=\begin{pmatrix}0 & 1 & 0\\ 1 & 0 & 0\\ 0 & 0 & 1\end{pmatrix} \qquad S=\{x: x_3=0\}.$$

Obviously, condition (3.1) is satisfied for this particular case; but condition (3.2) is violated at the vector $(1, -1, 0)$. Nevertheless, condition (3.1) is no longer needed when either (PS) or (DS) is assumed.

THEOREM 3.3. (PS) *and* (DN)$\Rightarrow$ *H is positive semidefinite.*

Proof. We first show that (PS) implies condition (3.1). If x is any vector in $S\cap(HS)^{\perp}$ then $x^THx=0$. Notice that if (PS) holds then H is positive definite on S. Therefore, we have $x=0$. This proves $S\cap(HS)^{\perp}=\{0\}$ and condition (3.1) is trivially satisfied. The theorem then follows directly from Theorem 3.2. □

THEOREM 3.4. (PN) *and* (DS)$\Rightarrow$*H is positive definite.*

Proof. By the similar argument as in the proof of the Theorem 3.3, we can show that (DS) implies condition (3.1). Therefore, by Theorem 3.2, we have that H is positive semidefinite. On the other hand, if (DS) holds then, by Corollary 2.6, H is also nonsingular. Thus, H is actually positive definite. □

From Theorems 3.3 and 3.4, we notice that (P) and (D) are not interchangeable and the results are not perfectly symmetric with respect to (P) and (D). This is because

the dual second order sufficient condition is a much stronger condition, which implies the nonsingularity of the Hessian H. Under the nonsingularity assumption on the Hessian H and by the fact that any nonsingular positive semidefinite matrix is actually positive definite, the statement of Theorem 3.4 will still hold when (P) and (D) are interchanged; this is also a result given in [4].

We make a last remark that the implications of Theorems 2.4 and 3.1 are certainly not limited to the characterization of the definiteness of the Hessian H as in Theorems 3.2, 3.3 and 3.4. They can also be used in analyzing a Karush-Kuhn-Tucker point for solving both the primal and the dual problems. For instance, under condition (3.1) and a constraint qualification, if it is known that $\eta(H) > \dim(S)$ then we can immediately conclude that the considered Karush-Kuhn-Tucker point is not a local maximum point of the dual problem (D). Similarly, if $\eta(H) > \dim((HS)^{\perp})$ then the Karush-Kuhn-Tucker point cannot be a local minimum point of the primal problem (P).

Acknowledgment. The author would like to thank one of the referees for his suggestions concerning Theorem 2.7.

REFERENCES

[1] J. BOGNÁR, *Indefinite Inner Product Spaces*, Springer, Berlin, Heidelberg 1974.

[2] R. W. COTTLE AND G. B. DANTZIG, *Complementary pivot theory of mathematical programming*, Linear Algebra Appl., 1 (1968), pp. 103-125.

[3] O. FUJIWARA, *Morse programs: a topological approach to smooth constrained optimization*, I, Math. Oper. Res., 7 (1982), pp. 602-616.

[4] O. FUJIWARA, S.-P. HAN AND O. L. MANGASARIAN, *Local duality of nonlinear programs*, this Journal, 22 (1984), pp. 162-169.

[5] A. V. FIACCO AND G. P. MCCORMICK, *Nonlinear Programming*, Wiley, New York, 1969.

[6] S.-P. HAN AND O. FUJIWARA, *An inertia theorem and its application to nonlinear programs*, manuscript, March 1984. (To appear in Linear Algebra Appl.)

[7] S.-P. HAN AND O. L. MANGASARIAN, *Conjugate cone characterization of positive and semidefinite matrices*, Linear Algebra Appl., 56 (1984), pp. 89-103.

[8] ———, *Characterization of positive definite and semidefinite matrices via quadratic programming duality*, SIAM J. Alg. Disc. Meth., 5 (1984), pp. 26-32.

[9] I. KAPLANSKY, *Linear Algebra and Geometry*, Allyn and Bacon, Boston, 1969.

[10] C. E. LEMKE, *A survey of complementary theory*, in Variational Inequalities and Complementarity Problems, R. W. Cottle, F. Giannessi and J.-L. Lions, eds., Wiley, New York, 1980, pp. 213-239.

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

[12] P. WOLFE, *A duality theorem for nonlinear programming*, Quart. Appl. Math., 19 (1961), pp. 239-244.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 3, May 1986

002

NECESSARY CONDITIONS FOR A DOMAIN OPTIMIZATION PROBLEM IN ELLIPTIC BOUNDARY VALUE PROBLEMS*

NOBUO FUJII†

Abstract. This paper deals with a domain optimization problem suggested by a real physical problem, the shape optimization problem for high-beta plasmas in nuclear fusion research. The function of the space dependent variable, which is the solution of an elliptic boundary value problem defined on a variable domain, should be optimized. The fundamental equation which evaluates the variation of the solution according to the boundary variation is given. From this equation, a variational equation and its equivalent which relate the variation of the solution to the variation of the boundary are derived. This variational equation is exploited to derive a necessary condition for the optimality of the domain. The necessary condition is composed of the Euler-Lagrange equation in the wider sense and the transversality condition. Another form of the necessary condition is also obtained from the equivlaent variational equation.

Key words. domain optimization, variational equation, boundary value problem, partial differential equation

1. Introduction. In research on nuclear fusion the investigation of the high-beta plasma has received attention in these years [4], [7], [11], [13], [20] [21]. In particular, Miller and Moore [11] and others [13] investigated the optimal shape of the cross-section of plasmas in axisymmetric toroids. Here "optimal" means the highest possible β-value, the ratio of the kinetic pressure to the magnetic pressure, of equilibrium plasmas. The β-value varies with the shape of the cross-section as well as the current profile of the plasma ring. Their approaches are intuitive and heuristic, though their results are very valuable. These investigations suggest that we develop a more systematic way for designing the shape of plasma cross-sections.

In this paper we shall consider a domain optimization problem which is suggested by the above physical problem. Our problem to be considered is, however, abstract, simplified, and only a step for solving the real physical problem.

Let us consider an elliptic boundary value problem:

$$\begin{aligned} &\Delta u(x) - (\mathbf{v}(x)\cdot\nabla)u(x) = f(x, u(x)), && x\in\Omega, \\ &u(x) = \kappa(\text{const.}), && x\in\Gamma. \end{aligned} \tag{1.1}$$

Here the symbol Δ denotes the Laplacian operator in R^2, the two-dimensional Euclidean space, and ∇, the gradient operator in R^2. The given nonlinear function $f(x, \xi)$ satisfies suitable conditions which will be clear in the next section. The domain Ω, which is bounded and has a sufficiently smooth boundary Γ, is assumed to be free to vary within a larger domain Ω_0. We want to determine Ω such that the functional $J(\Omega; u)$ defined by

$$J(\Omega; u) = \int_\Omega g(x, u)\, d\Omega \tag{1.2}$$

takes the largest (or smallest) value subject to (1.1) and

$$\int_\Omega h(x)\, d\Omega = M \text{ (const.)}, \tag{1.3}$$

* Received by the editors August 9, 1983, and in revised form January 7, 1985.

† Department of Control Engineering, Faculty of Engineering Science, Osaka University, Toyonaka, Osaka 560, Japan.

where the functions $g(x, \xi)$ and $h(x)$ are suitably defined. Note that the elliptic boundary value problem (1.1) is a generalization of the Grad-Shafranov equation [1, p. 66] which the equilibrium plasma must obey.

Pironneau studied similar problems in fluid mechanics, namely, the minimum-drag problems [15], [16]. In these problems the system of partial differential equations, Stokes or Navier-Stokes equations, is the main constraint, and the objective function is the energy dissipation in the fluid. He derived a necessary optimality condition for each problem with subsidiary constraints; he dealt with not the genuine (classical) but the weak solution of the system of partial differential equations, however. Other authors studied the related or similar problems [9], [17], [18], [19], [22]. Cea [3] enumerated various examples of the domain optimization problems. In particular, Zolesio [22] proposed in the general framework the method for calculating the first variation of objective functions. These authors, except Zolesio, never introduced the variation of the solution of the boundary value problem. Introduction of this notion seems to be natural in the context of the theory of variation.

We shall define the variation of the solution of the boundary value problem and derive a necessary condition for the solution Ω of the domain optimization problem (1.1)-(1.3), assuming the existence of the solution. In order to obtain the necessary condition, we must derive a variational equation which relates the first order variation δu of the solution of (1.1) to the variation δn of the boundary Γ. In § 2, the existence of the well defined δu will be shown first. The equation of variation and its equivalent will, then, be given. Using these results, in § 3 we shall derive the necessary condition for the optimal domain Ω. An equivalent form of the necessary condition will be given as well.

2. Equation of variation. The solution of the boundary value problem (1.1) is, of course, dependent on the domain Ω. If the domain varies (slightly), then the solution is forced to vary even in the common part of the domains. In this section, we shall derive an equation of variation, i.e. a relation between the first-order variation δu of the solution and the variation δn of the domain. We shall show an equivalent form of the equation, too. As far as the author knows, these formulae do not seem to be known in the literature. Once these formulae are obtained, it is not so difficult to derive a necessary condition which the optimal domain must satisfy.

Let Ω be a bounded domain in R^2 whose boundary Γ is sufficiently smooth; we henceforth fix it. Consider another domain Ω_0 large enough to contain Ω in its interior. Let the function $f(x, \xi)$ be defined on $\Omega_0 \times R$, continuous in x, and twice continuously differentiable with respect to ξ. Furthermore let us assume that $f(x, \xi)$ satisfies

$$|f(x, \xi)| \leqq c_0, \qquad \frac{\partial f(x, \xi)}{\partial \xi} \geqq 0, \qquad (x, \xi) \in \Omega_0 \times R, \tag{2.1}$$

where c_0 is a constant. The vector valued function $\mathbf{v}(x)$ is assumed to be defined on Ω_0 and is continuously differentiable. We assume that

$$\operatorname{div} \mathbf{v}(x) \leqq 0; \tag{2.2}$$

note that this condition is always satisfied in the Grad-Shafranov equation. Let us define on Γ a continuous function $\rho(s)$ of arclength s which is piecewise continuously differentiable. We erect at each point on Γ the normal and plot on it normal segments $\varepsilon\rho(s)$ such that the positive segment lies on the exterior normal. If ε is sufficiently small, the endpoints of the segments will form a continuous curve Γ_ε. Obviously Γ_ε is piecewise continuously differentiable. Let Ω_ε be the domain bounded by Γ_ε. We obtain

a sequence of domains which converges to Ω as ε tends to zero; we may, hence, assume that each Ω_ε is contained in Ω_0.

Let us consider the following sequence of elliptic boundary value problems:

$$\begin{aligned} &\Delta u_\varepsilon - (\mathbf{v}(x)\cdot\nabla)u_\varepsilon = f(x, u_\varepsilon), && x\in\Omega_\varepsilon, \\ &u_\varepsilon = \kappa, && x\in\Gamma_\varepsilon, \end{aligned} \tag{2.3}$$

and the boundary value problem

$$\begin{aligned} &\Delta u - (\mathbf{v}(x)\cdot\nabla)u = f(x, u), && x\in\Omega, \\ &u = \kappa, && x\in\Gamma. \end{aligned} \tag{2.4}$$

As is well known [5, Chap. IV], [14], each of these nonlinear boundary value problems has, owing to (2.1), a unique solution which is continuous on $\Omega_\varepsilon+\Gamma_\varepsilon(\Omega+\Gamma)$ and is twice continuously differentiable in $\Omega_\varepsilon(\Omega)$. Let us denote the solution of (2.3) by u_ε and denote that of (2.4) by u. What can we say about the difference between u and u_ε? In response to this question, we have the following

THEOREM 1. *Let $\rho(s)$ be a continuous variation of the boundary Γ of Ω and be piecewise continuously differentiable. Then there exists a well defined function $\phi(x)$ which is continuous in Ω such that*

$$u(x)-u_\varepsilon(x)=\varepsilon\phi(x)+o(\varepsilon), \qquad x\in\Omega\cap\Omega_\varepsilon, \tag{2.5}$$

where $o(\varepsilon)$ is a quantity such that $\varepsilon^{-1}o(\varepsilon)\to 0$ $(\varepsilon\to 0)$ and is uniformaly bounded.

If $f(x, u)$ does not depend on u and $\mathbf{v}(x)\equiv 0$, Theorem 1 is immediately proven (see Appendix 1), since we know the following lemma [2], [6, p. 292]:

LEMMA 1. *Let $G(x, y)$ be Green's function on Ω and $G_\varepsilon(x, y)$ be Green's function on Ω_ε. Then we have*

$$\begin{aligned} G(x, y)-G_\varepsilon(x, y) &= \frac{\varepsilon}{2\pi}\oint_\Gamma \frac{\partial G(z, x)}{\partial n_z}\frac{\partial G(z, y)}{\partial n_z}\rho\, d\Gamma_z+\varepsilon^2\gamma(x, y; \varepsilon) \\ &\equiv \varepsilon\delta G(x, y)+\varepsilon^2\gamma(x, y; \varepsilon), \qquad x, y\in\Omega\cap\Omega_\varepsilon, \end{aligned} \tag{2.6}$$

where, of course, $\partial/\partial n_z$ stands for the normal derivative and $\gamma(x, y; \varepsilon)$ is uniformly bounded in $\Omega\cap\Omega_\varepsilon$.

Unfortunately, however, our equations are nonlinear; it is never obvious that (2.5) holds. In the rest of this section we shall prove Theorem 1 step by step. In § 2.1, we shall assume that $\rho(s)$ is nonnegative to show (2.5). In § 2.2, we shall then show (2.5) in the general case. At the same time an equation of variation and its equivalent will be derived.

2.1. Convergence from above. Let $\rho(s)$ be nonnegative, i.e. $\rho(s)\geqq 0$. Then each Ω_ε contains Ω. Furthermore $\{\Omega_\varepsilon\}$ converges monotonically to Ω as ε approaches zero. In view of these, we may assume from the outset that ε is bounded, say $0<\varepsilon\leqq\varepsilon_1$.

Let us define a function $\phi_\varepsilon(x)$ which depends not only on $x\in\Omega$ but also on ε by

$$\phi_\varepsilon(x)=\frac{1}{\varepsilon}(u(x)-u_\varepsilon(x)). \tag{2.7}$$

Now we can assert the following proposition which will play a fundamental role for showing (2.5). In what follows, all the constants do not depend on ε.

PROPOSITION 1. (i) *There exist constants K_1 and K_2 such that*

$$|u_\varepsilon|\leqq K_1, \qquad x\in\Omega_\varepsilon, \tag{2.8}$$

$$|\phi_\varepsilon| < K_2, \qquad x \in \Omega_\varepsilon. \tag{2.9}$$

As a consequence of (2.9)

$$u_\varepsilon \to u \quad \textit{uniformly on } \Omega(\varepsilon \to 0). \tag{2.10}$$

(ii) *There exists a constant* K_3 *such that*

$$|\nabla u_\varepsilon| \leqq K_3, \qquad x \in \Omega_\varepsilon. \tag{2.11}$$

Furthermore,

$$\nabla u_\varepsilon \to \nabla u \quad \textit{for every } x \in \Omega(\varepsilon \to 0). \tag{2.12}$$

Proof. There evidently exists a constant c_1 such that $|u_1(x) - \kappa| \leqq c_1$, $u_1(x)$ being the solution corresponding to Ω_{ε_1}. It is immediate to show that $u_\varepsilon - u_1$ satisfies

$$\Delta(u_\varepsilon - u_1) - \mathbf{v} \cdot \nabla(u_\varepsilon - u_1) = \int_0^1 \frac{\partial f}{\partial \xi}(x, tu_\varepsilon + (1-t)u_1)\, dt (u_\varepsilon - u_1), \qquad x \in \Omega_\varepsilon.$$

Assumption (2.1) and the celebrated maximum principle [5, p. 326] tell us that

$$\max_{x \in \Omega_\varepsilon} |u_\varepsilon - u_1| \leqq \max_{x \in \Gamma_\varepsilon} |u_\varepsilon - u_1| = \max_{\Gamma_\varepsilon} |u_1 - \kappa| \leqq c_1.$$

From this estimate (2.8) follows at once.

Again we easily obtain, for $x \in \Omega$,

$$\Delta \phi_\varepsilon - \mathbf{v}(x) \cdot \nabla \phi_\varepsilon - \left(\int_0^1 \frac{\partial f}{\partial \xi}(x, tu + (1-t)u_\varepsilon)\, dt \right) \phi_\varepsilon = 0,$$

where ϕ_ε is defined by (2.7). From this it follows that

$$\max_{x \in \Omega} |\phi_\varepsilon| \leqq \max_{x \in \Gamma} \left| \frac{1}{\varepsilon}(u - u_\varepsilon) \right|. \tag{2.13}$$

To proceed further, the following lemma is useful. This lemma is a version of the assertion found in Courant and Hilbert [5, p. 370]; the proof of it may hence be omitted.

LEMMA 2. (i) *Let* $V_\varepsilon(x)$ *be a continuous function on* $\Omega_\varepsilon + \Gamma_\varepsilon$ *and be twice continuously differentiable in* Ω_ε. *If* $V_\varepsilon(x) \equiv 0$ *on* Γ_ε, *then there exist constants* c_2 *and* c_3 *such that*

$$\sup_{x \in \Omega_\varepsilon} \left| \frac{\partial V_\varepsilon}{\partial x_i} \right| \leqq c_2 \sup_{\Omega_\varepsilon} |\Delta V_\varepsilon - \mathbf{v}(x) \cdot \nabla V_\varepsilon| + c_3 \sup_{\Omega_\varepsilon} |V_\varepsilon(x)|, \qquad i = 1, 2. \tag{2.14}$$

(ii) *Let* $V(x)$ *be continuous on* $\Omega + \Gamma$ *and be twice continuously differentiable in* Ω. *Then, for any closed subdomain* G *in* Ω, *there exist constants* $\bar{c}_2$ *and* $\bar{c}_3$ *such that*

$$\max_{x \in G} \left| \frac{\partial V}{\partial x_i} \right| \leqq \bar{c}_2 \sup_{\Omega} |\Delta V - \mathbf{v}(x) \cdot \nabla V| + \bar{c}_3 \sup_{\Omega} |V(x)|, \qquad i = 1, 2. \tag{2.15}$$

The constants $\bar{c}_2$ *and* $\bar{c}_3$ *may depend on* G.

Going back to (2.13), we want to estimate the right-hand side. For each point $x(s)$ on Γ, let $x(s) + a\rho(s)$ denote a point of distance $a\rho(s)$ from $x(s)$ along the normal. Thanks to the mean value theorem, we can obtain

$$\begin{aligned} u(x(s)) - u_\varepsilon(x(s)) &= u_\varepsilon(x(s) + \varepsilon\rho(s)) - u_\varepsilon(x(s)) \\ &= \mathbf{n} \cdot \operatorname{grad} u_\varepsilon(x(s) + \alpha\varepsilon\rho(s))\varepsilon\rho(s), \end{aligned}$$

where $\mathbf{n}$ is the unit outward normal and α satisfies $0 < \alpha < 1$. In view of Lemma 2, the

following estimate is easy:

$$|\mathbf{n}\cdot\operatorname{grad} u_\varepsilon(x(s)+\alpha\varepsilon\rho(s))|\leqq 2c_2\sup_{\Omega_\varepsilon}|\Delta u_\varepsilon-\mathbf{v}\cdot\nabla u_\varepsilon|+2c_3\sup_{\Omega_\varepsilon}|u_\varepsilon-\kappa|$$
$$\leqq 2c_2\sup_{\Omega_\varepsilon}|f(x,u_\varepsilon)|+2c_3\{\sup_{\Omega_\varepsilon}|u_\varepsilon|+|\kappa|\}.$$

From (2.8) and the fact that $f(x,\xi)$ is bounded, it readily follows that for a constant c_4

$$\max_\Gamma\left|\frac{1}{\varepsilon}(u-u_\varepsilon)\right|\leqq c_4\max_\Gamma|\rho(s)|.$$

This and (2.13) yield (2.9) at once. From Lemma 2, we easily obtain (2.11) in the same fashion as above.

Applying the second part of Lemma 2 to $V(x)=u-u_\varepsilon$, we obtain

$$\max_{x\in G}|\operatorname{grad}(u-u_\varepsilon)|\leqq 2\bar{c}_2\sup_\Omega|\Delta(u-u_\varepsilon)-\mathbf{v}\cdot\nabla(u-u_\varepsilon)|$$
$$+2\bar{c}_3\sup_\Omega|u-u_\varepsilon|.$$

The second term of the right-hand side converges to zero due to (2.10). As to the first term, we can estimate as follows

$$|\Delta(u-u_\varepsilon)-\mathbf{v}\cdot\nabla(u-u_\varepsilon)|\leqq\left|\int_0^1\frac{\partial f}{\partial\xi}(x,tu+(1-t)u_\varepsilon)\,dt\right||u-u_\varepsilon|.$$

Thus, in view of (2.10), the first term approaches zero as ε tends to zero. Since G is an arbitrary closed subdomain in Ω, (2.12) follows from these facts. Proposition 1 is thereby proved.

We now wish to show (2.5). First we shall show $\phi_\varepsilon(x)$ converges to a function as ε tends to zero. In what follows we can assume that $\kappa=0$ without loss of generality. In fact, by setting $u_\varepsilon=v_\varepsilon+\kappa$ $(u=v+\kappa)$ we have a new boundary value problem with zero boundary data for v_ε (resp. v). Doing this affects neither the definition of ϕ_ε nor the properties of $f(x,\xi)$ stated above.

It is possible to represent u and u_ε by integral equations with the help of Green's functions, i.e.

$$u(x)=-\frac{1}{2\pi}\left\{\int_\Omega G(x,y)(\mathbf{v}\cdot\nabla)_y u(y)\,d\Omega_y\right.$$
$$\left.+\int_\Omega G(x,y)f(y,u)\,d\Omega_y\right\},\qquad x\in\Omega, \tag{2.16}$$

$$u_\varepsilon(x)=-\frac{1}{2\pi}\left\{\int_{\Omega_\varepsilon}G_\varepsilon(x,y)(\mathbf{v}\cdot\nabla)_y u_\varepsilon(y)\,d\Omega_y\right.$$
$$\left.+\int_{\Omega_\varepsilon}G_\varepsilon(x,y)f(y,u_\varepsilon)\,d\Omega_y\right\},\qquad x\in\Omega_\varepsilon. \tag{2.17}$$

From these we easily obtain

$$\phi_\varepsilon(x)=-\frac{1}{2\pi}\int_\Omega\left\{G(x,y)\frac{\partial f(y,u)}{\partial\xi}-\operatorname{div}_y(\mathbf{v}(y)G(x,y))\right\}\phi_\varepsilon\,d\Omega_y$$
$$-\frac{1}{2\pi\varepsilon}\int_\Omega\{G(x,y)-G_\varepsilon(x,y)\}\{f(y,u_\varepsilon)+(\mathbf{v}\cdot\nabla)_y u_\varepsilon\}\,d\Omega_y$$

$$+\frac{1}{2\pi\varepsilon}\int_{\Omega_\varepsilon-\Omega} G_\varepsilon(x,y)\{f(y,u_\varepsilon)+(\mathbf{v}\cdot\nabla)_y u_\varepsilon\}\,d\Omega_y \tag{2.18}$$
$$+\frac{1}{4\pi}\int_\Omega G(x,y)\frac{\partial^2 f(y,u^*)}{\partial\xi^2}\phi_\varepsilon(u-u_\varepsilon)\,d\Omega_y,$$

where u^* lies between u and u_ε.

Let us introduce the integral kernel by

$$K(x,y)\equiv -\frac{1}{2\pi}\left\{G(x,y)\frac{\partial f(y,u)}{\partial\xi}-\mathrm{div}_y\,(\mathbf{v}(y)G(x,y))\right\},\qquad x,y\in\Omega, \tag{2.19}$$

and introduce functions $F(x,\varepsilon)$ etc. by

$$\begin{aligned}F(x,\varepsilon)=&-\frac{1}{2\pi\varepsilon}\int_\Omega\{G(x,y)-G_\varepsilon(x,y)\}\{f(y,u_\varepsilon)+(\mathbf{v}\cdot\nabla)_y u_\varepsilon\}\,d\Omega_y\\&+\frac{1}{2\pi\varepsilon}\int_{\Omega_\varepsilon-\Omega} G_\varepsilon(x,y)\{f(y,u_\varepsilon)+(\mathbf{v}\cdot\nabla)_y u_\varepsilon\}\,d\Omega_y\\&+\frac{1}{4\pi}\int_\Omega G(x,y)\frac{\partial^2 f(y,u^*)}{\partial\xi^2}\phi_\varepsilon(u-u_\varepsilon)\,d\Omega_y\\&\equiv F_1(x,\varepsilon)+F_2(x,\varepsilon)+F_3(x,\varepsilon).\end{aligned} \tag{2.20}$$

Then (2.18) is rewritten as an integral equation on Ω:

$$\phi_\varepsilon(x)=\int_\Omega K(x,y)\phi_\varepsilon(y)\,d\Omega_y+F(x,\varepsilon). \tag{2.21}$$

In order to investigate (2.21), let us examine the properties of the integral kernel $K(x,y)$. Noting that $G(x,y)$ is the Green's function in two-dimensional domain Ω, we can easily conclude that the kernel has the following properties:

(i) $K(x,y)$ is continuous in both variables x and y except at the points $x=y$;

(ii) there exist constants c_5, c_6 and c_7 such that

$$|K(x,y)|\leqq\frac{c_5}{|x-y|}+c_6\ln\frac{c_7}{|x-y|},\qquad x,y\in\Omega. \tag{2.22}$$

Here we used the assumptions for $f(x\cdot\xi)$, $\mathbf{v}(x)$, and used the fact that $G(x,y)$ has a logarithmic singularity. The estimate (2.22) enables us to make an argument along the classical potential theory [8, Chap. 11], [12, Chap. 8]. Thus, we can say that the new integral kernel defined by $K(x,y;\lambda)\equiv\lambda K(x,y)$ has the resolvent kernel $R(x,y;\lambda)$ which satisfies

$$K(x,y;\lambda)=R(x,y;\lambda)-\int_\Omega K(x,z;\lambda)R(z,y;\lambda)\,d\Omega_z,$$

λ being a (complex) parameter. Of course, $R(x,y;\lambda)$ is expressed by [8, Chap. 11], [12, Chap. 8]

$$R(x,y;\lambda)=K(x,y;\lambda)+K_1(x,y;\lambda)+\frac{M_1(x,y;\lambda)}{\eta(\lambda)},\qquad x,y\in\Omega. \tag{2.23}$$

Here $K_1(x,y;\lambda)$ is the first iterated kernel of $K(x,y;\lambda)$, $M_1(x,y;\lambda)$ is continuous in

x, y and analytic in λ, and $\eta(\lambda)$ is an analytic function of λ. The poles of $R(x, y; \lambda)$ are just the poles of M_1/η.

Now we have the following lemma; the proof will be given in Appendix 2.

LEMMA 3. *The resolvent* $R(x, y; \lambda)$ *does not have pole* $\lambda = 1$. *Hence, the kernel* $K(x, y)$ *has the resolvent* $R(x, y) = R(x, y; 1)$.

As is well known in the classical potential theory, the resolvent has the same singularity as that of the kernel $K(x, y)$; more precisely, $R(x, y)$ has the following estimate

$$|R(x, y)| \leqq c_8 \frac{1}{|x-y|} + c_9 \ln \frac{c_{10}}{|x-y|}, \tag{2.24}$$

c_8, c_9 and c_{10} being constants.

Thus, the solution ϕ_ε of the integral equation (2.21) can be expressed by

$$\phi_\varepsilon(x) = F(x, \varepsilon) + \int_\Omega R(x, z) F(z, \varepsilon)\, d\Omega_z. \tag{2.25}$$

If we show that each term of $F(x, \varepsilon)$ converges to a certain function and is bounded, then, thanks to (2.24), $\phi_\varepsilon(x)$ is shown to approach a function.

Let us examine each term of $F(x, \varepsilon)$ in turn. In view of Lemma 1, (2.10) and (2.12) it is obvious that

$$F_1(x, \varepsilon) \to F_0(x) \equiv \frac{1}{2\pi} \int_\Omega \delta G(x, y)\{f(y, u) + (\mathbf{v} \cdot \nabla)_y\}\, d\Omega_y \quad (\varepsilon \to 0), \quad \text{for every } x \in \Omega. \tag{2.26}$$

Since the logarithmic singularities of Green's functions are eliminated by subtraction and u_ε is uniformly bounded, $F_1(x, \varepsilon)$ is uniformly bounded; i.e. for a constant c_{11}

$$|F_1(x, \varepsilon)| \leqq c_{11}. \tag{2.27}$$

Obviously,

$$F_2(x, \varepsilon) \to 0 \quad (\varepsilon \to 0) \quad \text{for every } x \in \Omega, \tag{2.28}$$

because $G_\varepsilon(x, y) = 0$ provided $y \in \Gamma_\varepsilon$. The modulus of $F_2(x, \varepsilon)$ can be bounded by

$$\begin{aligned} |F_2(x, \varepsilon)| &\leqq \frac{1}{\varepsilon}(c_0 + \max_{\Omega_0} |\mathbf{v}(x)| K_3) \int_{\Omega_\varepsilon - \Omega} |G_\varepsilon(x, y)|\, d\Omega_y \\ &\leqq \Big\{ (c_0 + K_3 \max_{\Omega_0} |\mathbf{v}(x)|) c_{12} \ln \frac{c_{13}}{\zeta} \oint_\Gamma \rho(s)\, ds \\ &\qquad + 2c_g \max_\Gamma |\rho(s)|(\zeta - \zeta \ln \zeta) + c_{14} \Big\}, \end{aligned}$$

where c_{12}, c_{13}, c_{14} are constants, ζ is a sufficiently small fixed number, and c_g is a geometric factor dependent on only Ω. This estimate shows that $F_2(x, \varepsilon)$ is uniformly bounded; i.e., for a constant c_{15},

$$|F_2(x, \varepsilon)| \leqq c_{15}. \tag{2.29}$$

Finally, since $\partial^2 f/\partial \xi^2$ is continuous and u_ε converges uniformly to u, we obtain for a constant c_{16}

$$\left| \frac{\partial^2 f}{\partial \xi^2}(y, u^*) \right| \leqq c_{16}.$$

This, (2.9) and (2.10) yield

$$|F_3(x, \varepsilon)| \leqq c_{17} \tag{2.30}$$

and at the same time

$$F_3(x, \varepsilon) \to 0 \quad (\varepsilon \to 0) \quad \text{for } x \in \Omega, \tag{2.31}$$

c_{17} being a constant.

Combining (2.26)-(2.31), we can conclude that

$$\phi_\varepsilon(x) \to \phi(x) \equiv F_0(x) + \int_\Omega R(x, z) F_0(z)\, d\Omega_z, \qquad x \in \Omega. \tag{2.32}$$

In other words, we established (2.5) for the case where every Ω_ε contains Ω.

An easy but tedious examination of the above derivation shows that the term $o(\varepsilon)$ in (2.5) is uniformly bounded.

2.2. Convergence in the general case and the equation of variation. We have just proved (2.5) in the case of nonnegative $\rho(s)$. From this result, it is possible to prove Theorem 1 in the general case, by using the Hadamard's device [6, p. 295], [2]. However this procedure calls for very tedious calculations, since domain dependent quantites appear in this procedure. So we shall prove the theorem along the argument in the former subsection, which will be a great help for reading this subsection.

Let $\rho(s)$ be an arbitrary, piecewise continuously differentiable function. In this case, the boundary curves Γ and Γ_ε may intercross each other. We can assert the following proposition in this case, too. The proof is almost similar to that of Proposition 1; it may therefore be omitted.

PROPOSITION 2. *Let the positive ε be bounded.*

(i) *Then, there exist constants K_4, K_5 and K_6 such that*

$$|u_\varepsilon(x)| \leqq K_4, \quad |\nabla u_\varepsilon| \leqq K_5, \quad |\phi_\varepsilon(x)| \leqq K_6, \qquad x \in \Omega_\varepsilon, \tag{2.33}$$

where, of course, ϕ_ε is defined by (2.7). *The last inequality of* (2.33) *implies that u_ε converges to u uniformly in every compact subdomain of Ω.*

(ii) *For every $x \in \Omega$, ∇u_ε converges to ∇u.*

Let us define Ω_ε^0, Ω_ε^1 and Ω_ε^2 by

$$\Omega_\varepsilon^0 = \Omega \cap \Omega_\varepsilon, \quad \Omega_\varepsilon^1 = \Omega - \Omega_\varepsilon^0, \quad \Omega_\varepsilon^2 = \Omega_\varepsilon - \Omega_\varepsilon^0. \tag{2.34}$$

Obviously, when ε approaches zero,

$$\Omega_\varepsilon^0 \to \Omega, \quad \Omega_\varepsilon^1 \to \varnothing, \quad \Omega_\varepsilon^2 \to \varnothing,$$

$\varnothing$ being the empty set. It is obvious that there exists a positive $\varepsilon(x)$ for any $x \in \Omega$ such that $x \in \Omega_\varepsilon^0$ $(0 < \varepsilon < \varepsilon(x))$. Therefore, in what follows, we shall understand that x belongs to Ω_ε^0 unless otherwise stated. We may assume that $\kappa = 0$ in this subsection, too. From (2.6), (2.16) and (2.17) we can easily obtain

$$u(x) - u_\varepsilon(x) = -\frac{1}{2\pi} \int_{\Omega_\varepsilon^0} \left\{ G(x, y) \frac{\partial f(y, u)}{\partial \xi} - \operatorname{div}_y (\mathbf{v}(y) G(x, y)) \right\} (u - u_\varepsilon)\, d\Omega_y$$

$$-\frac{1}{2\pi} \int_{\Omega_\varepsilon^0} \varepsilon \delta G(x, y) \{ f(y, u) + (\mathbf{v} \cdot \nabla)_y u \}\, d\Omega_y$$

$$-\frac{1}{2\pi} \int_{\Omega_\varepsilon^0} \varepsilon^2 \gamma(x, y; \varepsilon) \{ f(y, u) + (\mathbf{v} \cdot \nabla)_y u \}\, d\Omega_y$$

$$
\begin{aligned}
(2.35)\qquad & -\frac{1}{2\pi}\int_{\Omega_\varepsilon^0}(G-G_\varepsilon)\{f(y,u_\varepsilon)-f(y,u)+(\mathbf{v}\cdot\nabla)_y(u_\varepsilon-u)\}\,d\Omega_y \\
& -\frac{1}{2\pi}\Bigg[\int_{\Omega_\varepsilon^1}G(x,y)\{f(y,u)+(\mathbf{v}\cdot\nabla)_y u\}\,d\Omega_y \\
& \qquad -\int_{\Omega_\varepsilon^2}G_\varepsilon(x,y)\{f(y,u_\varepsilon)+(\mathbf{v}\cdot\nabla)_y u_\varepsilon\}\,d\Omega_y\Bigg] \\
& +\frac{1}{4\pi}\int_{\Omega_\varepsilon^0}G(x,y)\frac{\partial^2 f(y,u^*)}{\partial\xi^2}(u-u_\varepsilon)^2\,d\Omega_y \\
& -\frac{1}{2\pi}\int_{\Gamma_\varepsilon\cap\Omega}G(x,y)u(y)\mathbf{v}\cdot\mathbf{n}\,d\Gamma_y,
\end{aligned}
$$

where u^* lies between u and u_ε. If we introduce $\bar{r}(x,\varepsilon)$ by

$$
\begin{aligned}
& \bar{r}(x,\varepsilon)\equiv\sum_{i=1}^{5}r_i(x,\varepsilon), \\
& r_1(x,\varepsilon)\equiv-\frac{1}{2\pi}\int_{\Omega_\varepsilon^0}\varepsilon\gamma(x,y;\varepsilon)\{f(y,u)+(\mathbf{v}\cdot\nabla)_y u\}\,d\Omega_y, \\
& r_2(x,\varepsilon)\equiv-\frac{1}{2\pi\varepsilon}\int_{\Omega_\varepsilon^0}(G-G_\varepsilon)\{f(y,u_\varepsilon)-f(y,u)+(\mathbf{v}\cdot\nabla)_y(u_\varepsilon-u)\}\,d\Omega_y, \\
(2.36)\qquad & r_3(x,\varepsilon)\equiv-\frac{1}{2\pi\varepsilon}\Bigg[\int_{\Omega_\varepsilon^1}G(x,y)\{f(y,u)+(\mathbf{v}\cdot\nabla)_y u\}\,d\Omega_y \\
& \qquad -\int_{\Omega_\varepsilon^2}G_\varepsilon(x,y)\{f(y,u_\varepsilon)+(\mathbf{v}\cdot\nabla)_y u_\varepsilon\}\,d\Omega_y\Bigg], \\
& r_4(x,\varepsilon)\equiv\frac{1}{4\pi\varepsilon}\int_{\Omega_\varepsilon^0}G(x,y)\frac{\partial^2 f(y,u^*)}{\partial\xi^2}(u-u_\varepsilon)^2\,d\Omega_y, \\
& r_5(x,\varepsilon)\equiv-\frac{1}{2\pi\varepsilon}\int_{\Gamma_\varepsilon\cap\Omega}G(x,y)u(y)\mathbf{v}\cdot\mathbf{n}\,d\Gamma_y,
\end{aligned}
$$

we can obtain, from (2.35) and (2.19),

$$
\begin{aligned}
(2.37)\qquad & \phi_\varepsilon(x)=\int_{\Omega_\varepsilon^0}K(x,y)\phi_\varepsilon(y)\,d\Omega_y \\
& \qquad -\frac{1}{2\pi}\int_{\Omega_\varepsilon^0}\delta G(x,y)\{f(y,u)+(\mathbf{v}\cdot\nabla)_y u\}\,d\Omega_y+\bar{r}(x,\varepsilon).
\end{aligned}
$$

Equation (2.37) can be regarded as an integral equation for ϕ_ε. This integral equation is, however, difficult to be handled, since the domain of definition varies with ε. We shall therefore use a little device instead of dealing directly with (2.37).

Consider the following integral equation for ϕ:

$$
\begin{aligned}
(2.38)\qquad & \phi(x)=\int_{\Omega}K(x,y)\phi(y)\,d\Omega_y \\
& \qquad -\frac{1}{2\pi}\int_{\Omega}\delta G(x,y)\{f(y,u)+(\mathbf{v}\cdot\nabla)_y u\}\,d\Omega_y, \qquad x\in\Omega.
\end{aligned}
$$

It is seen from Lemma 3 that this integral equation admits a unique solution $\phi(x)$ which is a continuous function on Ω. Let $\Phi_\varepsilon(x)$ be defined by

$$\Phi_\varepsilon(x) \equiv \phi(x) - \phi_\varepsilon(x). \tag{2.39}$$

Subtracting (2.37) from (2.38), we obtain the following integral equation for $\Phi_\varepsilon(x)$:

$$\begin{aligned}\Phi_\varepsilon(x) = &\int_{\Omega_\varepsilon^0} K(x, y)\Phi_\varepsilon(y)\, d\Omega_y + \int_{\Omega-\Omega_\varepsilon^0} K(x, y)\phi(y)\, d\Omega_y \\ &- \frac{1}{2\pi}\int_{\Omega-\Omega_\varepsilon^0} \delta G(x, y)\{f(y, u) + (\mathbf{v}\cdot\nabla)_y u\}\, d\Omega_y - \bar{r}(x, \varepsilon).\end{aligned} \tag{2.40}$$

Let $r(x, \varepsilon)$ be given by

$$\begin{aligned} r(x, \varepsilon) &= -\bar{r}(x, \varepsilon) + r_6(x, \varepsilon) + r_7(x, \varepsilon), \\ r_6(x, \varepsilon) &= \int_{\Omega-\Omega_\varepsilon^0} K(x, y)\phi(y)\, d\Omega_y, \\ r_7(x, \varepsilon) &= -\frac{1}{2\pi}\int_{\Omega-\Omega_\varepsilon^0} \delta G(x, y)\{f(y, u) + (\mathbf{v}\cdot\nabla)_y u\}\, d\Omega_y. \end{aligned} \tag{2.41}$$

Integral equation (2.40) is then rewritten as

$$\Phi_\varepsilon(x) = \int_{\Omega_\varepsilon^0} K(x, y)\Phi_\varepsilon(y)\, d\Omega_y + r(x, \varepsilon). \tag{2.42}$$

If the resolvent $R_\varepsilon(x, y)$ of $K(x, y)$ on Ω_ε^0 exists, the solution of (2.42) can be given by

$$\Phi_\varepsilon(x) = r(x, \varepsilon) + \int_{\Omega_\varepsilon^0} R_\varepsilon(x, z) r(z, \varepsilon)\, d\Omega_z. \tag{2.43}$$

Let us show that $R_\varepsilon(x, y)$ really exists provided that ε is small enough. Consider the resolvent $R(x, y; \lambda; \varepsilon)$ of $K(x, y; \lambda) \equiv \lambda K(x, y)$ on Ω_ε^0. By the discussion similar to that in § 2.1, we see that $R(x, y; \lambda; \varepsilon)$ is expressed by

$$R(x, y; \lambda; \varepsilon) = K(x, y; \lambda) + K_1(x, y; \lambda; \varepsilon) + \frac{M(x, y; \lambda; \varepsilon)}{\eta(\lambda; \varepsilon)}. \tag{2.44}$$

Here, $K_1(x, y; \lambda; \varepsilon)$ is the first iterated kernel of $K(x, y; \lambda)$ on Ω_ε^0, and $M(x, y; \lambda; \varepsilon)$ is continuous in both x and y. The analytic function $\eta(\lambda; \varepsilon)$ is given by a series each term of which is the iterated integral of the Fredholm determinant on Ω_ε^0 [8, Chap. 11], [12, Chap. 8]. As is easily seen, the Fredholm determinant is continuous with respect to ε, and hence the integral of it is continuous in ε. While the series is uniformly convergent, $\eta(\lambda; \varepsilon)$ is continuous in ε. Since $\eta(\lambda; \varepsilon)$ is not only continuous in ε but also analytic with respect to λ, it follows immediately from Lemma 3 that $M(x, y; \lambda; \varepsilon)/\eta(\lambda; \varepsilon)$ does not have pole $\lambda = 1$ provided that ε is small enough. This means that $R_\varepsilon(x, y)$ exists and is given by $R_\varepsilon(x, y) = R(x, y; 1; \varepsilon)$ if ε is sufficiently small. Moreover, it is easy to show that there exist constants c_{17}, c_{18} and c_{19} such that

$$|R_\varepsilon(x, y)| \leqq c_{17}\frac{1}{|x-y|} + c_{18}\ln\frac{c_{19}}{|x-y|}. \tag{2.45}$$

In view of Proposition 2, it is possible to show that $r(x, \varepsilon)$ is uniformly bounded and converges to zero; more precisely, for a constant K_7 and for every $x \in \Omega$,

$$|r(x, \varepsilon)| \leqq K_7 \quad \text{and} \quad r(x, \varepsilon) \to 0 \qquad (\varepsilon \to 0). \tag{2.46}$$

In fact, it is obvious from Lemma 1 and Proposition 2 that $r_1(x, \varepsilon)$, $r_2(x, \varepsilon)$ are bounded and approach zero, respectively. As for $r_3(x, \varepsilon)$ and $r_4(x, \varepsilon)$, we can employ the same argument as that in § 2.2; hence, they are bounded and converge to zero. Noting that $G(x, y)$, $u(y)$ are zero at $y \in \Gamma$ and that $\Gamma_\varepsilon \cap \Omega$ is close to Γ, we can conclude that $r_5(x, \varepsilon)$ is bounded and approaches zero. Finally, it is easy to see that $r_6(x, \varepsilon)$, $r_7(x, \varepsilon)$ are bounded and converge to zero, since $K(x, y)$, $\delta G(x, y)$ are integrable and the measure of $\Omega - \Omega_\varepsilon^0$ approaches zero. Thus, (2.46) has been established.

From (2.43), (2.45) and (2.46), it readily follows that $\Phi_\varepsilon \to 0$ $(\varepsilon \to 0)$. This means that $\phi_\varepsilon \to \phi$ $(\varepsilon \to 0)$ for every $x \in \Omega$; i.e. we have proved (2.5) for arbitrary $\rho(s)$. Again, it is possible to show that the term $o(\varepsilon)$ in (2.5) is uniformly bounded in this case, too. The proof of Theorem 1 is thereby completed.

Now we wish to derive the relation between the first variation $\phi(x)$ of the solution and the boundary variation $\rho(s)$. In what follows and in the next section, we shall rewrite $\rho(s)$ as $\delta n(s)$ (or simply δn) and rewrite $\phi(x)$ as $\delta u(x)$ (or δu). As an immediate corollary to Theorem 1, we have

COROLLARY 1. *The variation $\delta u(x)$ is related to $\delta n(s)$ through*

$$\delta u(x) = \int_\Omega K(x, y)\delta u(y)\, d\Omega_y - \frac{1}{2\pi} \oint_\Gamma \frac{\partial u(y)}{\partial n_y} \delta n(y) \frac{\partial G(x, y)}{\partial n_y}\, d\Gamma_y, \qquad x \in \Omega. \tag{2.47}$$

Proof. We give only the outline of the proof. Using a fundamental solution $U(x, y)$ of the Laplacian, we can represent $u = u_\varepsilon$ by an integral equation similar to (2.18). The formula (2.5) helps us to obtain an integral equation for $\delta u(x)$ by the limiting process. After replacing $U(x, y)$ by $G(x, y)$ we obtain (2.47). This completes the proof.

Note that we can obtain (2.47) also from (2.38) by changing the order of integration, and that (2.47) admits a unique solution owing to Lemma 3.

From Corollary 1, we can easily derive an equivalent characterization for $\delta u(x)$.

COROLLARY 2. *The variation $\delta u(x)$ is the solution of the following boundary value problem:*

$$\begin{aligned} &(\Delta - \mathbf{v}(x) \cdot \nabla)\delta u(x) = \frac{\partial f(x, u)}{\partial \xi} \delta u(x), && x \in \Omega, \\ &\delta u(x) = \frac{\partial u(x)}{\partial n} \delta n(x), && x \in \Gamma. \end{aligned} \tag{2.48}$$

Proof. First note that (2.48) has a unique solution owing to (2.1). On the other hand, the solution of (2.47) is unique. Thus the solution of (2.48) is just the variation $\delta u(x)$ determined by (2.47). This completes the proof.

3. Multiplier rule. In this section we shall derive a necessary condition, which is, so to speak, a multiplier rule, for optimality of the domain in the optimization problem posed in § 1, assuming that the problem has a solution. This condition is based on (2.47). Another form of the necessary condition will be obtained from (2.48), too.

Let the domain Ω with the sufficiently smooth boundary Γ be an optimal domain. Let $u(x)$ be the corresponding solution of (1.1). Thanks to (2.5), it is possible to calculate the first order variation δJ of the functional J defined by (1.2). In fact, let δn be the boundary variation with the properties stated in the previous section and let δu be the corresponding variation of u. It is easy, with the help of (2.5), to obtain

$$\delta J = \int_\Omega \frac{\partial g(x, u)}{\partial \xi} \delta u\, d\Omega_x + \oint_\Gamma g(x, u)\delta n\, d\Gamma_x, \tag{3.1}$$

where we supposed that $g(x, \xi)$ is continuous in x and continuously differentiable in ξ.

Let us define a function $\lambda(x)$ by

$$\lambda(x) - \int_\Omega \lambda(y) K(y, x)\, d\Omega_y = \frac{\partial g(x, u)}{\partial \xi}, \qquad x \in \Omega \tag{3.2}$$

In view of Lemma 3, (3.2) has the unique solution $\lambda(x)$ for given $u(x)$; the equation (3.2), thus, really defines the continuous function $\lambda(x)$. Substitution of (3.2) into (3.1) yields

$$\delta J = \int_\Omega \lambda(x) \left\{ \delta u(x) - \int_\Omega K(x, y) \delta u(y)\, d\Omega_y \right\} d\Omega_x + \oint_\Gamma g(x, u) \delta n\, d\Gamma_x. \tag{3.3}$$

Again, substituting (2.47) into (3.3), we obtain

$$\delta J = \oint_\Gamma \left\{ \frac{1}{2\pi} \frac{\partial u(x)}{\partial n_x} \left(\int_\Omega \lambda(y) \frac{\partial G(y, x)}{\partial n_x}\, d\Omega_y \right) + g(x, u) \right\} \delta n(x)\, d\Gamma_x. \tag{3.4}$$

Thus $\delta u(x)$ has been eliminated from δJ. From the constraint (1.3), $\delta n(x)$ should satisfy

$$\oint_\Gamma h(x) \delta n(x)\, d\Gamma_x = 0. \tag{3.5}$$

Combining (3.4) and (3.5), we can conclude that there exists a constant λ_0 such that

$$\frac{1}{2\pi} \frac{\partial u(x)}{\partial n} \left(\int_\Omega \lambda(y) \frac{\partial G(y, x)}{\partial n_x}\, d\Omega_y \right) + g(x, u) = \lambda_0 h(x), \qquad x \in \Gamma. \tag{3.6}$$

In summary, we obtained the following necessary condition for the optimality of Ω.

THEOREM 2. *If Ω is an optimal domain and $u(x)$ is the corresponding solution of* (1.1), *then there exist a continuous function $\lambda(x)$ and a constant λ_0 such that*

$$\lambda(x) - \int_\Omega \lambda(y) K(y, x)\, d\Omega_y = \frac{\partial g(x, u)}{\partial \xi}, \qquad x \in \Omega, \tag{3.2}$$

$$\frac{1}{2\pi} \frac{\partial u(x)}{\partial n} \left(\int_\Omega \lambda(y) \frac{\partial G(y, x)}{\partial n_x}\, d\Omega_y \right) + g(x, u) = \lambda_0 h(x), \qquad x \in \Gamma. \tag{3.6}$$

We may call (3.2) an Euler-Lagrange equation in the wider sense. Equation (3.6) is the transversality condition.

From (2.48) we can obtain another form of the necessary condition.

THEOREM 3. *If Ω is an optimal domain and $u(x)$ is the corresponding solution of* (1.1), *then there eixst a function $p(x)$ and a constant λ_0 such that*

$$\begin{aligned} &\Delta p(x) + \operatorname{div}(\mathbf{v}(x) p(x)) - \frac{\partial f(x, u)}{\partial \xi} p(x) = \frac{\partial g(x, u)}{\partial \xi}, \qquad x \in \Omega, \\ &p(x) = 0, \qquad x \in \Gamma, \end{aligned} \tag{3.7}$$

$$-\frac{\partial u}{\partial n} \frac{\partial p}{\partial n} + g(x, u) = \lambda_0 h(x), \qquad x \in \Gamma. \tag{3.8}$$

Proof. Boundary value problem (3.7) admits a unique solution $p(x)$ due to (2.2). Substitution of (3.7) into (3.1) and integration by parts yield

$$\delta J = \oint_{\Gamma} \left\{ -\frac{\partial u}{\partial n}\frac{\partial p}{\partial n} + g(x, u) \right\} \delta n(x)\, d\Gamma_x. \tag{3.9}$$

From this and (3.5) we obtain (3.8). The proof is thereby completed.

We derived a necessary condition and its equivalent, which will play an important role in the investigation of the properties of the optimal domain as well as in the development of numerical methods for the optimization problem.

Recently, Zolesio [22] has pointed out that the first variation of the objective function takes the form similar to (3.9) in the case of linear boundary value problems, assuming the existence of $\delta u(x)$. As far as the author knows, no author showed the existence of $\delta u(x)$, especially in the case of the nonlinear boundary value problem.

4. Concluding remarks. We have derived the variational equations (2.47) and (2.48), each of which relates the first order variation δu of the solution of the elliptic boundary value problem to the variation of the domain. These equations are based on the relation (2.5), which also plays a fundamental role for deriving the first order variation δJ of the functional J. From the variational equation (2.47) and the expression for δJ a necessary condition for optimality of the domain Ω is obtained. This condition is composed of the Euler-Lagrange equation in the wider sense and the transversality condition. Another equivalent form of the necessary condition is derived from (2.48), too.

These necessary conditions are expected to be used in the investigation of the properties of the optimal domain and to be useful for developing the numerical method to seek the optimal domain. Our results are, however, only the first step for dealing with the real physical problem stated in § 1. The application of our results to this physical problem will be reported in a further paper.

From the theoretical point of view, there remain various problems. Namely, the existence of an optimal domain should be made clear. Lions [10] gives an existence result for a domain optimization problem which is different from that considered in this paper. We must give a sufficient condition, too. Further necessary conditions will be important for the investigation of the properties of the optimal domain as well as for the development of the numerical method. Our method and results are to be extended to a wider class of problems including optimal designs of vibrating plates or rods in structural mechanics and even to the dynamical problems described by, e.g., parabolic partial differential equations.

Appendix 1. Let us prove (2.5) for the case where $f(x, u)$ does not depend on u and $\mathbf{v}(x) \equiv 0$. Let u and u_ε be the solutions of the boundary value problems:

$$\begin{aligned} &\Delta u = f(x), && x \in \Omega, \\ &u = \kappa, && x \in \Gamma, \end{aligned} \tag{A.1}$$

$$\begin{aligned} &\Delta u_\varepsilon = f(x), && x \in \Omega_\varepsilon, \\ &u = \kappa, && x \in \Gamma_\varepsilon, \end{aligned} \tag{A.2}$$

respectively. Also in this case we can assume that $\kappa = 0$ without loss of generality. The solutions u and u_ε are expressed by

$$u(x) = -\frac{1}{2\pi}\int_G G(x, y) f(y)\, d\Omega_y, \tag{A.3}$$

$$u_\varepsilon(x) = -\frac{1}{2\pi}\int_{\Omega_\varepsilon} G_\varepsilon(x, y) f(y)\, d\Omega_y, \tag{A.4}$$

respectively. Let Ω_ε^0, Ω_ε^1 and Ω_ε^2 be as in the text (§ 2.2). Subtracting (A.4) from (A.3), we readily obtain

$$\begin{aligned} u(x) - u_\varepsilon(x) = &-\frac{1}{2\pi}\int_{\Omega_\varepsilon^0} (G - G_\varepsilon) f(y)\, d\Omega_y \\ &-\frac{1}{2\pi}\left[\int_{\Omega_\varepsilon^1} G(x, y) f(y)\, d\Omega_y - \int_{\Omega_\varepsilon^2} G_\varepsilon(x, y) f(y)\, d\Omega_y\right]. \end{aligned}$$

From this and Lemma 1, it follows that

$$\begin{aligned} \phi_\varepsilon(x) = &-\frac{1}{2\pi}\int_{\Omega_\varepsilon^0} \delta G(x, y) f(y)\, d\Omega_y \\ &-\frac{1}{2\pi}\varepsilon\int_{\Omega_\varepsilon^0} \gamma(x, y; \varepsilon) f(y)\, d\Omega_y \\ &-\frac{1}{2\pi\varepsilon}\left[\int_{\Omega_\varepsilon^1} G(x, y) f(y)\, d\Omega_y - \int_{\Omega_\varepsilon^2} G_\varepsilon(x, y) f(y)\, d\Omega_y\right]. \end{aligned} \tag{A.5}$$

Let ε approach zero in (A.5). It is easy to see that

$$-\frac{1}{2\pi}\int_{\Omega_\varepsilon^0} \delta G(x, y) f(y)\, d\Omega_y \to -\frac{1}{2\pi}\int_{\Omega} \delta G(x, y) f(y)\, d\Omega_y.$$

The second term of the right-hand side of (A.5) obviously tends to zero. As for the third term, we can see by the same reasoning as in the text that it approaches zero. These facts show that (2.5) for the case where $f(x, u)$ does not depend on u and $\mathbf{v}(x) \equiv 0$.

Appendix 2.

Proof of Lemma 3. We shall give the outline of the proof. Suppose that $\lambda = 1$ were a pole of $R(x, y; \lambda)$. Then there would be a function $\phi(x)$, continuous, not identically zero, which satisfies

$$\phi(x) = \int_\Omega K(x, y; 1)\phi(y)\, d\Omega_y = \int_\Omega K(x, y)\phi(y)\, d\Omega_y. \tag{A.6}$$

From the definition (2.19) of $K(x, y)$ and by integration by parts, (A.6) turns out to be

$$\phi(x) = -\frac{1}{2\pi}\int_\Omega G(x, y)\left\{\frac{\partial f(y, u)}{\partial \xi}\phi + (\mathbf{v}\cdot\nabla)_y\phi\right\} d\Omega_y. \tag{A.7}$$

Thus, the function $\phi(x)$ would satisfy

$$\Delta\phi(x) - (\mathbf{v}\cdot\nabla)\phi(x) - \frac{\partial f(x, u)}{\partial \xi}\phi(x) = 0, \qquad x \in \Omega. \tag{A.8}$$

By letting $x \to z \in \Gamma$ in (A.7), the boundary values of $\phi(x)$ would be zero, i.e.,

$$\phi(x) = 0, \qquad x \in \Gamma. \tag{A.9}$$

In view of (2.1), the solution of the boundary value problem (A.8), (A.9) is unique; $\phi(x) \equiv 0$ is the solution. This is contrarry to the assumption, and hence $\lambda = 1$ is not a pole of $R(x, y; \lambda)$. Lemma 3 is thereby proved.

Acknowledgments. The author thanks the reviewers for their critical readings and helpful comments. He wishes to express his thanks to Professor Y. Sakawa of Osaka University and to Mr. A. Myslinski of System Research Institute of Polish Academy of Science for their comments.

REFERENCES

[1] G. BATEMAN, MHD *Instabilities*, MIT Press, Cambridge, MA 1978.

[2] S. BERGMAN AND M. SCIFFER, *A representation of Green's and Neumann's functions in the theory of partial differential equations of second order*, Duke Math. J., 14 (1947), pp. 609-638.

[3] J. CEA, *Problems of shape optimal design*, in Optimization of Distributed Parameter Structures, vol. 2, E. J. Haug and J. Cea, eds., Sijthoff and Noordhoff, Alphen aan den Rijn, 1980, pp. 1005-1048.

[4] J. W. CONNER, R. J. HASTIE AND J. B. TAYLOR, *Shear, periodicity and plasma ballooning modes*, Phys. Rev. Lett., 40 (1978), pp. 396-399.

[5] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics*, Vol. 2, Interscience, New York, 1962.

[6] R. COURANT, *Dirichlet's Principle, Conformal Mapping, and Minimal Surfaces*, Interscience, New York, 1950.

[7] D. DOBROT et al., *Theory of ballooning modes in tokamaks with finite shear*, Phys. Rev. Lett., 39 (1977), pp. 943-946.

[8] O. D. KELLOG, *Foundations of Potential Theory* (republication), Dover, New York, 1954.

[9] M. KODA, *Optimum design in distributed parameter systems*, Preprints of IFAC 3rd Symposium-Control of Distributed Parameter Systems, J. P. Babary and L. LeLetty, eds., IFAC, Toulouse, 1982, pp. XIV.1-XIV.6.

[10] J. L. LIONS, *Some Aspects of Optimal Control of Distributed Parameter Systems*, CBMS Regional Conference Series in Applied Mathematics 6, Society for Industrial and Applied Mathematics, Philadelphia, 1972.

[11] R. L. MILLER AND R. W. MOORE, *Shape optimizatiion of tokamak plasmas to localized magneto-hydrodynamic modes*, Phys. Rev. Lett., 43 (1979), pp. 765-768.

[12] S. MIZOHATA, *The Theory of Partial Differential Equations*, Cambridge Univ. Press, Cambridge, 1973.

[13] R. W. MOORE et al., *Optimization and control of plasma shape and current profile in noncircular cross-section tokamaks*, in Plasma Physics and Controlled Nuclear Fusion Research 1980, Nuclear Fusion Supplement 1981, IAEA, Vienna, 1981, pp. 283-290.

[14] M. NAGUMO, *On principally linear elliptic differential equations of the second order*, Osaka Math. J., 6 (1954), pp. 207-229.

[15] O. PIRONNEAU, *On optimum problems in Stokes flow*, J. Fluid Mech., 59 (1973), pp. 117-128.

[16] ———, *On optimum design in fluid mechanics*, J. Fluid Mech., 64 (1974), pp. 97-110.

[17] B. ROUSSELET, *Réponse dynamique et optimisation de domaine*, Preprints of IFAC 3rd Symposium-Control of Distributed Parameter Systems, J. P. Babary and L. LeLetty, eds., IFAC, Toulouse, 1982, pp. XII.14-XII.17.

[18] ———, *Shape design sensitivity of a membrane*, J. Optim. Theory Appl., 40 (1983), pp. 595-623.

[19] C. SAGUEZ, *Contrôle optimal d'un système gouverné par une inéquation variationelle parabolique. Observation du domaine de contact*, C.R. Acad. Sc. Paris, 287 (1978), pp. 957-959.

[20] H. R. STRAUSS et al., *Stability of high-beta tokamaks to ballooning modes*, Nuclear Fusion, 20 (1980), pp. 638-641.

[21] K. T. TSANG AND D. J. SIGMAR, *Stabilization of ballooning modes by energetic particles in tokamaks*, Nuclear Fusion, 21 (1981), pp. 1227-1233.

[22] J. P. ZOLESIO, *The material derivative (or speed) method for shape optimization*, in Optimization of Distributed Parameter Structures, vol. 2, E. J. Haug and J. Cea, eds., Sijthoff and Noordhoff, Alphen aan den Rijn, 1980, pp. 1089-1151.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 3, May 1986

003

DIFFERENTIAL GAMES OF GENERALIZED PURSUIT AND EVASION*

LEONARD D. BERKOVITZ†

Abstract. Differential games of generalized pursuit and evasion are studied by comparing them with differential games of fixed duration, for which a theory already has been established. It is shown that if the Isaacs condition holds and the data satisfy reasonable hypotheses, then the games have values which are continuous functions of the initial time and state. If the data satisfy appropriate Lipschitz conditions, then the value is Lipschitz continuous and satisfies the Isaacs equation at all points of differentiability of the value.

Key words. differential games, generalized pursuit and evasion, value, Isaacs equation

1. Introduction. In [1] we developed a theory of differential games of fixed duration in which the existence of value and saddle point in the presence of the Isaacs condition was obtained by relatively elementary methods, without recourse to arguments involving stochastic processes and partial differential equations. In [1], the definition of strategy followed that of Friedman [2], [3], while the definition of payoff followed that of Krasovskii and Subbotin [4]. In this paper we shall study differential games of generalized pursuit and evasion under these notions of strategy and payoff. We assume that the reader is familiar with [1] and shall freely use concepts, notations, definitions, etc. introduced there.

Formulated intuitively, the game that we shall study here has its state $x(t)$ at time t determined by the system of differential equations

$$\frac{dx}{dt} = f(t, x, u(t), v(t)), \qquad x(t_0) = x_0, \tag{1.1}$$

where $u(t) \in Y$ is chosen by Player I at each instant of time t and $v(t) \in Z$ is chosen by Player II at each instant of time t. The payoff is

$$\int_{t_0}^{t_f} f^0(s, \phi(s), u(s), v(s))\, ds, \tag{1.2}$$

where ϕ is the solution of (1.1) and t_f is the first time that $(t, \varphi(t))$ reaches some preassigned terminal set $\mathcal{T}$. Note that if we adjoin a coordinate x^0 to x and a differential equation $dx^0/dt = f^0(t, x, u(t), v(t))$ to (1.1), then if $\hat{x} = (x^0, x)$, we can write the payoff as $\varphi^0(t_f)$.

We shall show that under appropriate conditions on the data of the problem, if the Isaacs condition holds, then the game of generalized pursuit and evasion has a continuous value. We shall obtain these results by comparing our game with a family of games of fixed duration, adapting the idea introduced in Friedman [3] to our definition of the game. Many details will be simpler than in [3]. We shall also show that if we further require the data of the problem to be Lipschitz, then the value function will be Lipschitz continuous and will satisfy the Isaacs equation at points of differentiability.

2. Assumptions and notation. As noted in the introduction we shall use the notation established in [1, § 2]. In addition, we shall let $\hat{x} = (x^0, x)$ and shall let $\hat{f} = (f^0, f^1, \cdots, f^n)$. Concerning $\hat{f}$, we assume the following.

* Received by the editors June 27, 1984, and in revised form December 18, 1984.
† Department of Mathematics, Purdue University, West Lafayette, Indiana 47907.

Assumption I. Statements (i) to (iii) of Assumption I of [1] hold with f replaced by $\hat{f}$.

(iv): $f^0(t, x, y, z) \geqq 0$ for all (t, x, y, z) in $\mathscr{D}$, where $\mathscr{D}$ is as in (2.1) of [1].

In the sequel we shall only need to consider the function f^0 evaluated at points (t, x, y, z) of the form $(t, \varphi(t), u(t), v(t))$ where $\varphi(\cdot) = \varphi(\cdot, t_0, x_0, u, v)$ is an nth state trajectory of (1.1) corresponding to the controls u and v. It follows from (i)-(iii) of Assumption I that for all initial conditions (t_0, x_0) in a given compact set in (t, x)-space, the set of all possible nth stage trajectories resulting from all possible strategies is such that all points $(t, \varphi(t), u(t), v(t))$ lie in a compact subset of $\mathscr{D}$. Hence, since f^0 is continuous, f^0 is bounded below on this set by some constant $-C$, $C > 0$. Hence there is no loss of generality in assuming that $f^0(t, x, y, z) \geqq 0$ on $\mathscr{D}$, for we may replace f^0 by $f^0 + C$.

Assumption I′. Statements (i)-(iii) are as in Assumption I′ of [1] with f replaced by $\hat{f}$.

Assumption II is concerned with the terminal set.

Assumption II. Let F_1 be a closed domain in $(T_0, \infty) \times R^n$ with $C^{(2)}$ boundary ∂F_1. At each point $(t, x) \in \partial F_1$ let

$$\min_z \max_y [\nu_0 + \langle \nu, f(t, x, y, z) \rangle] < 0, \tag{2.1}$$

where (ν_0, ν) is the normal to ∂F_1 at (t, x) pointing to the exterior of F_1, the min is taken over $z \in Z$ and the max over y in Y. Let F denote the intersection of F_1 with the slab $T_0 \leqq t \leqq T$, and let F be bounded. Let the terminal set $\mathscr{T}$ be given by $\mathscr{T} = F \cup ([T, \infty) \times R^n)$.

Assumption II′. Let F_1 be as in Assumption I and let F_1 further satisfy the condition $F_1 \supset ([T, \infty) \times R^n)$. Let $F = F_1$ and let $\mathscr{T} = F$.

3. Definition of game. Let (t_0, x_0) with $(t_0, x_0) \notin \mathscr{T}$ be an initial point of the game. Let Γ be a strategy for Player I on the interval $[t_0, T]$ and Δ a strategy for Player II on the interval $[t_0, T]$, where strategies are defined as in [1]. Corresponding to (Γ, Δ) we obtain a sequence of controls (u_n, v_n) and a sequence of nth stage trajectories $\hat{\varphi}_n(\cdot, t_0, x_0, u_n, v_n)$, where $\hat{\varphi}_n = (\varphi_n^0, \varphi_n)$ satisfies

$$\begin{aligned} \varphi_n^{0\prime}(t) &= f^0(t, \varphi_n(t), u_n(t), v_n(t)), \qquad \varphi_n^0(t_0) = x_n^0, \\ \varphi_n'(t) &= f(t, \varphi_n(t), u_n(t), v_n(t)), \qquad \varphi_n(t_0) = x_n, \end{aligned} \tag{3.1}$$

where $(x_n^0, x_n) \to (0, x_0)$. Note that to determine φ_n we must solve a system of differential equations. Once φ_n is determined, the zeroth component φ_n^0 is determined by a quadrature. A motion $\hat{\varphi}[\cdot, t_0, x_0, \Gamma, \Delta]$ is a limit of nth stage trajectories $\hat{\varphi}_n(t_0, x_{0n}, u_n, v_n)$, where (u_n, v_n) are the controls resulting from (Γ_n, Δ_n).

Let $\mathscr{T}$ be either as in Assumption II or Assumption II′. By the *capture time* $t_{f,n}$ of the nth stage trajectory $\hat{\varphi}_n(\cdot)$, we mean the first time such that $(t, \varphi_n(t))$ belongs to $\mathscr{T}$. More precisely, let $C_{t_0,n} = \{t\colon t_0 < t \leqq T\colon (t, \varphi_n(t)) \in \mathscr{T}\}$. Clearly $C_{t_0,n}$ is not empty. Let $t_{f,n} = \inf\{t\colon t \in C_{t_0,n}\}$. The *capture time* t_f of a motion $\hat{\varphi}[\cdot, t_0, x_0, \Gamma, \Delta]$ is defined similarly. Note that all points $(t_{f,n}, \varphi_n(t_{f,n}))$ and $(t_f, \varphi[t_f])$ belong to $\mathscr{T}$.

We now define the payoff $P(t_0, x_0, \Gamma, \Delta)$ resulting from a pair of strategies (Γ, Δ) as follows:

$$P(t_0, x_0, \Gamma, \Delta) \equiv \cup\, \varphi^0[t_f, t_0, x_0, \Gamma, \Delta], \tag{3.2}$$

where the union is taken over all motions $\hat{\varphi}[\cdot, t_0, x_0, \Gamma, \Delta]$ resulting from (Γ, Δ). Note that t_f in general will be different for different motions. Having defined the strategies

and payoff, the definition of the game is complete. We shall designate this game as "the game G" or "the game $G(t_0, x_0)$", when we wish to call attention to the initial position. We shall let $W^+(t_0, x_0)$ denote the upper value of $G(t_0, x_0)$ and let $W^-(t_0, x_0)$ denote the lower value of $G(t_0, x_0)$. Thus:

$$W^+(t_0, x_0) = \inf_{\Delta} \sup_{\Gamma} P(t_0, x_0, \Gamma, \Delta)\, W^-(t_0, x_0) = \sup_{\Gamma} \inf_{\Delta} P(t_0, x_0, \Gamma, \Delta).$$

4. Comparison with game of fixed duration. We make two simple observations that we shall use. First, since $f^0 \geqq 0$, we have that for any nth stage trajectory $\hat{\varphi}_n(\cdot)$, the zeroth component $\varphi_n^0(\cdot)$ is a nondecreasing function. Similarly, the zeroth component $\varphi^0[\cdot]$ of any motion $\hat{\varphi}[\cdot]$ is nondecreasing. The second observation we state as a Lemma.

LEMMA 4.1. *Let $\{\hat{\varphi}_n(\cdot)\}$ be a sequence of nth stage trajectories converging uniformly to a motion $\hat{\varphi}[\cdot]$. Let $\{t_{f,n}\}$ be the corresponding sequence of capture times for the trajectories and let t_f be the capture time of $\hat{\varphi}[\cdot]$. Then* $\liminf_n \varphi_n^0(t_{f,n}) \geqq \varphi^0[t_f]$. *A similar conclusion holds for any sequence of motions $\{\hat{\varphi}_n[\cdot]\}$ with capture times $\{t_{f,n}\}$ converging uniformly to a motion $\varphi[\cdot]$ with capture time t_f.*

We prove the first statement; the proof of the second is similar. Let $\lambda = \liminf_n \varphi_n^0(t_{f,n})$. Then there exists a subsequence, which we again label as $\hat{\varphi}_n(\cdot)$, such that $\varphi_n^0(t_{f,n}) \to \lambda$ and $t_{f,n} \to \bar{t}$ for some $\bar{t}$ in $[t_0, T]$. Since $\hat{\varphi}_n(\cdot)$ converges uniformly to $\hat{\varphi}[\cdot]$, we have that $\hat{\varphi}_n(t_{f,n}) \to \hat{\varphi}[\bar{t}]$. Since each point $(t_{f,n}, \varphi_n(t_{f,n}))$ belongs to $\mathcal{T}$, and $\mathcal{T}$ is closed, we have that $(\bar{t}, \varphi[\bar{t}]) \in \mathcal{T}$. Hence by the definition of t_f, $\bar{t} \geqq t_f$, and by the monotonicity of φ^0, $\lambda = \varphi^0[\bar{t}] \geqq \varphi^0[t_f]$, which is the desired conclusion.

COROLLARY. *Let $\{t_{f,n}\}$ and t_f be as in the lemma. Then* $\liminf_n t_{f,n} \geqq t_f$.

Let Assumption II hold. Let $\hat{\partial}F$ denote the boundary points of F that are not interior to F_1, i.e.

$$\hat{\partial}F = (\partial F) \cap (\partial F_1).$$

For a point (t, x) with $T_0 \leqq t \leqq T$, let $\rho(t, x)$ denote the signed distance of (t, x) to $\hat{\partial}F$, with negative values assigned to points (t, x) in the interior of F.

For $\varepsilon > 0$, let $(\hat{\partial}F)_\varepsilon = \{(t, x)\colon T_0 \leqq t \leqq T,\ |\rho(t, x)| < \varepsilon\}$. Then since $\hat{\partial}F$ is a $C^{(2)}$ manifold and since F is bounded, it follows that there exists an $\varepsilon_0 > 0$ such that ρ is C' on $(\hat{\partial}F)_{\varepsilon_0}$. Moreover, if (τ, ξ) is a point of $\hat{\partial}F$ with $\tau < T$ and if (ν^0, ν) is the unit normal to $\hat{\partial}F$ at (τ, ξ) pointing to the exterior of F, then

$$\lim_{(t,x)\to(\tau,\xi)} (\rho_t(t, x), \rho_x(t, x)) = (\nu^0, \nu). \tag{4.1}$$

Let Assumption II′ hold. For a point (t, x) let $\rho(t, x)$ now denote the signed distance of (t, x) to ∂F, with negative values assigned to points interior to F. Let $(\partial F)_\varepsilon = \{(t, x)\colon |\rho(t, x)| < \varepsilon\}$. Then for each $R > 0$ there exists an $\varepsilon_0 > 0$ such that ρ is C' in $(\partial F)_{\varepsilon_0} \cap \{(t, x)\colon |x| < R\}$ and (4.1) holds for all (τ, ξ) in $(\partial F)_{\varepsilon_0} \cap \{(t, x)\colon |x| < R\}$.

Let Assumption II or II′ hold. Let $F_\mu = \{(t, x)\colon (t, x) \in F, |\rho(t, x)| < \mu\}$. If Assumption II holds, there exists a $\mu_0 > 0$ such that for all $0 < \mu < \mu_0$, $F - F_\mu \neq \phi$. If Assumption II′ holds, there exists a μ_0 such that for all $0 < \mu < \mu_0$, $(F - F_\mu) \cap \{(t, x)\colon |x| < R\} \neq \phi$.

We now define a family of games of fixed duration. Let Assumption II or II′ hold. Let $\gamma(r) = 1 - r$ if $0 \leqq r \leqq 1$ and let $\gamma(r) = 0$ if $r > 1$. For each $0 < \mu < \mu_0$ let

$$f_\mu^0(t, x, y, z) = \begin{cases} f^0(t, x, y, z)\gamma(|\rho(t, x)|/\mu) & \text{if } (t, x) \in F, \\ f^0(t, x, y, z) & \text{if } (t, x) \notin F. \end{cases}$$

The function f_μ^0 is continuous on $[T_0, T_1] \times R^n \times Y \times Z \times (0, \mu_0)$, and $f_\mu^0(t, x, y, z) = 0$ if $(t, x) \in F - F_\mu$. For each μ in $(0, \mu_0)$ we consider the game G_μ of fixed duration with

terminal time T, with dynamics given by (1.1) and with payoff

$$P_\mu(t_0, x_0, \Gamma, \Delta) = \int_{t_0}^{T} f_\mu^0(s, \varphi(s), u(s), v(s))\, ds.$$

If we introduce a zeroth coordinate x^0 by means of the state equation $dx^0/dt = f_\mu^0(t, x, y, z)$, then the games of fixed duration G_μ will be in the format of [1] with $g(\hat{x}) = x^0$. It is readily verified that Assumption I or I′ of [1] holds for the games G_μ whenever the corresponding assumption holds for the generalized pursuit evasion game G.

Let $W^+(t_0, x_0, \mu)$ and $W^-(t_0, x_0, \mu)$ denote the upper and lower values respectively of the games G_μ, $0 < \mu < \mu_0$, with initial position (t_0, x_0). Let

$$R = (([T_0, \infty) \times R^n) - \mathcal{T}) \cup (\partial \mathcal{T}). \tag{4.2}$$

LEMMA 4.2. *Let Assumption* I *and either Assumption* II *or* II′ *hold. Then for each* (t_0, x_0) *in* R *and each* $0 < \mu < \mu_0$

$$W^-(t_0, x_0) \leqq W^-(t_0, x_0, \mu), \qquad W^+(t_0, x_0) \leqq W^+(t_0, x_0, \mu).$$

For $(t_0, x_0) \in \partial \mathcal{T}$ the result is immediate, since $W^-(t_0, x_0)$ and $W^+(t_0, x_0)$ are equal to zero, and $f_\mu^0 \geqq 0$. We henceforth suppose that $(t_0, x_0) \notin \partial \mathcal{T}$.

Let $\hat{\varphi}[\cdot, t_0, x_0, \Gamma, \Delta]$ be a motion in the game $G(t_0, x_0)$ and let $\{\hat{\varphi}_n(\cdot, t_0, x_0, u_n, v_n)\}$ be the sequence of nth stage trajectories converging uniformly to $\hat{\varphi}[\cdot]$. The sequence $\hat{\varphi}_n(\cdot)$ determines a sequence $\{\hat{\varphi}_{\mu,n}(\cdot, t_0, x_0, u_n, v_n)\}$ of nth stage trajectories in the game $G_\mu(t_0, x_0)$ where

$$\begin{aligned}
\varphi_{\mu,n}(t) &= \varphi_n(t), \qquad t_0 \leqq t \leqq T, \\
\varphi_{\mu,n}^0(t) &= \varphi_n^0(t), \qquad t_0 \leqq t \leqq t_{f,n}, \\
&= \varphi_n^0(t_{f,n}) + \int_{t_{f,n}}^{t} f_\mu^0(s, \varphi_n(s), u_n(s), v_n(s))\, ds
\end{aligned}$$

for $t_{f,n} \leqq t \leqq T$. The capture time of $\hat{\varphi}_{\mu,n}(\cdot)$ is obviously also $t_{f,n}$. There exists a subsequence of $\{\hat{\varphi}_{\mu,n}(\cdot)\}$, which we again label as $\{\hat{\varphi}_{\mu,n}(\cdot)\}$ which converges uniformly to a motion $\hat{\varphi}_\mu[\cdot, t_0, x_0, \Gamma, \Delta]$ in the game G_μ, where

$$\varphi_\mu[t, t_0, x_0, \Gamma, \Delta] = \varphi[t, t_0, x_0, \Gamma, \Delta], \qquad t_0 \leqq t \leqq T.$$

Thus $\hat{\varphi}_\mu[\cdot]$ and $\hat{\varphi}[\cdot]$ have the same capture time t_f.

Let $t_n = \min(t_f, t_{f,n})$. Using the corollary to Lemma 4.1, we conclude that $t_n \to t_f$. For $t_0 \leqq t \leqq t_n$, $\varphi_{\mu,n}^0(t) = \varphi_n^0(t)$ and for $t_n \leqq t \leqq T$

$$\varphi_{\mu,n}^0(t) = \varphi_n^0(t_n) + \int_{t_n}^{t} f_\mu^0(s, \varphi_n(s), u_n(s), v_n(s))\, ds.$$

If we now set $t = T$ in the preceding equation and then let $n \to \infty$, we get that $\varphi_\mu^0[T] = \varphi^0[t_f] + E$, where $E \geqq 0$. To summarize, we have shown that for every motion $\hat{\varphi}[\cdot, t_0, x_0, \Gamma, \Delta]$ in the game G, there is a motion $\hat{\varphi}_\mu[\cdot, t_0, x_0, \Gamma, \Delta]$ (same Γ, Δ) in the game G_μ such that

$$\varphi^0[t_f, t_0, x_0, \Gamma, \Delta] \leqq \varphi_\mu^0[T, t_0, x_0, \Gamma, \Delta]. \tag{4.3}$$

In a similar fashion we can show that for every motion $\hat{\varphi}_\mu[\cdot, t_0, x_0, \Gamma, \Delta]$ in the game G_μ there is a motion $\hat{\varphi}[\cdot, t_0, x_0, \Gamma, \Delta]$ in the game G such that (4.3) holds. Hence for

each fixed Γ,

$$\inf_{\Delta} P(t_0, x_0, \Gamma, \Delta) \leqq \inf_{\Delta} P_\mu(t_0, x_0, \Gamma, \Delta),$$

and for each fixed Δ

$$\sup_{\Gamma} P(t_0, x_0, \Gamma, \Delta) \leqq \sup_{\Gamma} P_\mu(t_0, x_0, \Gamma, \Delta).$$

The conclusion of the lemma follows from these inequalities and from the definitions of W^+ and W^-.

LEMMA 4.3. *Let Assumption* I *and either Assumption* II *or* II′ *hold. Let $\mathscr{C}$ be a compact set in R. Then there exists a constant $C>0$ such that for all μ sufficiently small and all (t_0, x_0) in $\mathscr{C}$*

$$W^-(t_0, x_0) \geqq W^-(t_0, x_0, \mu) - C\mu.$$

The proof of the lemma involves an adaptation to our definition of the game of ideas used by Friedman [2], [3]. The proof will be given in several steps and will be carried out under the supposition that $(t_0, x_0) \notin \partial\mathscr{T}$ and that Assumption II holds. The reader will see that the proof is also valid for $(t_0, x_0) \in \partial\mathscr{T}$ if he keeps in mind that if $(t_0, x_0) \in \partial\mathscr{T}$ then $W^-(t_0, x_0) = 0$, that $t_f = t_0$, and consequently that $\varphi[t_f, t_0, x_0, \Gamma, \Delta] = x_0$, $\varphi^0[t_f, t_0, \Gamma, \Delta] = 0$. He should also look upon the strategy Δ^* to be defined below as being a strategy over $[t_0, T]$ in the game $G_\mu(t_0, x_0)$, rather than a modification of Δ. He will have to change the definition of $\Delta^*_{n,1}$ to be any control u_1 in I_1. The proof if Assumption II′ holds is then the same as that if Assumption II holds, except that for $\hat{\partial}F$ one should read ∂F.

We noted earlier that there exists an $\varepsilon_0 > 0$ such that ρ is C' on $(\hat{\partial}F)_{\varepsilon_0}$. Define a function on $(\hat{\partial}F)_{\varepsilon_0} \times Y \times Z$ as follows:

$$\Lambda(t, x, y, z) = \frac{\partial \rho}{\partial t}(t, x) + \sum_{i=1}^{n} \frac{\partial \rho}{\partial x^i}(t, x) f^i(t, x, y, z). \tag{4.4}$$

Let

$$\Lambda_1(t, x) = \min_z \max_y \Lambda(t, x, y, z),$$

where the min is taken over Z and the max over Y. From the continuity of (ρ_t, ρ_x) on $(\hat{\partial}F)_{\varepsilon_0}$, from the compactness of $\hat{\partial}F$, from (2.1) and from (4.1) it follows that there exists an $\varepsilon_1 < \varepsilon_0$ and a constant $A > 0$ such that

$$\Lambda_1(t, x) \leqq -A_1 \qquad (t, x) \in (\hat{\partial}F)_{\varepsilon_1}. \tag{4.5}$$

Let $\Lambda_0(t, x, z) = \max_y \Lambda(t, x, y, z)$ and let $z^*(t, x)$ be an element in Z such that

$$\Lambda_0(t, x, z^*(t, x)) = \min_z \Lambda_0(t, x, z) = \min_z \max_y \Lambda(t, x, y, z)$$

$$= \Lambda_1(t, x).$$

Now,

$$\Lambda(t, x, y, z^*(t, x)) \leqq \max_y \Lambda(t, x, y, z^*(t, x)) = \Lambda_0(t, x, z^*(t, x)).$$

Thus, for all $(t, x) \in (\hat{\partial}F)_{\varepsilon_0}$ and y in Y

$$\Lambda(t, x, y, z^*(t, x)) \leqq \Lambda_1(t, x). \tag{4.6}$$

The function Λ is uniformly continuous on the closure of $(\hat{\partial}F)_{\varepsilon_1} \times Y \times Z$. Hence there exists an $\varepsilon_2 < \varepsilon_1$ and a $\delta > 0$ such that for all (t, x) in $(\hat{\partial}F)_{\varepsilon_2}$ and all (t', x') with

$|(t', x')-(t, x)|<\delta$ we have $(t', x')\in(\hat{\partial}F)_{\varepsilon_1}$ and $|\Lambda(t', x', y, z)-\Lambda(t, x, y, z)|\leqq A/2$ for all $y\in Y$ and $z\in Z$. From this and from (4.6) and (4.5) we get that for all (t, x) in $\operatorname{cl}[(\hat{\partial}F)_{\varepsilon_2}]$ and all $|(t', x')-(t, x)|<\delta$,

$$\Lambda(t', x', y, z^*(t, x))\leqq\Lambda(t, x, y, z^*(t, x))+A/2\leqq -A/2. \tag{4.7}$$

Since ρ is C' on $(\hat{\partial}F)_{\varepsilon_0}$, it follows that for any absolutely continuous function θ whose graph lies in $(\hat{\partial}F)_{\varepsilon_0}$ the mapping $t\to\rho(t, \theta(t))$ is absolutely continuous and

$$\frac{d}{dt}\rho(t, \theta(t))=\rho_t(t, \theta(t))+\langle\rho_x(t, \theta(t)), \theta'(t)\rangle \quad \text{a.e.}$$

(See [2, Lemma 3.2.2].) In particular if φ is a trajectory of (1.1), then

$$\frac{d}{dt}\rho(t, \varphi(t))=\Lambda(t, \varphi(t), u(t), v(t)) \quad \text{a.e.} \tag{4.8}$$

as long as the trajectory lies in $(\hat{\partial}F)_{\varepsilon_0}$.

LEMMA 4.4. *Let (Γ, Δ) be a pair of strategies and let $\{\hat{\varphi}_n(\cdot)=\hat{\varphi}_n(\cdot, t_0, x_{0n}, \Gamma, \Delta)\}$ be a corresponding sequence of nth stage trajectories. Let $\bar{t}\in(t_0, T)$. Then there exists a strategy Δ^*, depending on Δ and $\bar{t}$, and a sequence $\varepsilon(n)\to 0$ as $n\to\infty$ such that if $\{\hat{\varphi}_n^*(\cdot)\}=\{\hat{\varphi}_n^*(\cdot, t_0, x_{0n}, u_n^*, v_n^*)\}$ is the sequence of nth state trajectories resulting from (Γ, Δ^*) and $\{x_{0n}\}$, then the following are true.*

(i) *$\hat{\varphi}_n^*(t)=\hat{\varphi}_n(t)$ for $t_0\leqq t\leqq\bar{t}$.*

(ii) *If $\{t_n\}$ is a sequence converging to $\bar{t}$ and if for all sufficiently large n, $t_n\leqq\bar{t}$ and $(\bar{t}, \varphi_n(\bar{t}))\in(\hat{\partial}F)_{\varepsilon_2}$, then the inequality*

$$\rho(t, \varphi_n^*(t))-\rho(t_n, \varphi_n^*(t_n))\leqq\varepsilon(n)-A(t-t_n)/2 \tag{4.9}$$

holds for these values of n and for all t such that the graph of φ_n^* on the interval $[\bar{t}, t]$ lies in $(\hat{\partial}F)_{\varepsilon_2}$.

Let $\{\pi_n\}$ be a sequence of partitions of $[t_0, T]$ and let π_n have partition points $t_0=\tau_0<\tau_1<\cdots<\tau_n=T$. Let $I_i=[\tau_{i-1}, \tau_i)$, $i=1, \cdots, n$. Let $k=k(n)$ denote the integer such that $\bar{t}\in I_{k+1}$. We define the strategy $\Delta^*=\{\Delta_n^*\}$. Let $\Delta_{n,1}^*=\Delta_{n,1}$. For $2\leqq j\leqq k+1$, let

$$\Delta_{n,j}^*(\mathcal{Y}_1\times\mathcal{Z}_1\times\cdots\times\mathcal{Y}_{j-1}\times\mathcal{Z}_{j-1})=\Delta_{n,j}(\mathcal{Y}_1\times\mathcal{Z}_1\times\cdots\times\mathcal{Y}_{j-1}\times\mathcal{Z}_{j-1}).$$

Let $j>k+1$. For $i=1, \cdots, j-1$, let $u_{n,i}\in\mathcal{Y}_i$, $v_{n,i}\in\mathcal{Z}_i$, and for $t\in I_i$, let $u_n'(t)=u_{n,i}(t)$ and $v_n'(t)=v_{n,i}(t)$. Then u_n' and v_n' are controls on $[t_0, \tau_{j-1})$ and they determine an nth stage trajectory $\hat{\varphi}_n^*(\cdot)=\hat{\varphi}_n^*(\cdot, t_0, x_{0n}, u_n', v_n')$ on $[t_0, \tau_{j-1}]$. We now define $\Delta_{n,j}^*$. If $(\tau_{j-1}, \varphi^*(\tau_{j-1}))\in(\hat{\partial}F)_{\varepsilon_2}$, let

$$(\Delta_{n,j}^*(u_{n,1}, v_{n,1}, \cdots, u_{n,j-1}, v_{n,j-1}))(t)=z^*(\tau_{j-1}, \varphi_n^*(\tau_{j-1}))$$

for all $t\in I_j$. If $(\tau_{j-1}, \varphi_n^*(\tau_{j-1}))\notin(\hat{\partial}F)_{\varepsilon_2}$ let

$$\begin{aligned}&\Delta_{n,j}^*(u_{n,1}, v_{n,1}, \cdots, u_{n,j-1}, v_{n,j-1})\\&\quad=\Delta_{n,j}(u_{n,1}, v_{n,1}, \cdots, u_{n,j-1}, v_{n,j-1}).\end{aligned}$$

By construction $\hat{\varphi}_n^*(t)=\hat{\varphi}_n(t)$ for $t_0\leqq t\leqq\bar{t}$. By assumption $(\bar{t}, \varphi_n(\bar{t}))\in(\hat{\partial}F)_{\varepsilon_2}$ for n sufficiently large, so that for each n there is an interval $[\bar{t}, \bar{t}+\alpha_n)$ on which the graph

of $\hat{\varphi}_n^*(\cdot)$ lies in $(\hat{\partial}F)_{\varepsilon_2}$. It now follows from (4.8) that for $t \in [\bar{t}, \bar{t}+\alpha_n)$

$$
\begin{aligned}
&\rho(t, \varphi_n^*(t)) - \rho(t_n, \varphi_n^*(t_n)) \\
&\quad = \int_{t_n}^{t} \Lambda(s, \varphi_n^*(s), u_n^*(s), v_n^*(s))\, ds \\
&\quad = \int_{t_n}^{\tau_{k+1}} \Lambda(s, \varphi_n(s), u_n^*(s), v_n^*(s))\, ds \\
&\qquad + \sum_{j=1}^{J} \int_{\tau_{k+1}}^{\tau_{k+j+1}} \Lambda(s, \varphi_n^*(s), u_n^*(s), z_j^*)\, ds + \int_{\tau_{k+J}}^{t} \Lambda(s, \varphi_n^*(s), u_n^*(s), z_J^*)\, ds,
\end{aligned} \tag{4.10}
$$

where $J = J(n, t)$ and $z_j^* = z^*(\tau_{k+j}, \varphi_n^*(\tau_{k+j}))$.

We now estimate each of the terms on the right. It follows from Assumption I that there exists a constant $K > 0$, independent of controls u, v and depending only on the compact set $\mathscr{C}$ of initial values such that for all trajectories φ of the system (1.1) we have $|\varphi(t) - \varphi(t')| \leqq K|t - t'|$. Hence the first integral on the right is bounded by $K|\tau_{k+1} - t_n|$, which $\to 0$ as $n \to \infty$. It also follows that there exists an integer n_0 which depends only on $\mathscr{C}$ such that for all partition points τ_i, all nth stage trajectories, $n \geqq n_0$, $\hat{\varphi}_n^*(\cdot)$ resulting from all strategies Γ, Δ^*, we have $|(t, \varphi_n^*(t)) - (\tau_i, \varphi_n^*(\tau_i))| < \delta$, for all t in I_{i+1}, where δ is as in (4.7). Applying this and (4.7) to each integrand in (4.10) other than the first, gives $\Lambda(s, \varphi_n^*(s), u_n^*(s), z_j^*) \leqq -A/2, j = k+1, \cdots, k+J$. Combining this with the bound $K|\tau_{k+1} - t_n|$ on the first integral in (4.10) establishes (4.9) and the lemma.

Remark 4.1. We shall refer to the strategy Δ^* defined above as the modification of Δ by means of the forcing function z^*.

COROLLARY 4.4.1. *Let $\tilde{t} > \bar{t}$. For all n sufficiently large let $(\tilde{t}, \varphi_n^*(\tilde{t})) \in (\hat{\partial}F)_{\varepsilon_2}$ and let the graph of $\varphi_n^*(\cdot)$ on $[\tilde{t}, t]$ lie in $(\hat{\partial}F)_{\varepsilon_2}$. Then for sufficiently large n,*

$$
\rho(t, \varphi_n^*(t)) - \rho(\tilde{t}, \varphi_n^*(\tilde{t})) \leqq -A(t - \tilde{t})/2. \tag{4.11}
$$

Since $\bar{t} < \tilde{t}$, for n sufficiently large, $\tau_{k+1} < \tilde{t}$. Thus, if in (4.10) we replace t_n by $\tilde{t}$, we obtain that the integrand on the right in (4.10) is $\leqq -A/2$ for all $\tilde{t} \leqq s \leqq t$.

LEMMA 4.5. *Let μ satisfy $0 < \mu \leqq \varepsilon_2/2$ and let (Γ, Δ) be strategies in G. Let $\hat{\varphi}[\cdot, t_0, x_0, \Gamma, \Delta]$ be a motion with capture time t_f. Then there exists a $\Delta^* = \Delta^*(\Gamma, \Delta, \hat{\varphi}[\cdot])$ and a motion $\hat{\varphi}_\mu[\cdot, t_0, x_0, \Gamma, \Delta^*]$ in the game G_μ such that for $t_0 \leqq t \leqq t_f$,*

$$
\hat{\varphi}[t, t_0, x_0, \Gamma, \Delta] = \hat{\varphi}_\mu[t, t_0, x_0, \Gamma, \Delta^*], \tag{4.12}
$$

and

$$
\varphi^0[t_f, t_0, x_0, \Gamma, \Delta] \geqq \varphi_\mu^0[T, t_0, x_0, \Gamma, \Delta^*] - C\mu, \tag{4.13}
$$

where C is a constant independent of Γ, Δ, μ and the motion $\hat{\varphi}[\cdot]$, but dependent on the set $\mathscr{C}$ of initial conditions.

We first note that if $t_f = T$, then $\varphi^0[T] = \varphi^0[t_f] = \varphi_\mu^0[t_f] = \varphi_\mu^0[T]$. Thus we can take $\Delta^* = \Delta$ and (4.12) and (4.13) hold for any $C \geqq 0$. Hence we need only consider the case $t_f < T$.

Let $\{\hat{\varphi}_n(\cdot)\} = \{\hat{\varphi}_n(\cdot, t_0, x_{0n}, u_n, v_n)\}$ be the sequence (after relabelling) of nth stage trajectories converging to the motion $\hat{\varphi}[\cdot, t_0, x_0, \Gamma, \Delta]$. Let $t_{f,n}$ be the capture time of $\hat{\varphi}_n(\cdot)$ and let $t_n = \min(t_{f,n}, t_f)$. Then $t_n \leqq t_f$ and, as was seen in the proof of Lemma 4.1, $t_n \to t_f$. Applying Lemma 4.4 with $\bar{t} = t_f$ and $\{t_n\}$ as above, we obtain a strategy Δ^*, a sequence of nth stage trajectories $\{\hat{\varphi}_n^*(\cdot)\} = \{\hat{\varphi}_n^*(\cdot, t_0, x_{0n}, u_n^*, v_n^*)\}$ resulting from (Γ, Δ^*) and a motion $\hat{\varphi}^*[\cdot]$ having the properties stated in Lemma 4.4.

Let

$$\varphi_{\mu,n}(t, t_0, x_0, u_n^*, v_n^*) = \varphi_n^*(t, t_0, x_0, u_n^*, v_n^*) \tag{4.14}$$

for all $t_0 \leqq t \leqq T$. Since for $t_0 \leqq t \leqq t_n$, $\hat{\varphi}_n^*(t) = \hat{\varphi}_n(t)$, we have that

$$\varphi_{\mu,n}(t, t_0, x_0, u_n^*, v_n^*) = \varphi_n(t, t_0, x_0, u_n, v_n) \tag{4.15}$$

for $t_0 \leqq t \leqq t_n$. For $t_0 \leqq t \leqq t_n$, let

$$\varphi_{\mu,n}^0(t, t_0, x_0, u_n^*, v_n^*) = \varphi_n^{*0}(t, t_0, x_0, u_n^*, v_n^*) = \varphi_n^0(t, t_0, x_0, u_n, v_n). \tag{4.16}$$

For $t_n \leqq t \leqq T$, let

$$\varphi_{\mu,n}^0(t) = \varphi_n^0(t_n) + \int_{t_n}^{t} f_\mu^0(s, \varphi_n^*(s), u_n^*(s), v_n^*(s))\, ds. \tag{4.17}$$

The sequence $\{\hat{\varphi}_{\mu,n}(\cdot)\}$ thus defined is clearly a sequence of nth stage trajectories in the game G_μ corresponding to the strategies (Γ, Δ^*). We select a further subsequence so that we may assume that $\hat{\varphi}_{\mu,n}(\cdot)$ converges to a motion $\hat{\varphi}_\mu[\cdot, t_0, x_0, \Gamma, \Delta^*]$ of the game G_μ.

We shall show that the motion $\hat{\varphi}_\mu[\cdot]$ of the previous paragraph has the desired properties. First, we conclude immediately from (4.15) and (4.16) that (4.12) holds.

Now suppose that $t_f < T - (3\mu/A)$. Let $\varepsilon(n)$ be as in Lemma 4.4 and let

$$t_{\mu,n} = t_n + 2[\varepsilon(n) + \rho(t_n, \varphi_n^*(t_n)) + \mu]/A. \tag{4.18}$$

Then for n sufficiently large $t_{\mu,n} < T$. We assert that for each sufficiently large n, there exists an s_n in $[t_n, t_{\mu,n}]$ such that

$$\rho(s_n, \varphi_n^*(s_n)) < -\mu. \tag{4.19}$$

First we suppose that on the interval $[t_n, t_{\mu,n}]$ the graph of $\varphi_n^*(\cdot)$ lies entirely in $(\hat{\partial}F)_{\varepsilon_2}$. We can then set $t = t_{\mu,n}$ in (4.9) and obtain that $\rho(t_{\mu,n}, \varphi_n^*(t_{\mu,n})) < -\mu$.

For those n such that the graph of $\varphi_n^*(\cdot)$ on $[t_n, t_{\mu,n}]$ does not lie entirely in $(\hat{\partial}F)_{\varepsilon_2}$ we proceed as follows. We first note that by virtue of (4.12) we can find a $\tilde{t} > t_f$ such that for all n sufficiently large, the graph of $\varphi_n^*(\cdot)$ on $[t_f, \tilde{t}]$ will lie in $(\hat{\partial}F)_{\varepsilon_2}$. Let $\bar{t} = t_f$ and apply Corollary 4.4.1 to observe that if $t > \tilde{t}$ and if the graph of $\varphi_n^*(\cdot)$ on $[\tilde{t}, t]$ lies in $(\hat{\partial}F)_{\varepsilon_2}$, then $\rho(t, \varphi_n^*(t)) < \rho(\tilde{t}, \varphi_n^*(\tilde{t}))$. From this we conclude that when the trajectory $\varphi_n^*(\cdot)$ leaves $(\hat{\partial}F)_{\varepsilon_2}$, it will do so in the interior of F. Since $\mu < \varepsilon_2/2$, (4.19) follows.

From (4.19) and (4.14) we see that for all n sufficiently large we are assured that the trajectories $\varphi_{\mu,n}(\cdot)$ will eventually leave the set $F_\mu = (\hat{\partial}F)_\mu \cap F$ at some time $t_n < \bar{s}_n < t_{\mu,n}$. If we use Corollary 4.41 with $\bar{t} = t_f$, we can easily show that for sufficiently large n, once the trajectory $\varphi_{\mu,n}(\cdot)$ leaves F_μ, then it will never reenter F_μ and will remain in $F - F_\mu$. Hence for $t_{\mu,n} \leqq t \leqq T$, $f_\mu^0(t, \varphi_{\mu,n}(t), u_n^*(t), v_n^*(t)) = 0$. From this and from (4.17) we get

$$\varphi_{\mu,n}^0(T) = \varphi^0(t_n) + \int_{t_n}^{t_{\mu,n}} f_\mu^0(s, \varphi_n^*(s), u_n^*(s), v_n^*(s))\, ds. \tag{4.20}$$

Recall that $f_\mu^0 \geqq 0$. The integral in (4.20) is bounded above by $K(t_{\mu,n} - t_n)$, where K is a positive constant which depends only on the set $\mathscr{C}$ of initial data. Therefore, if we let $n \to \infty$ in (4.20) and take into account (4.18), we get $\varphi_\mu^0[T] \leqq \varphi^0[t_f] + C\mu$, for some constant C depending only on $\mathscr{C}$. But this is precisely (4.13).

If $t_f \geqq T-(3\mu/A)$ then $T-t_f \leqq 3\mu/A$. In (4.17) replace t by T. Then the resulting integral will be bounded above by $L(T-t_n)$. The result again follows on letting $n\to\infty$, Lemma 4.5 is thus proved.

We now can easily establish Lemma 4.3. Fix Γ. From (4.13) we have

$$\begin{aligned}\inf_{\Delta}\varphi^0[t_f, t_0, x_0, \Gamma, \Delta] &\geqq \inf_{\Delta}\varphi^0_\mu[T, t_0, x_0, \Gamma, \Delta^*(\Delta)] - C_\mu \\ &\geqq \inf_{\tilde{\Delta}}\varphi^0_\mu[T, t_0, x_0, \Gamma, \tilde{\Delta}] - C_\mu,\end{aligned}$$

where $\tilde{\Delta}$ is an arbitrary strategy in G_μ over $[t_0, T]$. If we now take the supremum over Γ, we obtain the conclusion of Lemma 4.3.

5. Properties of W^-.

LEMMA 5.1. *Let Assumption* I *and Assumption* II *or* II′ *hold. Then for each* (t, x) *in* R

$$\lim_{\mu\to 0} W^-(t, x, \mu) = W^-(t, x), \tag{5.1}$$

and the convergence is uniform on compact subsets $\mathscr{C}$ *of* R. *The function* W^- *is continuous on* R.

Proof. The second statement is a consequence of the first. Let $\mathscr{C}$ be a compact subset of R. Then by Lemmas 4.2 and 4.3 there exists a constant C such that for all (t, x) in R and μ sufficiently small

$$W^-(t, x, \mu) - C\mu \leqq W^-(t, x) \leqq W^-(t, x, \mu).$$

The first statement follows from this.

Let

$$\tilde{R} \equiv R - \partial\mathscr{T} \equiv ([T_0, \infty)\times R^n) - \mathscr{T}.$$

LEMMA 5.2. *Let Assumptions* I′ *and* II′ *hold. Then the function* W^- *is Lipschitz continuous on compact subsets of* $\tilde{R}$.

Let (τ, ξ) and (τ', ξ') be two points in $\tilde{R}$. The crux of the proof is Lemma 5.3 below, which relates the payoff resulting from a motion with initial point (τ, ξ) to the payoff resulting from a motion with initial point (τ', ξ'). In [1, § 6] we showed that to each pair of strategies (Γ, Δ) in the game with initial point (τ, ξ) there corresponds a pair of strategies $(\theta\Gamma, \theta\Delta)$ in the game with initial point (τ', ξ'). Moreover the mappings $\theta\colon \Gamma\to\theta\Gamma$ and $\theta\colon \Delta\to\theta\Delta$ are one-to-one and onto.

LEMMA 5.3. *Let* X *be a compact subset of* $\tilde{R}$ *and let* (τ, ξ) *be a point in* X. *Then there exists a* $\delta_0 > 0$ *and a* $K > 0$, *both depending on the set* X *but not on the point* (τ, ξ), *such that for all* (τ', ξ') *satisfying* (τ', ξ') *in* X,

$$|(\tau', \xi') - (\tau, \xi)| \leqq \delta_0 \tag{5.2}$$

the following is true. Given a pair of strategies (Γ, Δ) *in* $G(\tau, \xi)$ *and a motion* $\hat{\varphi}[\cdot, \tau, \xi, \Gamma, \Delta]$ *with capture time* t_f, *then there exists a strategy* Δ^* *in* $G(\tau', \xi')$ *and a motion* $\hat{\varphi}[\cdot, \tau', \xi', \theta\Gamma, \Delta^*]$ *with capture time* s_f^* *such that*

$$\varphi^0[t_f, \tau, \xi, \Gamma, \Delta] \geqq \varphi^0[s_f^*, \tau', \xi', \theta\Gamma, \Delta^*] + E(\tau, \xi, \tau', \xi'), \tag{5.3}$$

where

$$|E(\tau, \xi, \tau', \xi')| \leqq K[|\tau-\tau'| + |\xi-\xi'|] \tag{5.4}$$

for all (τ, ξ, τ', ξ') *satisfying* (5.2).

Proof. Let $\{\hat{\varphi}_n\} = \{\hat{\varphi}_n(\cdot, \tau, \xi_n, u_n, v_n)\}$ denote the sequence of nth stage trajectories converging uniformly to the motion $\varphi[\cdot, \tau, \xi, \Gamma, \Delta]$ and let $\{t_{f,n}\}$ denote the corresponding sequence of capture times of the trajectories. Let $\{\theta\hat{\varphi}_n\} = \{\theta\hat{\varphi}_n(\cdot, \tau', \xi'_n, \theta u_n, \theta v_n)\}$ denote the sequence of nth stage trajectories resulting from $(\theta\Gamma, \theta\Delta)$ and let $\{s_{f,n}\}$ denote the corresponding sequence of capture times. We suppose, after choosing a subsequence if necessary, that $\{\theta\hat{\varphi}_n\}$ converges to a motion $(\theta\hat{\varphi}[\cdot]) = \hat{\varphi}[\cdot, \tau', \xi', \theta\Gamma, \theta\Delta]$ with capture time s_f. Recall that the mapping θ utilizes the relationship $s = s(t)$ between the time variable t on $[\tau, T]$ and the time variable s on $[\tau', T]$ given by (6.5) of [1]. It follows from (6.5) of [1] that for $\tau \leqq t \leqq T$, $|s(t) - t| \leqq |\tau - \tau'|$.

We consider three cases. First we suppose that $s(t_f) < s_f$. Let $t_n = \min(t_f, t_{f,n})$, and let $s_n = \min(s_f, s_{f,n})$. Then $t_n \to t_f$ and $s_n \to s_f$. Hence $s(t_n) \leqq s(t_f)$ and $s(t_n) \to s(t_f)$. Thus for n sufficiently large, $s(t_n) < s_n$, and consequently $\rho(s(t_n), \theta\varphi(s(t_n))) > 0$. From the preceding, the triangle inequality and Lemma 6.4 of [1] we see that there exists a constant K_1, which depends on X, such that

$$\begin{aligned} \rho(s(t_n), \theta\varphi_n(s(t_n))) &\leqq |(s(t_n), \theta\varphi_n(s(t_n))) - (t_f, \varphi[t_f])| \\ &\leqq |(s(t_n), \theta\varphi_n(s(t_n))) - (t_n, \varphi_n(t_n))| + |(t_n, \varphi_n(t_n)) - (t_f, \varphi[t_f])| \\ &\leqq K_1[|\tau - \tau'| + |\xi - \xi'|] + \varepsilon_n, \end{aligned} \tag{5.5}$$

where $\varepsilon_n \to 0$ as $n \to \infty$, uniformly for (τ, ξ), (τ', ξ') in X. It follows that there exists a $\delta_0 > 0$, which depends on X, such that whenever $(\tau', \xi') \in X$ and $|(\tau', \xi') - (\tau, \xi)| \leqq \delta_0$, then $(s(t_n), \theta\varphi_n(s(t_n))) \in (\partial F)_{\varepsilon_2}$ for all n sufficiently large. Thus $(s(t_f), \theta\varphi[s(t_f)]) \in (\partial F)_{\varepsilon_2}$.

Let $(\theta\Delta_n)^*$ denote the modification of $\theta\Delta_n$ after $s(t_f)$ by means of the forcing function $z^*(t, x)$. Let $(\theta\hat{\varphi}_n)^*(\cdot)$ denote the corresponding nth stage trajectory and let $(\theta\hat{\varphi})^*[\cdot]$ denote any motion $(\theta\hat{\varphi})^*[\cdot, \tau', \xi', \theta\Gamma, (\theta\Delta)^*]$. Note that for $0 \leqq s \leqq s(t_n)$ we have $(\theta\hat{\varphi}_n)(s) = (\theta\hat{\varphi}_n)^*(s)$, and for $0 \leqq s \leqq s(t_f)$ we have $(\theta\hat{\varphi})[s] = (\theta\hat{\varphi})^*[s]$. It follows from (4.9) that for each n there exists a maximal interval $[s(t_f), st_f) + \alpha_n)$ such that for s in this interval,

$$\rho(s, (\theta\varphi_n)^*(s)) - \rho(s(t_n), (\theta\varphi_n)^*(s(t_n))) \leqq \varepsilon_n - A(s - s(t_n))/2. \tag{5.6}$$

Since for $s > s(t_f)$, $d\rho/ds < 0$ along $(\theta\varphi_n)^*(s)$, and since $s(t_n) \to s_f$, it follows that for n sufficiently large the trajectory $(\theta\varphi_n)^*(s)$ has a capture time $s^*_{f,n} \in [s(t_f), s(t_f) + \alpha_n)$. Let s^*_f denote the capture time of $(\theta\hat{\varphi})^*[\cdot]$, and let $s^*_n = \min(s^*_f, s^*_{f,n})$. If we now set $s = s^*_n$ in (5.6), note that $(\theta\varphi_n)^*(s_n) \to (\theta\varphi)^*[s^*_f]$ and use (5.5) we get that

$$s^*_n - s(t_n) \leqq \varepsilon'_n + (2K_1/A)[|\tau - \tau'| + |\xi - \xi'|], \tag{5.7}$$

where $\varepsilon'_n \to 0$ as $n \to \infty$, uniformly with respect to (Γ, Δ) and (τ, ξ, τ', ξ') satisfying (5.2).

If we denote the outcomes of $(\theta\Gamma_n(\theta\Delta_n)^*)$ by $((\theta u_n)^*, (\theta v_n)^*)$, then for n sufficiently large we have

$$\begin{aligned} &(\theta\varphi^0_n)^*(s^*_n, \tau', \xi'_n, (\theta u_n)^*(\theta v_n)^*) \\ &\quad = \varphi^0_n(s(t_n), \tau', \xi'_n, \theta u_n, \theta v_n) \\ &\qquad + \int_{s(t_n)}^{s^*_n} f^0(s, (\theta\varphi_n)^*(s), (\theta u_n)^*(s), (\theta v_n)^*(s))\, ds. \end{aligned}$$

It now follows from Lemma 6.5 of [1], from (5.7) and from the fact that the integrand on the right is bounded that

$$(\theta\varphi^0_n)^*(s^*_n, \tau', \xi'_n, (\theta u_n)^*, (\theta v_n)^*) = \varphi^0_n(t_n, \tau, \xi_n, u_n, v_n) + E'(\tau, \xi, \tau', \xi') + \varepsilon''_n,$$

where E' is as in (5.4) and $\varepsilon_n'' \to 0$ uniformly with respect to (Γ, Δ) and (τ, ξ, τ', ξ') satisfying (5.2). If we now let $n \to \infty$, and take subsequences, if necessary, we obtain

$$\varphi^0[t_f, \tau, \xi, \Gamma, \Delta] = (\theta\varphi^0)^*[s_f^*, \tau', \xi', \theta\Gamma, (\theta\Delta)^*] + E(\tau, \xi, \tau', \xi'),$$

where $E = -E'$. If we now set $\Delta^* \equiv (\theta\Delta)^*$ and set $\hat{\varphi}[\cdot, \tau', \xi', \theta\Gamma, \Delta^*] = (\theta\hat{\varphi})^*[\cdot, \tau', \xi', \theta\Gamma, (\theta\Delta)^*]$ we obtain (5.3) and (5.4) in the case under consideration.

Let us now suppose that $s_f < s(t_f)$. Then $t_f > t(s_f)$ and also $t_n > t(s_n)$ for n sufficiently large. Hence, since $f^0 \geqq 0$, we have that $\varphi^0(t_n, \tau, \xi_n, u_n, v_n) \geqq \varphi^0(t(s_n), \tau, \xi_n, u_n, v_n)$. If we now use Lemma 6.4 of [1] we get

$$\varphi^0(t_n, \tau, \xi_n, u_n, v_n) \geqq \theta\varphi^0(s_n, \tau', \xi_n', \theta u_n, \theta v_n) + E,$$

where E is as in (5.4). If we let $n \to \infty$ we get

$$\varphi^0[t_f, \tau, \xi, \Gamma, \Delta] \geqq \theta\varphi^0[s_f, \tau', \xi', \theta\Gamma, \theta\Delta] + E.$$

If we set $s_f^* = s_f$, $\Delta^* = \theta\Delta$ and $\hat{\varphi}[\cdot, \tau', \xi', \theta\Gamma, \Delta^*] = \theta\hat{\varphi}[\cdot, \tau', \xi', \theta\Gamma, \theta\Delta]$, we obtain statements (5.3) and (5.4).

If $s_f = s(t_f)$, the lemma follows immediately from Lemma 6.4 of [1] with $s_f^* = s_f$ and $\Delta^* = \theta\Delta$.

We return to the proof of Lemma 5.2. Since W^- is continuous on R, to prove that it is Lipschitz continuous on a compact subset X of R it suffices to show that there exists a $\delta_0 \geqq 0$ and a constant K such that if (τ, ξ) and (τ', ξ') are as in (5.2) then

$$|W^-(\tau, \xi) - W^-(\tau', \xi')| \leqq K[|\tau - \tau'| + |\xi - \xi'|]. \tag{5.8}$$

Let $(\tau, \xi) \in X$, let $\varepsilon > 0$ be arbitrary and let Γ be arbitrary. Then there exists a Δ_Γ and a motion $\hat{\varphi}[\cdot, \tau, \xi, \Gamma, \Delta_\Gamma]$ such that $\varepsilon + W^-(\tau, \xi) \geqq \varphi^0[t_f, \tau, \xi, \Gamma, \Delta_\Gamma]$. Let (τ', ξ') be as in (5.2). Then by Lemma 5.3, there exists a strategy Δ^* in $G(\tau', \xi')$ and a motion $\hat{\varphi}[\cdot, \tau', \xi', \theta\Gamma, \Delta^*]$ with capture time s_f^* such that

$$\varepsilon + W^-(\tau, \xi) \geqq \varphi^0[s_f^*, \tau', \xi', \theta\Gamma, \Delta^*] + E(\tau, \xi, \tau', \xi'),$$

where E is as in (5.4). Hence

$$\varepsilon + W^-(\tau, \xi) \geqq \inf_{\Delta} \varphi^0[s_f, \tau', \xi', \theta\Gamma, \Delta] + E(\tau, \xi, \tau', \xi'),$$

where the infinimum is taken over all strategies Δ on $[\tau', T]$ and all motions $\hat{\varphi}[\cdot, \tau', \xi', \theta\Gamma, \Delta]$. Since $\varepsilon > 0$ is arbitrary and the mapping $\theta: \Gamma \to \theta\Gamma$ is one-to-one onto, it now follows that $W^-(\tau, \xi) \geqq W^-(\tau', \xi') + E$. We may also start with (τ', ξ') and obtain that $W^-(\tau', \xi') \geqq W^-(\tau, \xi) + E'$. From the last two inequalities (5.8) follows.

6. Existence of value. Let

$$H(t, x, \lambda^0, \lambda, y, z) = \lambda^0 f^0(t, x, y, z) + \langle \lambda, f(t, x, y, z) \rangle. \tag{6.1}$$

DEFINITION 6.1. The Isaacs condition holds at a point (t, x) for the game G if

$$\min_z \max_y H(t, x, 1, \lambda, y, z) = \max_y \min_z H(t, x, 1, \lambda, y, z) \tag{6.2}$$

for all $\lambda \in R^n$, where the max is taken over y in Y and the min is taken over $z \in Z$. The Isaacs condition holds on a set in R^{n+1} if it holds at each point of the set.

The Isaacs condition (6.2) holds if and only if for all $\lambda^0 \geqq 0$

$$\min_z \max_y H(t, x, \lambda^0, \lambda, y, z) = \max_y \min_z H(t, x, \lambda^0, \lambda, y, z). \tag{6.3}$$

If (6.3) holds for all $\lambda^0 \geqq 0$, then (6.2) surely holds. Since for $\lambda^0 > 0$, $H(t, x, \lambda^0, \lambda, y,$

$z) = \lambda^0 H(t, x, 1, \lambda/\lambda^0, y, z)$, it follows that if (6.2) holds, then (6.3) holds for $\lambda^0 > 0$. To obtain (6.3) for $\lambda^0 = 0$, write (6.3) in the equivalent form of a saddle point inequality and let $\lambda^0 \to 0$.

Assumption III. The Isaacs condition (6.2) holds at all points of $[T_0, T_1] \times R^n$.

At each point (t, x) of $[T_0, T_1] \times R^n$, let

$$\min_z \max_y [s^0 f_\mu^0 + \langle s, f \rangle] = \max_y \min_z [s^0 f_\mu^0 + \langle s, f \rangle], \tag{6.4}$$

for all $\hat{s} = (s^0, s)$, where the argument of f_μ^0 and f is (t, x, y, z). It then follows from [1] that each game G_μ has a value and saddle point. We shall show in the appendix that we only need (6.4) to hold for those vectors $\hat{s}$ with $s^0 \geqq 0$ in order for G_μ to have a value and saddle point.

For fixed (t, x), $f_\mu^0(t, x, y, z) = \alpha(t, x) f^0(t, x, y, z)$, where $0 \leqq \alpha(t, x) \leqq 1$. It follows from the equivalence of (6.3) and (6.2) that if the Isaacs condition (6.2) holds for the game G, then (6.4) holds for all vectors $\hat{s}$ with $s^0 \geqq 0$. Thus, if Assumption III holds then each game G_μ has a value and a saddle point.

THEOREM 6.1. *Let Assumptions* I, II *or* II′ *and* III *hold. Then for each* (t, x) *in* R *the game* $G(t, x)$ *has value* $W(t, x)$ *and the function* W *is continuous on* r. *If Assumptions* I′, II′ and III *hold, then* W *is Lipschitz continuous on compact subsets of* $\tilde{R}$.

Let (t, x) be fixed in R. Then by Lemma 4.2, $W^+(t, x) \leqq W^+(t, x, \mu)$. Since each game $G_\mu(t, x)$ has value, we have $W^+(t, x, \mu) = W^-(t, x, \mu)$ and so $W^+(t, x) \leqq W^-(t, x, \mu)$. It now follows from (5.1) that $W^+(t, x) \leqq W^-(t, x)$, and thus $G(t, x)$ has value. The continuity of W now follows from Lemma 5.1 and the Lipschitz property from Lemma 5.2.

7. The Isaacs equation. In studying the game G we have essentially converted the game in $[T_0, T_1] \times R^n$ with integral payoff (1.2) into a game in $[T_0, T_1] \times R^{n+1}$ with terminal payoff $g(t_f, \hat{x}_f) = x_f^0$. For the game in $[T_0, T_1] \times R^{n+1}$ we have always taken the initial condition $(t_0, \hat{x}_0)$ to be such that $\hat{x}_0 = (0, x_0)$, where $(t_0, x_0) \in R$. In developing the Isaacs equation it will be necessary to consider initial conditions with $\hat{x}_0 = (x_0^0, x_0)$, where x_0^0 is arbitrary. If we consider such arbitrary initial conditions, then it is easy to see that if $\hat{W}^+(t_0, \hat{x}_0)$ denotes the upper value of this game and if $\hat{W}^-(t_0, \hat{x}_0)$ denotes the lower value, then $\hat{W}^\pm(t_0, \hat{x}_0) = x_0^0 + W^\pm(t_0, x_0)$. Thus, if the value $W(t_0, x_0)$ exists, then for any $\hat{x}_0$ so does $\hat{W}(t_0, \hat{x}_0)$ and

$$\hat{W}(t_0, \hat{x}_0) = x_0^0 + W(t_0, x_0). \tag{7.1}$$

The principal result of this section is Theorem 7.1 below, which states that if the data of the problem is Lipschitz then the value function W satisfies the Isaacs equation. The first step in the proof of this result is to establish the following modification of Lemma 8.3 of [1].

Let $(t_0, \hat{x}_0)$ be fixed, let $v_0 = \hat{W}^-(t_0, \hat{x}_0)$ and let $v^0 = \hat{W}^+(t_0, \hat{x}_0)$. Let $\hat{C}(v_0) = \{(\tau, \hat{\xi}) : (\tau, \xi) \notin \mathcal{T},\ \hat{W}^-(\tau, \hat{\xi}) \leqq v_0\}$ and let $\hat{C}(v^0) = \{(\tau, \hat{\xi}) : (\tau, \xi) \notin \mathcal{T},\ \hat{W}^+(\tau, \hat{\xi}) \geqq v^0\}$.

LEMMA 7.1. *Let* $(\tau, \hat{\xi})$ *be a point of* $\hat{C}(v_0)$. *Let* t_1 *satisfy* $\tau < t_1 < T$ *and let* u *be any control for Player I on* $[\tau, t_1]$. *Then either there exists a relaxed control* $\zeta' = \zeta'(u)$ *such that the relaxed trajectory* $\hat{\psi}(\cdot, \tau, \hat{\xi}, u, \zeta')$ *has the property that* $(t_f, \psi(t_f)) \in \mathcal{T}$ *for some* $t_f \in [\tau, t_1]$ *or there exists a relaxed control* $\zeta = \zeta(u)$ *such that* $(t, \psi(t)) \notin \mathcal{T}$ *for all* $t \in [\tau, t_1]$ *and* $(t_1, \psi(t_1)) \in \hat{C}(v_0)$.

The proof is a modification of the proof of Lemma 8.3 of [1] and will be omitted.

Using Lemma 7.1 in the way Lemma 8.3 of [1] was used to establish Lemma 13.1 of [1], we can establish the analogue of Lemma 13.1 for $\hat{W}^+$ and $\hat{W}^-$. From this we obtain Theorem 7.1.

THEOREM 7.1. *Let Assumptions* I′, II′ *and* III *hold. then at almost every point of* $\tilde{R}$:

$$\begin{aligned} -W_t(t,x) &= H(t,x,1,W_x(t,x),y^*,z^*) \\ &= \min_z \max_y H(t,x,1,W_x(t,x),y,z) \\ &= \max_y \min_z H(t,x,1,W_x(t,x),y,z), \end{aligned}$$

where H is as in (6.1), *the* max *is taken over Y, the* min *is taken over Z, and* $y^* = y^*(t,x)$, $z^* = z^*(t,x)$ *is the saddle point for the game over* $Y \times Z$ *with payoff* $H(t,x,1,W_x(t,x),y,z)$.

Appendix. Games $G(t_0, x_0)$ in $[T_0, T_1] \times R^n$ with fixed terminal time T and payoffs of the form $g(x_f) + \int_{t_0}^{T} f^0(t,x,u,v)\,dt$ can be written as games $\hat{G}(t_0, \hat{x}_0)$ in $[T_0, T_1] \times R^{n+1}$ with terminal payoff $g(x_f) + x_f^0$ by adjoining a zeroth coordinate as in (3.1). If we let $\hat{W}^-(\tau, \hat{\xi})$ with $\hat{\xi} = (\xi^0, \xi)$ denote the upper value of the game $\hat{G}(\tau, \hat{\xi})$ and let $W^-(\tau, \xi)$ denote the lower value of the game $G(\tau, \xi)$ then $\hat{W}^-(\tau, \hat{\xi}) = \xi^0 + W^-(\tau, \xi)$. A similar relationship holds for $\hat{W}^+(\tau, \xi)$.

We now refer to [1, § 10], interpreted for the game $\hat{G}(t_0, \hat{x}_0)$. In constructing the extremal aiming strategies for this game the Isaacs condition is required to hold only for those vectors $\hat{s} = (s^0, s)$ of the form $\hat{s} = \hat{s}^* = \hat{x}^* - \hat{w}^*$, where $(t^*, \hat{x}^*) \notin \hat{C}(v_0)$ (or $\notin \hat{C}(v^0)$) and $(t^*, \hat{w}^*)$ is a point of $S(t^*) = H(t^*) \cap \hat{C}(v_0)$ of minimum distance to $(t^*, \hat{x}^*)$. We shall now show that for $\hat{s}^*$ we always have $s^{*0} \geqq 0$.

Let $(\tau, \hat{\xi})$ be a point $\notin \hat{C}(v_0)$. Let $(\tau, \hat{\xi}^*)$ be a point of $S(\tau) = H(\tau) \cap \hat{C}(v_0)$ at which the distance from $S(\tau)$ to $(\tau, \hat{\xi})$ is achieved. We first note that $\hat{W}^-(\tau, \hat{\xi}^*) = v_0$. For if $\hat{W}^-(\tau, \hat{\xi}^*) < v_0$, then since W^- is continuous there would exist points of $\hat{C}(v_0)$ closer to $(\tau, \hat{\xi})$ than $(\tau, \hat{\xi}^*)$. Thus, $W^-(\tau, \xi^*) + \xi^{*0} = v_0$. Now $\hat{s}^* = \hat{\xi} - \hat{\xi}^*$, so if $s^{*0} < 0$, we must have $\xi^0 < \xi^{*0}$. But if this were the case, then $W^-(\tau, \xi^*) + \xi^0 < v_0$. Hence $(\xi^*, \xi^0) \in S(\tau)$. Clearly, $|(\xi^*, \xi^0) - (\xi, \xi^0)| < |(\xi^*, \xi^{*0}) - (\xi, \xi^0)|$. But this contradicts the definition of (ξ^*, ξ^{*0}). Hence $s^{*0} < 0$ is not possible.

REFERENCES

[1] L. D. BERKOVITZ, *The existence of value and saddle point in games of fixed duration*, this Journal, 23 (1985), pp. 172–196.

[2] A. FRIEDMAN, *Differential Games*, John Wiley, New York, London, Sydney, Toronto, 1971.

[3] ———, *Differential Games*, CBMS Regional Conference Series in Mathematics, 18, American Mathematical Society, Providence, RI, 1974.

[4] N. N. KRASOVSKII AND A. I. SUBBOTIN, *Positional Differential Games*, Nauka, Moscow, 1974. (In Russian.)

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 3, May 1986

004

DYNAMICAL REALIZATIONS OF HOMOGENEOUS HAMILTONIAN SYSTEMS*

P. E. CROUCH† AND M. IRVING‡

Abstract. In this work we consider homogeneous input-output systems defined by a single Volterra kernel, which have an internal state space realization in the form of a Hamiltonian system. Previous work by one of the authors considered the state space realization of input-output maps determined by finite Volterra series, the homogeneous case being characterized by an internal homogeneity. The present paper shows that minimal Hamiltonian realizations exist which exhibit the same polynomic structure whilst simultaneously displaying the canonical symplectic structure. The homogeneity is conveniently handled by extensive use of graded vector spaces, and their relation with nilpotent Lie algebras. The additional relation with symplectic structures is also utilized here. The work also provides a specialized version of the Darboux-Weinstein theorem, which is globally valid.

Key words. Hamiltonian, symplectic, Volterra series, realization theory, canonical form

1. Introduction.

1.1. One of the authors has previously considered minimal state space representations of input-output systems defined by finite Volterra series [4]. These dynamical realizations, or controlled dynamical systems with outputs, are characterized by certain polynomic vector fields on a vector space generating solvable Lie algebras, with codimension one nilpotent ideal. When the Volterra series consists of a single term this external homogeneity is reflected in a particular type of internal homogeneity.

Such input-output systems also determine an intrinsic graded vector space structure on the state space which may be exploited to describe the nilpotent Lie algebras which occur. The relationship between graded vector spaces and nilpotent Lie algebras is well understood; see, for example, [7].

A controlled dynamical system on a symplectic manifold may take a special form, being defined by Hamiltonian vector fields, with associated output maps being the Hamiltonian functions. The input-output maps of such Hamiltonian systems therefore represent a subclass of all input-output systems, and in particular those with finite Volterra series form a subclass of finite Volterra series. In Crouch and Irving [5] necessary and sufficient conditions are given for a finite Volterra series to be the input-output map of a Hamiltonian system. However, the state space representation used there for such systems was that of the bilinear system. The problem of obtaining minimal representations, as for arbitrary finite Volterra series in Crouch [4] had remained unanswered. That minimal Hamiltonian realizations exist at all have been answered affirmatively by van der Schaft [17], and under more restrictive conditions in Goncalves [6], although in this latter work a technique is exploited which to some extent is used in this paper.

In this paper we obtain minimal Hamiltonian realizations in the form of [4] but only in the restricted case of homogeneous Volterra series. The case of Volterra series consisting of only the term linear in "u" corresponds to minimal linear systems. Realizability and structural questions for minimal Hamiltonian systems within this class are dealt with in van der Schaft [22]. This paper can be viewed as a partial generalization of that work. The general case remains a formidable problem.

* Received by the editor March 13, 1984, and in revised form January 29, 1985.

† Department of Mathematics, Arizona State University, Tempe, Arizona 85287, and Department of Engineering, University of Warwick, Coventry CV4 7AL, England.

‡ The work of this author was supported in part by S.E.R.C. under grant GR/B/9116.7 while at the Department of Engineering, University of Warwick, Coventry CV4 7AL, England.

Symplectic geometry has already exposed much of the structure in transitive Lie group actions on symplectic manifolds where the corresponding Lie algebra is represented by Hamiltonian vector fields. We review some of this material briefly in § 2. See also Souriau [15], Kostant [10], Kirillov [9], and the work of Wallach [18]. In this work we are obviously interested in the case of nilpotent and solvable group actions on noncompact manifolds. In fact such a transitive, nilpotent group action on a compact symplectic manifold implies that the group is Abelian, Zwart and Boothby [20]. The results most clearly allied to ours are contained in Pukansky [14] and again Wallach [18]. However there is a lot of work on the associated algebraic and representation problems in Kirillov [9] and Auslander and Kostant [1]. In fact we use a few of the results in this latter reference to a large extent. We feel that although the problem studied here is motivated by a problem in control theory, it is of independent mathematical interest, and as such we give both control theoretic and abstract mathematical interpretations of the results.

Two main techniques are used to derive the results. The first, studied in § 4, relies on the decomposition of systems with nilpotent Lie algebras, as in Crouch [4], and itself based on earlier work of Krener [11] and Chen [3]. This gives a coordinate representation of the abstract system on a graded vector space with compatible symplectic structure. This defines a graded symplectic vector space which we analyze in § 2.

The second technique depends on a new and global version of the Darboux-Weinstein theorem in [19], and is described in § 5.

Many of the ideas here evolved during the Ph.D. thesis of M. Irving [5]. However the approach taken by Irving is somewhat different, and the main aim is to produce local results but in the more general case of arbitrary finite Volterra series.

In the remainder of the section we make the structure and problems raised here more specific.

1.2. We assume all data is real analytic and introduce the class of systems defined by the equations

$$\begin{aligned} &\dot{x}(t) = X_0(x(t)) + \sum_{i=1}^{m} u_i(t) X_i(x(t)), \qquad x(0) = x_0, \\ &y_j(t) = h_j(x(t)), \qquad 1 \leqq j \leqq p, \quad X_0(x_0) = 0, \end{aligned} \tag{1}$$

where the state x belongs to a manifold M and $X_0, X_1, \cdots, X_m$ are complete vector fields on M with $h_1, \cdots, h_p$ real valued functions on M. In this paper we are interested in the subclass of systems which exhibit the following relation between the output functions $t \mapsto y_i(t)$, and the input or control functions $t \mapsto u_i(t)$, called the input-output map.

$$y_j(t) = \int_0^t \cdots \int_0^{\sigma_{N-1}} \sum_{i_1, \cdots, i_N} W_N^{j i_1 \cdots i_N}(t, \sigma_1, \cdots, \sigma_N) u_{i_1}(\sigma_1) \cdots u_{i_N}(\sigma_N)\, d\sigma_N \cdots d\sigma_1 \tag{2}$$

for suitable continuously differentiable, and hence analytic functions $W_N^{j i_1 \cdots i_N}$ in the variables $t, \sigma_1, \cdots, \sigma_N$. To describe a natural class of systems which exhibit such input-output maps we briefly recall some results from graded vector spaces; see Goodman [7].

Let V be an n-dimensional real vector space viewed as a direct sum of N vector spaces V_i, $V = V_1 \oplus V_2 \oplus \cdots \oplus V_N$ and write $x \in V$ as $(x_1, x_2, \cdots, x_N)$ where $x_i \in V_i$. We define for each $t > 0$ a dilation δ_t by

$$\delta_t(x_1, x_2, \cdots, x_N) = (t x_1, t^2 x_2, \cdots, t^N x_N).$$

We call the pair (V, δ_t) a graded vector space of degree N. We say a function h on V is homogeneous of degree k if

$$h \circ \delta_t = t^k h.$$

Let Q^k denote the vector space of all homogeneous polynomials of degree k on V. A vector field X on V with polynomial coefficients is said to be homogeneous of degree m, $0 \leqq m \leqq N$, if for each $k \geqq 0$ and $h \in Q^k$

$$L_X(h) \in Q^{k-m}$$

where L denotes the Lie derivative, and $Q^k = \{0\}$ for $k < 0$. Equivalently, a vector field X is homogeneous of degree m if

$$\delta_{t*} X = (X \circ \delta_t) t^m.$$

We denote the space of all vector vector fields homogneeous of degree m by P^m. We easily deduce that

$$[P^k, P^m] \subset p^{m+k}, \qquad Q^k \otimes P^m \subset P^{m-k}$$

where $[\cdot, \cdot]$ denotes Lie bracket.

From the results of Crouch [4] one may deduce the following result.

THEOREM 1. *If there exists a system* (1) *whose input-output map coincides with a given input-output map as in* (2), *then there exists another system defined on a graded vector space V and represented by*

$$\begin{aligned} &\dot{z}(t) = Z_0(z(t)) + \sum_{i=1}^{m} u_i(t) Z_i(z(t)), \qquad z(0) = 0, \\ &y_j(t) = H_j(z(t)), \qquad 1 \leqq j \leqq p, \quad Z_0(0) = 0, \end{aligned} \tag{3}$$

with $z \in V = V_1 \oplus V_2 \oplus \cdots \oplus V_N$ satisfying

(i) $Z_0 \in P^0$, $Z_i \in P^1$, $1 \leqq i \leqq m$ $\quad H_j \in Q^N$, $1 \leqq j \leqq p$.

(ii) *The input-output map relating the y_j and u_i functions coincides with the given one.*

(iii) *The system is accessible; that is the Lie algebra $\mathcal{L}$ generated by $Z_0, Z_1, \cdots, Z_m$ is transitive on V, and in particular the subalgebra $\mathcal{S} \subset \mathcal{L}$ defined as the ideal generated by $Z_1, \cdots, Z_m$ in $\mathcal{L}$ is also transitive on V.*

(iv) *The system is locally observable; that is the smallest linear space $\mathcal{H}$ of functions on V containing H_j, $1 \leqq j \leqq p$ and closed under Lie derivatives by $Z_0, Z_1, \cdots, Z_m$ satisfies $d\mathcal{H}(x) = \{dh(x); h \in \mathcal{H}\} = T_x^* V = V$ for all $x \in V$.*

We also have the following result which includes a partial converse to Theorem 1.

THEOREM 2. *Any system* (3) *defined on a graded vector space (V, δ_t) of degree N, which satisfies condition* (i) *of Theorem* 1, *has an input-output map in the form of equation* (2). *The Lie algebra $\mathcal{L}$ is solvable, finite-dimensional and the ideal $\mathcal{S}$ is nilpotent and has a codimension at most one in $\mathcal{L}$. The descending central series of $\mathcal{S}$ has length less than or equal to N. That is $\mathcal{S}^{N+1} = \{0\}$ where $\mathcal{S}^1 = \mathcal{S}$, $\mathcal{S}^k = [\mathcal{S}, \mathcal{S}^{k-1}]$.*

Proof. For general systems (1) the input-output map can be expanded in a Volterra series, Brockett [2], Lesiak and Krener [12], the Nth term of which is given by (2). From Krener and Lesiak [12], it can be shown that the kernel $W_r^I(t, \sigma_1, \cdots, \sigma_r)$ defining the rth term is given by

$$\sum_J \sigma_r^{j_r}/j_r! \cdots \sigma_1^{j_1}/j_1!\, t^{j_0}/j_0!\, \alpha_r^{J,I}(x_0), \tag{4}$$

where

$$\alpha_r^{J,I} = L_{adX_0^{j_r}(X_{i_r})} \cdots L_{adX_0^{j_1}(X_{i_1})} L_{X_0}^{j_0}(h_j),$$
$$J = (j_0, j_1, \cdots, j_r), \qquad I = (j, i_1, \cdots, i_r).$$

Here $ad\, X_0(X) = [X_0, X]$ and $ad\, X_0^j(X) = ad\, X_0(ad\, X_0^{j-1}(X))$ with a similar definition for the repeated Lie derivative $L_{X_0}^j(h)$.

By using this formula in the case of system (3) under conditions (i) of Theorem 1 it is clear that W_r^I vanishes for each r except $r = N$ for each I. It follows that the input-output map in this case is an expression as in equation (2). The remaining results follow directly from the properties of the spaces P^i and Q^i. □

1.3. If now M is a symplectic manifold with symplectic form ω, denote by X_h the Hamiltonian vector field on M with Hamiltonian h, that is

$$i(X_h)\omega = dh$$

where i is the inner product. Let β be the map $h \to X_h$ the Lie algebra homomorphism of functions on M under Poisson bracket, also denoted $[\cdot, \cdot]$, into the Lie algebra of vector fields on M under the Lie bracket.

Consider a controlled differential equation with outputs, on the symplectic manifold (M, ω)

$$\begin{aligned} &\dot{x}(t) = X_{h_0}(x(t)) + \sum_{i=1}^{m} u_i(t) X_{h_i}(x(t)), \qquad x(0) = x_0, \\ &y_i(t) = h_i(x(t)), \qquad 1 \leqq i \leqq m, \quad X_{h_0}(x_0) = 0. \end{aligned} \tag{5}$$

We call such a system a Hamiltonian system. Motivation for such a definition is made in van der Schaft [23], see also Brockett [24]. For example, if one takes a set of Lagrangian equations with generalized inputs u_i and observations $y_i = q_i$, the configuration space generalized coordinates, we obtain

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{q}_i} - \frac{\partial L}{\partial q_i} = -u_i, \qquad 1 \leqq i \leqq n, \qquad y_i = q_i.$$

Under suitable conditions the Legendre transformation then yields a set of Hamiltonian equations in R^{2n} of the form

$$\begin{aligned} \dot{q}_i &= \frac{\partial H}{\partial p_i} + \sum_{j=1}^{n} u_j \frac{\partial H_j}{\partial p_i}, \qquad 1 \leqq i \leqq n, \\ \dot{p}_i &= -\frac{\partial H}{\partial q_i} - \sum_{j=1}^{n} u_j \frac{\partial H_j}{\partial q_i}, \qquad 1 \leqq i \leqq n, \\ y_i &= H_i, \end{aligned}$$

where $H_i(p, q) = q_i$. This is just a specific case of the system (5).

If system (5) exhibits an input-output map as in equation (2), we may use the results of van der Schaft [17], or Goncalves [6] to deduce that there exists another Hamiltonian system defined on another symplectic manifold (M', ω') denoted

$$\begin{aligned} &\dot{z}(t) = Z_{H_0}(z(t)) + \sum_{i=1}^{m} u_i(t) Z_{H_i}(z(t)), \qquad z(0) = z_0, \\ &y_i(t) = H_i(z(t)), \qquad 1 \leqq i \geqq m, \quad Z_{H_0}(z_0) = 0, \end{aligned} \tag{6}$$

satisfying conditions (ii)-(iv) of Theorem 1. Thus we define corresponding spaces $\mathscr{H}'$, $\mathscr{S}'$, $\mathscr{L}'$. Note that conditions (ii) and (iv) are equivalent in this case because β maps $\mathscr{H}'$ onto $\mathscr{S}'$, and ω' is nondegenerate. This motivates the question which we answer in this paper: can we reproduce condition (i) whilst simultaneously retaining the symplectic structure, and conditions (ii), (iii) and (iv) of Theorem 1.

Before posing the corresponding abstract mathematical question we comment on the added structure possessed by system (6) by virtue of its Hamiltonian nature. The isomorphism theorem of Sussmann [16] ensures that system (6) and the system (3) obtained by applying Theorem 1 are isomorphic. In particular M' is diffeomorphic to a Cartesian space, and the Lie algebras $\mathscr{S}'$ and $\mathscr{L}'$ are isomorphic to $\mathscr{S}$ and $\mathscr{L}$, respectively. In particular by Theorem 2 $\mathscr{S}'$ is a nilpotent ideal of codimension one in $\mathscr{L}'$, which has step length less than or equal to N and is finite-dimensional. If the Poisson Lie algebra of functions on M generated by $H_0, H_1, \cdots, H_m$ is denoted $\mathscr{K}'$, then $\mathscr{H}'$ is the codimension one ideal in $\mathscr{K}'$ generated by $H_1, \cdots, H_m$, such that β maps $\mathscr{K}'$ onto $\mathscr{L}'$ and $\mathscr{H}'$ onto $\mathscr{S}'$. Since the kernel of β consists of the one-dimensional space of constant functions, $\mathscr{K}'$ and $\mathscr{H}'$ are solvable and nilpotent Lie algebras, respectively.

Using the identities for Hamiltonian vector fields

$$[Z_H, Z_G] = Z_{[H,G]}, \qquad L_{Z_H}(G) = [G, H],$$

we may rewrite the coefficients $\alpha_r^{J,I}(x_0)$ up to a choice in sign by

$$\alpha_r^{J,I} = [ad\, H_0^{j_r}(H_{i_r}), [\cdots [ad\, H_0^{j_1}(H_{i_1}), ad\, H_0^{j_0}(H_j)] \cdots]. \tag{7}$$

Since system (6) has the input-output map in (2) we see that $\alpha_r^{J,I}(z_0)$ are zero for all J, I and r except $r = N$. We also deduce [4, Thm. 3.2] that the functions $\alpha_r^{J,I}$ are identically zero for $r > N$ and constant for $r = N$. It follows that $\mathscr{H}'$ has descending central series of length exactly $N+1$, denoted $\mathscr{H}'^k$, $k = 0, \cdots, N$, $\mathscr{H}'^0 = \mathscr{H}'$, with $\mathscr{H}'^N$ consisting of constant functions only.

1.4. The discussion in the previous subsection motivates consideration of the following abstract system of data:

A symplectic manifold (M, ω), a specific point $x_0 \in M$, a solvable Lie algebra $\mathscr{K}$ of functions on M generated by the functions $H_0, H_1, \cdots, H_m$ with a codimension one nilpotent ideal $\mathscr{H}$ generated by $H_1, \cdots, H_m$ and additionally satisfying:

Assumption (i). $\beta(H_i)$ are complete vector fields $0 \leqq i \leqq m$.

Assumption (ii). $\beta(H_0)(x_0) = 0$, $d\mathscr{H}(x) = T_x^\infty M$, $\quad x \in M$.

Assumption (iii). $\mathscr{H}$ has a descending central series of length $N+1$, and in particular there exist multi-indices J and I such that $\alpha_N^{J,I}(x_0) \neq 0$ ($\alpha_r^{J,I}$ defined in (7)).

Assumption (iv). $\alpha_r^{J,I}(x_0) = 0$ for $r < N$.

We make some prelminary deductions about such a situation. Clearly this situation allows us to define a Hamiltonian system (6) satisfying conditions (ii)-(iv) of Theorem 1, for an input-output map (2) defined by the coefficients $\alpha_N^{J,I}(x_0)$. Thus as in § 1.3 we deduce that Lie algebra $\mathscr{L} = \beta(\mathscr{K})$ is finite-dimensional, and the Lie algebra $\mathscr{S} = \beta(\mathscr{H})$ is a finite-dimensional nilpotent ideal, with descending central series of length N (if it had length less than N then $\mathscr{H}$ would necessarily have step length less then $N+1$ which is a contradiction). Moreover M is diffeomorphic to a Cartesian space, and $\alpha_N^{J,I}$ are all constant functions on M.

Since $\mathscr{L}$ is finite-dimensional and generated by complete vector fields, we may assume that $\mathscr{L}$ is the infinitesimal generator of a Lie transformation group L acting on M, Palais [13]. Since $\mathscr{L}$ is transitive on M, L is transitive on M and then so too

is the nilpotent ideal $S \subset L$ with Lie algebra $\mathscr{S}$. Let K, $H \subset K$ be the simply connected Lie groups with algebras $\mathscr{K}$ and $\mathscr{H}$, respectively. The homomorphism β extends to a homomorphism of groups $b: K \to L$, which in turn exhibits both K and H as Lie transformation groups of M.

2. Graded symplectic vector spaces. Let (V, δ_t) be a graded vector space of degree N, and assume that ω is a symplectic form on V. If ω satisfies $\delta_t^* \omega = t^{N+1}\omega$, or for vector fields X and Y on V we have $\omega(\delta_{t*}X \circ \delta_t^{-1}, \delta_{t*}Y \circ \delta_t^{-1}) = \omega(X, Y)t^{N+1}$; then we say ω is homogeneous symplectic form on V, and the triple (V, δ_t, ω) is a homogeneous graded symplectic vector space. We only deal with homogeneous structures in this paper so we shall not use the term homogeneous from now on. The motivation for considering such a definition comes from the following result.

LEMMA 1. *Given a graded symplectic vector space* (V, δ_t, ω) *then a function* $H \in Q^k$ *if and only if the Hamiltonian vector field* $X_H \in P^{N+1-k}$.

Proof. Assume $H \in Q^k$ is given. Then for any vector field Z on V we have

$$\begin{aligned} t^k\, dH(Z) &= d(H \circ \delta_t)(Z) = dH \circ \delta_t \delta_{t*} Z \\ &= \omega(X_H \circ \delta_t, \delta_{t*}Z) = \omega(\delta_{t*}\delta_{t*}^{-1}X_H \circ \delta_t, \delta_{t*}Z) \\ &= t^{N+1}\omega(\delta_{t*}^{-1}X_H \circ \delta_t, Z). \end{aligned}$$

Thus $dH(Z) = \omega(X_H, Z) = t^{N+1-k}\omega(\delta_{t*}^{-1}X_H \circ \delta_t, Z)$.

Since Z is arbitrary we deduce that

$$\delta_{t*}X_H = t^{N+1-k}X_H \circ \delta_t$$

and hence $X_H \in P^{N+1-k}$. The converse is completely analogous. □

As special cases we see that $X_H \in P^0(P^1)$ if and only if $H \in Q^{N+1}(Q^N)$. Thus in the case of $N = 1$, $X_H \in P^0$ if and only if $H \in Q^2$, and in this case $\delta_t^* \omega = t^2\omega$. It follows that ω is a constant symplectic form, X_H is a linear vector field and H is a quadratic function on V.

Given a vector space V which has the particular form

$$V = V_1 \oplus V_2 \oplus \cdots \oplus V_N \oplus W_N \oplus W_{N-1} \oplus \cdots \oplus W_1$$

with $\operatorname{Dim}(W_i) = \operatorname{Dim}(V_i) = n_i$, $1 \leqq i \leqq N$, we may select linear isomorphisms $\Omega_i: W_i \to V_i^*$ where V_i^* is the dual space of V_i and construct a symplectic form on V by setting

$$\begin{aligned} &\omega(p_i, p_j) = \omega(q_i, q_j) = 0, \qquad 1 \leqq i, j \leqq N, \\ &\omega(q_i, p_j) = \omega(p_j, q_i) = 0, \qquad i \neq j, \\ &\omega(q_i, p_i) = -\omega(p_i, q_i) = \Omega_i p_i(q_i), \qquad 1 \leqq i \leqq N, \end{aligned}$$

where $p_i \in W_i$, $q_i \in V_i$ are also identified with vectors in V all components of which are zero except those of $p_i \in W_i$ and $q_i \in V_i$, respectively. If we define a graded structure on V by setting

$$\delta_t(q_1, q_2, \cdots, q_N, p_N, p_{N-1}, \cdots, p_1) = (tq_1, \cdots, t^N q_N, tp_N, t^2 P_{N-1}, \cdots, t^N p_1)$$

then it is clear that ω is a constant symplectic form on V which satisfies $\delta_t^* \omega = t^{N+1}\omega$. It follows that (V, ω, δ_t) forms a graded symplectic vector space. By identifying W_i and V_i with R^{n_i} and Ω_i with the identity matrix on R^{n_i}, ω is represented in its canonical form which we denote by

$$\Omega_c = \sum_{i=1}^{N} dq_i \wedge dp_i.$$

THEOREM 3. *Given the graded symplectic vector space* $(R^n, \delta_t, \Omega_c)$ *of degree* N *where* R^n *has the form*

$$R^n = R^{n_1} \oplus R^{n_2} \oplus \cdots R^{n_N} \oplus \cdots \oplus R^{n_1}$$

and the dilation δ_t *has the form*

$$\delta_t(q_1, q_2, \cdots, q_N, p_N, \cdots, p_1) = (tq_1, t^2 q_2, \cdots, t^N q_N, tp_N, t^2 p_{N-1}, \cdots, t^N p_1)$$

then if N *is even* $N = 2r$ *there exists a symplectic diffeomorphism* Φ_1 *which intertwines the graded structures* $\Phi_1 : (R^n, \delta_t, \Omega_c) \to (R^n, \delta_t^1, \Omega_c)$ *where the graded structure on* (R^n, δ_t^1) *is defined by*

$$R^n = R^{n_1} \oplus R^{n_N} \oplus R^{n_2} \oplus R^{n_{N-1}} \cdots \oplus R^{n_N} \oplus R^{n_1}$$

$$\begin{aligned} (8) \qquad & \delta_t^1(q_1, q_2, \cdots, q_N, p_N, \cdots, p_1) \\ & = (tq_1, tq_2, \cdots, t^r q_{N-1}, t^r q_N, t^{r+1} p_N, t^{r+1} p_{N-1}, \cdots, t^N p_2, t^N p_1); \end{aligned}$$

if N *is odd* $N = 2r+1$ *there exists a symplectic diffeomorphism* Φ_2 *which intertwines the graded structures* $\Phi_2 : (R^n, \delta_t, \Omega_c) \to (R^n, \delta_t^2, \Omega_c)$ *where the graded structure on* (R^n, δ_t^2) *is defined by*

$$R^n = R^{n_1} \oplus R^{n_N} \oplus R^{n_2} \oplus R^{n_{N-1}} \oplus \cdots \oplus R^{n_N} \oplus R^{n_1}$$

$$\begin{aligned} (9) \qquad & \delta_t^2(q_1, q_2, \cdots, q_N, p_N, \cdots, p_1) \\ & = (tq_1, tq_2, \cdots, t^r q_{N-2}, t^r q_{N-1}, t^{r+1} q_N, t^{r+1} p_N, \cdots, t^N p_2, t^N p_1). \end{aligned}$$

Proof. If $N = 2r$ take Φ_1 to be the map

$$\begin{aligned} & q_i \to q_{2i-1}, \\ & p_{N-i+1} \to -q_{2i}, \qquad 1 \leqq i \leqq r, \\ & q_{N-i+1} \to p_{2i}, \\ & p_i \to p_{2i-1}. \end{aligned}$$

If $N = 2r+1$ take Φ_2 to be the map

$$\begin{aligned} & p_{N-i+1} \to -q_{2i}, \qquad 1 \leqq i \leqq r, \\ & q_{N-i+1} \to p_{2i}, \\ & q_i \to q_{2i-1}, \qquad 1 \leqq i \leqq r+1, \\ & p_i \to p_{2i-1}, \end{aligned}$$

Φ_1 and Φ_2 are clearly diffeomorphisms, and $\delta_t^1 = \Phi_1 \circ \delta_t \circ \Phi_1^{-1}$ satisfies (8) whilst $\delta_t^2 = \Phi_2 \circ \delta_t \circ \Phi_2^{-1}$ satifies (9). That Φ_i preserves Ω_c we take $N = 2r$, $i = 1$ and write

$$\Omega_c = \sum_{i=1}^{r} dq_i \wedge dp_i + \sum_{i=1}^{r} dq_{N-i+1} \wedge dp_{N-i+1}.$$

Thus

$$(\Phi_1^{-1})^* \Omega = \sum_{i=1}^{r} dq_{2i-1} \wedge dp_{2i-1} - \sum_{i=1}^{r} dp_{2i} \wedge dq_{2i} = \Omega_c.$$

The proof in the case $i = 2$, $N = 2r+1$ follows similarly. □

It is interesting to look at the effect of the maps Φ_1 and Φ_2 on functions on the graded vector space (R^n, δ_t). In the case N even all functions in P^N and P^{N+1} with

respect to (R^n, δ_t^1) are affine in the "p" variables, but certainly not with respect to (R^n, δ_t). In the case N odd all functions in P^N with respect to (R^n, δ_t^2) are affine in the "p" variables, but certainly not with respect to (R^n, δ_t). However functions in P^{N+1} with respect to (R^n, δ_t^2) are affine with respect to "p_1" $\cdots$ "p_{N-1}" variables and quadratic with respect to "p_N" variables. Again this is not the case with respect to (R^n, δ_t).

In the remaining sections we shall construct examples of the graded vector spaces $(R^n, \delta_t^1, \Omega_c)$ and $(R^n, \delta_t^2, \Omega_c)$, but by use of Theorem 3 we see that these are equivalent to (R^n, δ_t, Ω) graded vector spaces.

3. Symplectic geometry.

3.1. We briefly recall some facts from symplectic geometry as found in Souriau [15], Pukansky [14], Kostant [10] and nicely summarized in Wallach [18]. Let (M, ω) be a symplectic manifold, and let $\psi: K \times M \to M$ be the action of a Lie transformation group K on M, with Lie algebra $\mathcal{K}$. Let $X^{\#}$ denote the vector field on M induced by the action of K on M, and corresponding to the vector field $X \in \mathcal{K}$. Thus for $x \in M$.

$$X^{\#}(x) = \frac{d}{dt}\psi(\exp tX, x)\Big|_{t=0}.$$

The action of ψ of K on M is said to be Hamiltonian if (i) $X^{\#}$ is infinitesimally symplectic for each $X \in \mathcal{K}$, that is $L_{X^{\#}}\omega = 0$, (ii) there exists a Lie algebra homomorphism λ from $\mathcal{K}$ into the Lie algebra of functions on M under Poisson bracket satisfying $\beta \circ \lambda(X) = X^{\#}$, $X \in \mathcal{K}$.

Let $\mathcal{K}^*$ denote the dual space of $\mathcal{K}$, $k \to Ad\, k$ the adjoint representation of K on $\mathcal{K}$, and $k \to Ad\, k^*$ the coadjoint representation of K on $\mathcal{K}^*$ defined by

$$Ad\, k^* f(X) = f(Ad\, k^{-1} X), \qquad f \in \mathcal{K}^*, \quad X \in \mathcal{K}.$$

Denote the action of K on $\mathcal{K}^*(k, f) \to Ad\, k^* f$ by $\phi: K \times \mathcal{K}^* \to \mathcal{K}^*$. If O_f denotes the orbit of f under K in $\mathcal{K}^*$ then O_f has a symplectic manifold structure with symplectic form Ω^f. If $X \in \mathcal{K}$ let X^* denote the vector field on O_f induced by the action of K on O_f, then

$$\Omega^f(X^*(f), Y^*(f)) = f([X, Y]).$$

If the action of K on M is Hamiltonian, then we define the "moment" map $\tau: M \to \mathcal{K}^*$ by

$$\tau(x)(X) = \lambda(X)(x), \qquad X \in \mathcal{K}, \quad x \in M.$$

The moment map τ is equivariant for the actions ψ and ϕ of K on M and $\mathcal{K}^*$, respectively

$$\tau(\psi(k, x)) = \phi(k, \tau(x)).$$

In the case where K acts transitively on M and we specify $x_0 \in M$ with $f = \tau(x_0)$, then $\tau: M \to O_f$ is a covering map which preserves the symplectic structure, $\tau^* \Omega^f = \omega$.

In the case where K is a nilpotent Lie group then as shown in Wallach [18, p. 302], O_f is diffeomorphic to a cartesian space, and so in this case τ is a diffeomorphsim and M is diffeomorphic to a cartesian space also.

3.2. We now recall some results on polarizations of nilpotent Lie algebras from Kirillov [9] and Auslander and Kostant [1], and conclude by giving implications for solvable and nilpotent Hamiltonian group actions on symplectic manifolds.

If $\mathcal{H}$ is a real Lie algebra and $f \in \mathcal{H}^*$ then we define a subalgebra $\mathcal{H}_f$ by

$$\mathcal{H}_f = \{X \in \mathcal{H}; f([X, Y]) = 0, Y \in \mathcal{H}\}.$$

A real polarization for $\mathcal{H}$ at f is a subalgebra $\mathcal{P} \subset \mathcal{H}$ such that

$\mathcal{P}_R$ (i) $\quad \operatorname{Dim}\left(\frac{\mathcal{H}}{\mathcal{P}}\right) = \frac{1}{2}\operatorname{Dim}\left(\frac{\mathcal{H}}{\mathcal{H}_f}\right),$

$\mathcal{P}_R$ (ii) $\quad f([\mathcal{P}, \mathcal{P}]) = 0.$

Alternatively $\mathcal{P}$ is described as a maximal subspace of $\mathcal{H}$ satisfying $\mathcal{P}_R$ (ii). In particular $\mathcal{P} \supset \mathcal{H}_f$.

A fundamental result of Kirillov [9] is that nilpotent Lie algebras $\mathcal{H}$ admit polarizations for every $f \in \mathcal{H}^*$.

In the case of interest in this paper the nilpotent Lie algebra $\mathcal{H}$ is a codimension one ideal in $\mathcal{K}$ a solvable Lie algebra. If $X \in \mathcal{K}$ is such that $\mathcal{H}$ and X span $\mathcal{K}$ then $t \to \exp t \operatorname{ad} X$ defines a one parameter group Γ of automorphisms of $\mathcal{H}$. In general, given $f \in \mathcal{H}^*$ there may not exist any polarization for $\mathcal{H}$ at f which is invariant under Γ. To deal with this case we introduce complex polarizations, as in Auslander and Kostant [1].

Let $\mathcal{H}_c = \mathcal{H} + i\mathcal{H}$ be the complexification of the real Lie algebra $\mathcal{H}$, and extend $f \in \mathcal{H}^*$ to a linear functional on $\mathcal{H}_c$. A complex polarization for $\mathcal{H}$ at f is a complex subalgebra $\mathcal{P}_c \subset \mathcal{H}_c$ satisfying

$\mathcal{P}_c$ (i) $\quad \mathcal{H}_f \subset \mathcal{P}_c,$

$\mathcal{P}_c$ (ii) $\quad \operatorname{Dim}_{\mathbb{C}} \frac{\mathcal{H}_c}{\mathcal{P}_c} = \frac{1}{2}\operatorname{Dim}_{\mathbb{R}} \frac{\mathcal{H}}{\mathcal{H}_f},$

$\mathcal{P}_c$ (iii) $\quad f([\mathcal{P}_c, \mathcal{P}_c]) = 0,$

$\mathcal{P}_c$ (iv) $\quad \mathcal{P}_c + \bar{\mathcal{P}}_c$ is a Lie subalgebra of $\mathcal{H}_c$

(where $\bar{\mathcal{P}}_c$ denotes complex conjugate of $\mathcal{P}_c$). If θ_f is the bilinear form on $\mathcal{H}$ defined by $\theta_f(X, Y) = f([X, Y])$, then θ_f extends to a bilinear form on $\mathcal{H}_c$ by extending f to $\mathcal{H}_c$. A complex polarization $\mathcal{P}_c \subset \mathcal{H}_c$ may be defined as a maximal subspace of $\mathcal{H}_c$ satisfying $\mathcal{P}_c$ (iii), that is $\theta_f(\mathcal{P}_c, \mathcal{P}_c) = 0$.

If we set $\mathcal{D} = \mathcal{P}_c \cap \mathcal{H}$ and $\mathcal{E} = (\mathcal{P}_c + \bar{\mathcal{P}}_c) \cap \mathcal{H}$ then $\mathcal{D}$ and $\mathcal{E}$ are real subalgebras of $\mathcal{H}$ such that $\mathcal{D}$ is the orthogonal complement of $\mathcal{E}$ with respect to θ_f. Note that since $\mathcal{H}_f \subset \mathcal{D} \subset \mathcal{E}$, $\theta_f(\mathcal{D}, \mathcal{D}) = 0$ so θ_f induces a nondegenerate bilinear form on $\mathcal{E}/\mathcal{D}$. If $\mathcal{E} = \mathcal{D}$ then $\mathcal{P}_c = \mathcal{P} + i\mathcal{P}$ where $\mathcal{P}$ is a real polarization for $\mathcal{H}$ at f. Finally a complex polarization is said to be positive if either $\mathcal{E} = \mathcal{D}$ or

$$-i\theta_f(z, z) \geqq 0 \quad \forall z \in \mathcal{P}_c.$$

When $\mathcal{H}$ is a nilpotent Lie algebra given any $f \in \mathcal{H}^*$ and one parameter group of automorphisms Γ of $\mathcal{H}$ there exists a positive complex polarization for $\mathcal{H}$ at f which is invariant under Γ, ([1, Lem. II.3.1]). Moreover, $\mathcal{D}$ is an ideal in $\mathcal{E}$ such that $\mathcal{E}/\mathcal{D}$ is abelian ([1, Thm. 1.4.10]). We note that since Γ is real and $\mathcal{P}_c$ is invariant under Γ, both of the real subalgebras $\mathcal{D}$ and $\mathcal{E}$ are also invariant under Γ.

3.3. We now show how this structure may be introduced into the situation described in § 1.4. Let $\phi: K \times M \to M$ be the transitive action of the simply connected Lie group K on the symplectic manifold (M, ω), where K has Lie algebra $\mathcal{K}$. The action is clearly Hamiltonian since $\beta(\mathcal{K})$ consists of Hamiltonian vector fields, and

$\beta(X) = X^{\#}$ for each vector field $X \in \mathcal{K}$ by the construction of the action ψ. Clearly ψ restricts to the nilpotent ideal H also giving a transitive Hamiltonian action on M. Let $\phi: K \times \mathcal{K}^* \to \mathcal{K}^*$ be the corresponding action of K on $\mathcal{K}^*$ under the coadjoint representation. Notice that since $\mathcal{H}$ is an ideal in $\mathcal{K}$, ϕ restricts to an action $\phi: K \times \mathcal{H}^* \to \mathcal{H}^*$. If $\tau: M \to \mathcal{H}^*$ is the moment map, $f = \tau(x_0)$ and O_f the orbit of f under $H \subset K$, then since $\mathcal{H}$ is nilpotent $\tau: M \to O_f$ is a diffeomorphism with M and O_f diffeomorphic to cartesian spaces. This of course is in agreement with the observation in § 1.4.

Since τ is equivalent for the actions ψ and ϕ of K on M and $\mathcal{H}^*$, respectively, and $\beta(H_0)(x_0) = 0$, the one parameter group generated by H_0 is in the isotropy group $\{k \in K;\ \phi(k, f) = f\} = \{k \in K;\ \psi(k, x_0) = x_0\}$. In particular, the orbit of f under K coincides with O_f. Let $H_f = \{k \in H;\ \phi(k, f) = f\} = \{k \in H;\ \psi(k, x_0) = x_0\}$. Since H_f is a closed subgroup of a simply connected nilpotent group H, H_f is a simply connected subgroup with Lie algebra $\mathcal{H}_f$. Hence M and O_f may be viewed as the homogeneous space H/H_f with

$$\operatorname{Dim}\left(\frac{H}{H_f}\right) = \operatorname{Dim}\left(\frac{\mathcal{H}}{\mathcal{H}_f}\right) = \operatorname{Dim} O_f = \operatorname{Dim} M.$$

LEMMA 2.

$$ad\, H_0(\mathcal{H}_f) \subset \mathcal{H}_f.$$

Proof. If $X \in \mathcal{H}_f$ and $Y \in \mathcal{H}$ then $f([[H_0, X], Y]) = -f[[X, Y], H_0]) - f([[Y, H_0], X])$. Since $X \in \mathcal{H}_f$ and $\mathcal{H}$ is an ideal in $\mathcal{K}$, $f([[Y, H_0], X]) = 0$. Since the one parameter group generated by H_0 is in the isotropy group $\{k \in K;\ \phi(k, f) = f\} f([[X, Y], H_0]) = 0$. Thus $f([[H_0, X], Y]) = 0$ and since X and Y are arbitrary the result follows. □

Since τ preserves the symplectic structures on M and O_f, for each $X, Y \in \mathcal{H}$ we have

$$\omega(X^{\#}(x_0), Y^{\#}(x_0)) = \Omega^f(X^*(f), Y^*(f)) = \theta_f(X, Y). \tag{10}$$

Note that $X^{\#}(x_0)$ vanishes if and only if $X^*(f)$ vanishes, since then $X \in \mathcal{H}_f$. Now θ_f is in general a degenerate bilinear form on $\mathcal{H}$ whilst ω defines a nondegenerate bilinear form on $T_{x_0}M$. We define an orthogonal complement of subspaces of a vector space V with respect to a bilinear form Ω, degenerate or not, in the usual manner $W^{\perp} = \{X \in V;\ \Omega(X, W) = 0\}$. We than define isotropic (coisotropic) subspaces by $W \subset W^{\perp}$ ($W^{\perp} \subset W$). A Lagrangian subspace W is a maximally isotropic subspace or $W = W^{\perp}$.

We define the linear map $\beta_0: \mathcal{H} \to T_{x_0}M$ by setting $\beta_0(X) = X^{\#}(x_0)$. Thus β_0 has kernel $\mathcal{H}_f$ and gives the following easily verified result.

LEMMA 3. *With respect to the alternating bilinear form on $T_{x_0}M$ defined by ω and θ_f on $\mathcal{H}$ we have for any subspace $W \subset \mathcal{H}$*

$$(\beta_0(W))^{\perp} = \beta_0(W^{\perp}).$$

We deduce that if $\mathcal{P}$ is a polarization for $\mathcal{H}$ at f then $\beta_0(\mathcal{P})$ is a Lagrangian subspace of $T_{x_0}M$ with respect to ω, and if $\mathcal{P}_c$ is a complex polarization for $\mathcal{H}$ at f with corresponding real subalgebras $\mathcal{D} \subset \mathcal{E}$, then $\beta_0(D)^{\perp} = \beta_0(\mathcal{E})$ as subspaces of T_xM with respect to ω. We actually use a more sophisticated version of these results in the next section.

Finally we state for future reference, the following result which summarizes the invariance properties of polarizations described in § 3.2.

LEMMA 4. *Given the situation described in* § 1.4 *set $f = \tau(x_0)$ then either there exists a real polarization $\mathcal{P}$ for $\mathcal{H}$ at f, $\mathcal{P} = \mathcal{P}^{\perp}$, which is invariant under ad H_0, or there exists*

a complex polarization $\mathcal{P}_c$ for $\mathcal{H}$ at f, and associated real subalgebras $\mathcal{D} \subset \mathcal{E} \subset \mathcal{H}$ such that $\mathcal{D}^\perp = \mathcal{E}$ and both $\mathcal{D}$ and $\mathcal{E}$ are invariant under ad H_0.

4. Decompositions.

4.1. In this section we construct global coordinates for a Hamiltonian system constructed from the data described in § 1.4.

We must make definitions before we state the main result that we use. Assume that T is a transitive Lie algebra of vector fields on a manifold M, and a specific point $x_0 \in M$ is given. Let $\mathcal{N} \subset T$ be the subalgebra of vector fields on M which vanish at x_0. Let $\{T^k\}_{k=1}^{N+1}$ be a sequence of subspaces of T with

$$T = T^1 \supset T^2 \supset \cdots \supset T^N \supset \mathcal{N} = T^{N+1}$$

with $T^i(x_0)/T^{i+1}(x_0) \cong R^{n_i}$, $1 \leqq i \leqq N-1$, $T^N(x_0) \cong R^{n_N}$. Choose vector fields X_{ij} in T as follows: $X_{ij} \in T^i$ for $j = 1 \cdots n_i$, $X_{i1}(x_0) \cdots X_{in_i}(x_0)$ are linearly independent and together with $T^{i+1}(x_0)$ span $T^i(x_0)$.

Let $(t, x) \to \gamma_{ij}(t)(x)$ be the flow of X_{ij} and define a local coordinate chart (U, ϕ) for M about x_0 on a suitable neighbourhood U of x_0 where ϕ^{-1} is the map

$$(x_{11}, x_{12} \cdots x_{1n_1}, x_{21} \cdots x_{Nn_N}) \to \gamma_{11}(x_{11}) \circ \gamma_{12}(x_{12}) \circ \cdots \gamma_{1n_1}(x_{1n_1}) \circ \gamma_{21}(x_{21}) \circ \cdots \gamma_{Nn_N}(x_{Nn_n}).$$

We call (U, ϕ) a system of coordinates adapted to the sequence $\{T^k\}_{k=1}^{N+1}$.

In the situation of § 1.4 we have a solvable Lie algebra $\mathcal{L} = \beta(\mathcal{H})$ generated by vector fields $X_{H_i} = \beta(H_i)$ $0 \leqq i \leqq m$, and a nilpotent ideal $\mathcal{S} = \beta(\mathcal{H})$ generated by $X_{H_i} = \beta(H_i)$, $1 \leqq i \leqq m$. Let $\mathcal{O}$ be the subspace of $\mathcal{H}$ spanned by $H_1 \cdots H_m$ and closed under Poisson bracket with H_0. We set $\mathcal{O}^0 = \mathcal{O}$, $\mathcal{O}^k = [\mathcal{O}, \mathcal{O}^{k-1}]$. Thus $\mathcal{H}^k = \mathcal{O}^k + \mathcal{H}^{k+1}$ for $k = 0, \cdots, N$ where $\mathcal{H}^k$ is the descending central series of $\mathcal{H}$. Now $\beta(\mathcal{O}) = \mathcal{R}$, the subspace of $\mathcal{S}$ spanned by $X_{H_1} \cdots X_{H_m}$ and closed under Lie bracket with X_{H_0}. Setting $\mathcal{R}^{k+1} = \beta(\mathcal{O}^k)$ for $k = 0, \cdots, N-1$, $\mathcal{R} = \mathcal{R}^1$, $\mathcal{R}^k = [\mathcal{R}, \mathcal{R}^{k-1}]$ and $\mathcal{S}^k = \mathcal{R}^k + \mathcal{S}^{k+1}$ where $\mathcal{S}^k$ is the descending central seires of $\mathcal{S}$. (The indexing is chosen to coincide with that in Crouch [4]). Note that $\mathcal{R}^{N+1} = \{0\}$ which corresponds to the fact that $\mathcal{H}^N$ consists of constant functions. The following result, to which the above situation will be applied, follows from Crouch [4].

THEOREM 4. *Let $X_0, X_1, \cdots, X_m$ be complete vector fields on a manifold M which generate a solvable transitive Lie algebra, with nilpotent ideal $\mathcal{S}$ of codimension one generated by $X_1 \cdots X_m$. Assume $X_0(x_0) = 0$ for $x_0 \in M$. Let $\{\mathcal{R}^k\}$ be the sequence of subspaces of $\mathcal{S}$ as defined above, $\{\mathcal{S}^k\}$ the descending central series of $\mathcal{S}$ of length N, and the $\mathcal{N} \subset \mathcal{S}$ the subalgebra consisting of vector fields vanishing at x_0. Assume that $\mathcal{R}^k \cap \mathcal{N} = \mathcal{R}^k \cap (\mathcal{S}^{k+1} + \mathcal{N})$ for $k = 1 \cdots N-1$. Then there exists a global system of coordinates (M, Φ) adapted to the sequence of subalgebras*

$$\mathcal{S} = \mathcal{S}^1 \supset \mathcal{S}^2 + \mathcal{N} \supset \mathcal{S}^3 + \mathcal{N} \supset \cdots \supset \mathcal{S}^N + \mathcal{N} \supset \mathcal{N}$$

such that $\Phi(M) = R^n$ can be given a graded vector space structure $(R^n, \delta_t) R^n = R^{n_1} \oplus \cdots \oplus R^{n_N}$ $\delta_t(x_1 \cdots x_N) = (tx_1, t^2x_2, \cdots, t^N x_N)$ satisfying

(i) $\Phi_* X_0 \circ \Phi^{-1} \in P^0$.

(ii) $\Phi_* X_i \circ \Phi^{-1} \in P^1$, $\quad 1 \leqq i \leqq m$.

(iii) $\Phi(x_0) = 0$, Φ_* *maps the subspace $\mathcal{S}^k(x_0) \subset T_{x_0}M$ onto the subspace $R^{n_k} \oplus \cdots \oplus R^{n_N}$ of R^n.*

(iv) *The linear part of $\Phi_* X_0 \circ \Phi^{-1}$ is just the block diagonal matrix representation of $-\text{ad}\, X_0$ on S/N induced by the coordinate system and corresponding to the invariant subspaces $R^i + \mathcal{N}/\mathcal{N}$.*

The significance of the condition $R^k \cap \mathcal{N} = \mathcal{R}^k \cap \mathcal{S}^{k+1} + \mathcal{N})$ is indicated by the following result.

LEMMA 5. *In the situation of* § 1.4 *we have* $\mathcal{H}_f + \mathcal{H}^{n-k} = \mathcal{H}^{k\perp}$ *for* $k = 0, \cdots, N$ *and* $\mathcal{R}^k \cap \mathcal{N} = \mathcal{R}^k \cap (\mathcal{S}^{k+1} + \mathcal{N})$ *for* $k = 1, \cdots, N-1$. *Equivalently,* $T_{x_0}M$ *is the direct sum decomposition* $T_{x_0}M = R^1(x_0) \oplus \cdots \oplus R^N(x_0)$, *with* $\operatorname{Dim}(R^i(x_0)) = \operatorname{Dim}(R^{N+1-i}(x_0))$, *and* $\omega(R^i(x_0), R^j(x_0)) = 0$ *for* $i + j \neq N+1$.

Proof. By taking appropriate linear combinations and applying the Jacobi relation to the functions $\alpha_r^{J,I}$ defined in (7), it is clear that Assumption (iv) in § 1.4 is equivalent to

$$\theta_f(\mathcal{O}^j, \mathcal{O}^k) = 0 \quad \text{for } j + k \leqq N-2,$$

since $f = \tau(x_0)$ with $\tau(x_0)(X) = X(x_0)$ for $X \in \mathcal{H}$. Now Assumption (iii) in § 1.4 implies that $\theta_f(\mathcal{O}^j, \mathcal{O}^k) = 0$ for $j + k \geqq N$. In particular $\theta_f(\mathcal{O}^j, \mathcal{O}^k) \neq 0$ if and only if $j + k = N-1$.

From (10) we see that

$$\theta_f(X, Y) = \omega(\beta_0(X), \beta_0(Y)), \qquad X \in \mathcal{O}^j, \quad Y \in \mathcal{O}^k$$

so $\omega(\mathcal{R}^k(x_0), R^j(x_0)) = 0$ for $j + k \neq N+1$. By Assumption (ii) in § 1.4 $\beta_0(\mathcal{H}) = \mathcal{S}(x_0)$ spans $T_{x_0}M$. Thus by the nondegeneracy of ω given $Z \in \mathcal{O}^{N-j-1}$ such that $\beta_0(Z) \neq 0$ there exists $Y \in \mathcal{O}^j$ such that

$$\omega(\beta_0(Z), \beta_0(Y)) = \theta_f(Z, Y) \neq 0 \quad \text{for } j = 0, \cdots, N-1.$$

In particular, $\mathcal{R}^i(x_0) \cap \mathcal{R}^j(x_0) = \{0\}$ for $i \neq j$ since $\mathcal{R}^i(x_0) = \beta_0(\mathcal{O}^{i-1})$. It follows that $\mathcal{R}^i(x_0) \cap \mathcal{S}^{i+1}(x_0) = \{0\}$. That is

$$\{0\} = (\mathcal{R}^i + \mathcal{N}/\mathcal{N} \cap (\mathcal{S}^{i+1} + \mathcal{N})/\mathcal{N} = \frac{\mathcal{R}^i \cap (\mathcal{S}^{i+1} + \mathcal{N}) + \mathcal{N}}{\mathcal{N}}.$$

Hence $\mathcal{R}^i \cap (\mathcal{S}^{i+1} + \mathcal{N}) \subset \mathcal{N} \cap \mathcal{R}^i$. It is clear that this is equivalent to the direct sum decomposition of $T_{x_0}M$ by the spaces $R^i(x_0)$.

We must show $\mathcal{H}_f + \mathcal{H}^{N-k} = \mathcal{H}^{k\perp}$. Since $[\mathcal{H}^j, \mathcal{H}^k] \subset \mathcal{H}^{j+k+1}$, by definition of the descending central series, and $\mathcal{H}^{N+1} = \{0\}$ it is clear that $\mathcal{H}_f + \mathcal{H}^{N-k} \subset \mathcal{H}^{k\perp}$. Conversely, if $Z \in \mathcal{H}^{k\perp}$ then $\theta_f(Z, \mathcal{O}^j) = 0$ for $j \geqq k$. If $Z \notin \mathcal{H}_f + \mathcal{H}^{N-k}$, then $Z \in \mathcal{O}^0 + \cdots + \mathcal{O}^{N-k-1}$ and $\beta_0(Z) \neq 0$. If $Z = \sum_{i=0}^{N-k-1} Z_i$ with $Z_i \in \mathcal{O}^{N-k-i}$, then $0 = \theta_f(Z, \mathcal{O}^j) = \theta_f(Z_{N-j-1}, \mathcal{O}^j) = \omega(\beta_0(Z_{N-j-1}), \beta_0(\mathcal{O}^j))$ for $j \geqq k$. Now not all $\beta_0(Z_{N-j-1}) = 0$ since $\beta_0(Z) \neq 0$. But this contradicts our previous statement. Thus $\mathcal{H}_f + \mathcal{H}^{N-k} = \mathcal{H}^{k\perp}$. □

Remark 1. In the original context of Theorem 5 in Crouch [4], a less specific version of Lemma 5 is valid, but still yields the equivalence of $\mathcal{R}^k \cap \mathcal{N} = \mathcal{R}^k \cap (\mathcal{S}^{k+1} + \mathcal{N})$ with the direct sum decomposition of $T_{x_0}M$ by spaces $\mathcal{R}^k(X_0)$. The choice of coordinate system adapted to $\{\mathcal{S}^k + \mathcal{N}\}_{k=1}^{N+1}$ in Theorem 5 is characterized by the selection of the vector fields $X_{ij} \in \mathcal{R}^i$, which is possible by the above property.

4.2. We now construct certain sequences of subspaces of the Lie algebra $\mathcal{H}$, using the subalgebras $\mathcal{P}$, $\mathcal{D}$, $\mathcal{E}$ and $\mathcal{H}_f$ constructed in § 3. As in Assumption (iii) of § 1.4 we assume H has a descending central series of length $N+1$ and hence

$$\mathcal{H} = \mathcal{H}^0 \supset \mathcal{H}^1 \supset \cdots \supset \mathcal{H}^{N-1} \supset \mathcal{H}_f \supset \{0\};$$

here $\mathcal{H}^N$ consists of constant functions so $\mathcal{H}^N \subset \mathcal{H}_f$. We also have the orthogonal complement sequence, with respect to θ_f

$$\mathcal{H} = \mathcal{H}_f^{\perp} \supset \mathcal{H}^{N-1\perp} \supset \cdots \supset \mathcal{H}^{1\perp} \supset \mathcal{H}^{0\perp} = \mathcal{H}_f.$$

We must consider four cases.

Case 1. $N=2r$ there exists real polarization $\mathcal{P}$ at f invariant under $ad\, H_0$.

Case 2. $N=2r+1$ there exists real polarization $\mathcal{P}$ at f invariant under $ad\, H_0$.

Case 3. $N=2r$ there exists a complex polarization $\mathcal{P}_c$ at f invariant under $ad\, H_0$.

Case 4. $N=2r+1$ there exists a complex polarization $\mathcal{P}_c$ at f invariant under $ad\, H_0$.

In Cases 1 and 2 we have the following inclusions:

$$\mathcal{P}\cap\mathcal{H}^{k\perp}+\mathcal{H}^{N-K}\supset\mathcal{P}\cap\mathcal{H}^{k\perp}+\mathcal{H}^{N-k+1}\supset\mathcal{P}\cap\mathcal{H}^{k-1\perp}+\mathcal{H}^{N-k+1} \tag{11}$$

for $k=1,2,\cdots,N$.

In Case 3 we have the following inclusions:

$$E\cap\mathcal{H}^{k\perp}+\mathcal{H}^{N-k}\supset\mathcal{E}\cap\mathcal{H}^{k\perp}+\mathcal{H}^{N-k+1}\supset\mathcal{E}\cap\mathcal{H}^{k-1\perp}+\mathcal{H}^{N-k+1}$$

for $k=N,\ N-1,\cdots,r+1$,

$$\begin{gathered}\mathcal{E}\cap\mathcal{H}^{r\perp}+\mathcal{H}^{r}\supset\mathcal{D}\cap\mathcal{H}^{r\perp}+\mathcal{H}^{r},\\ \mathcal{D}\cap\mathcal{H}^{k\perp}+\mathcal{H}^{N-k}\supset\mathcal{D}\cap\mathcal{H}^{k\perp}+\mathcal{H}^{N-k+1}\supset\mathcal{D}\cap\mathcal{H}^{k-1\perp}+\mathcal{H}^{N-k+1}\end{gathered} \tag{12}$$

for $k=1,2,\cdots,r$.

In Case 4 we have the following inclusions:

$$\mathcal{E}\cap\mathcal{H}^{k\perp}+\mathcal{H}^{n-k}\supset\mathcal{E}\cap\mathcal{H}^{k\perp}+\mathcal{H}^{N-k+1}\supset\mathcal{E}\cap\mathcal{H}^{k-1\perp}+\mathcal{H}^{N-k+1}$$

for $k=N,\ N=1,\cdots,r+2$,

$$\begin{gathered}\mathcal{E}\cap\mathcal{H}^{r+1\perp}+\mathcal{H}^{r}\supset\mathcal{E}\cap\mathcal{H}^{r+1\perp}+\mathcal{H}^{r+1}\supset\mathcal{D}\cap\mathcal{H}^{r+1\perp}+\mathcal{H}^{r+1}\supset\mathcal{D}\cap\mathcal{H}^{r\perp}+\mathcal{H}^{r+1},\\ \mathcal{D}\cap\mathcal{H}^{k\perp}+\mathcal{H}^{N-k}\supset\mathcal{D}\cap\mathcal{H}^{k\perp}+\mathcal{H}^{N-k+1}\supset\mathcal{D}\cap\mathcal{H}^{k-1\perp}+\mathcal{H}^{N-k+1}\end{gathered} \tag{13}$$

for $k=1,2,\cdots,r$.

LEMMA 6. *For $N=2r$ and $1\leqq k,j\leqq r$ or for $N=2r+1$ and $1\leqq k\leqq r$, $1\leqq j\leqq r+1$ we have the following identities concerning subalgebras of $\mathcal{H}$.*

$$\begin{aligned}(\mathcal{P}\cap\mathcal{H}^{N-k\perp}+\mathcal{H}^{k})^{\perp}&=\mathcal{P}\cap\mathcal{H}^{k\perp}+\mathcal{H}^{N-k},\\ (\mathcal{P}\cap\mathcal{H}^{N-j+1\perp}+\mathcal{H}^{j})^{\perp}&=\mathcal{P}\cap\mathcal{H}^{j\perp}+\mathcal{H}^{N-j+1},\\ (\mathcal{E}\cap\mathcal{H}^{N-k\perp}+\mathcal{H}^{k})^{\perp}&=\mathcal{D}\cap\mathcal{H}^{k\perp}+\mathcal{H}^{N-k},\\ (\mathcal{E}\cap\mathcal{H}^{N-j+1\perp}+\mathcal{H}^{j})^{\perp}&=\mathcal{D}\cap\mathcal{H}^{j\perp}+\mathcal{H}^{N-j+1}.\end{aligned}$$

Proof. As in Lemma 5 $\mathcal{H}^{N-k}\subset\mathcal{H}^{k\perp}$, so $\mathcal{H}^{N-k+1}\subset\mathcal{H}^{k\perp}$ also. If $X,\ Y\in\mathcal{H}^{k\perp}$ and $Z\in\mathcal{H}^{k-1}$ then

$$\theta_f([X,Y],Z)=f([[X,Y],Z])=-f([[Y,Z],X])-f([[Z,X],Y]).$$

Since $[Y,Z]$ and $[Z,X]$ belong to $\mathcal{H}^{k}$ and $X,\ Y\in\mathcal{H}^{k\perp}$ the last two terms vanish. Thus $[\mathcal{H}^{k\perp},\mathcal{H}^{k\perp}]\subset\mathcal{H}^{k-1\perp}\subset\mathcal{H}^{k\perp}$. The lemma is now obvious once it is observed that $\mathcal{D}$, $\mathcal{E}$ and $\mathcal{P}$ are all subalgebras of $\mathcal{H}$ satisfying $\mathcal{P}=\mathcal{P}^{\perp}$, $\mathcal{D}=\mathcal{E}^{\perp}$. □

We may now apply Lemma 5 to the expressions in (11), (12) and (13), whilst remembering that $\mathcal{H}_f\subset\mathcal{P}$ and $\mathcal{H}_f\subset\mathcal{D}\subset\mathcal{E}$.

In Cases 1 and 2 we have the following inclusions:

$$\mathcal{H}^{N-k}+\mathcal{H}_f\supset\mathcal{P}\cap\mathcal{H}^{n-k}+\mathcal{H}^{N-k+1}+\mathcal{H}_f\supset\mathcal{H}^{N-k+1}+\mathcal{H}_f \tag{14}$$

for $k=1,2,\cdots,N$.

In Case 3 we have the following inclusions:

$$\mathcal{H}^{N-k}+\mathcal{H}_f \supset \mathcal{E}\cap\mathcal{H}^{N-k}+\mathcal{H}^{N-k+1}+\mathcal{H}_f \supset \mathcal{H}^{N-k+1}+\mathcal{H}_f \tag{15}$$

for $k=N,\ N-1,\cdots,r+1$,

$$\mathcal{H}^{N-k}+\mathcal{H}_f \supset \mathcal{D}\cap\mathcal{H}^{N-k}+\mathcal{H}^{N-k+1}+\mathcal{H}_f \supset \mathcal{H}^{N-k+1}+\mathcal{H}_f$$

for $k=1,2,\cdots,r$.

In Case 4 we have the following inclusions:

$$\mathcal{H}^{N-k}+\mathcal{H}_f \supset \mathcal{E}\cap\mathcal{H}^{N-k}+\mathcal{H}^{N-k+1}+\mathcal{H}_f \supset \mathcal{H}^{N-k+1}+\mathcal{H}_f$$

for $k=N,\ N-1,\cdots,r+2$,

$$\begin{aligned}&\mathcal{H}^{r}+\mathcal{H}_f \supset \mathcal{E}\cap\mathcal{H}^{r}+\mathcal{H}^{r+1}+\mathcal{H}_f \supset \mathcal{D}\cap\mathcal{H}^{r}+\mathcal{H}^{r+1}+\mathcal{H}_f \supset \mathcal{H}^{r+1}+\mathcal{H}_f,\\ &\mathcal{H}^{N-k}+\mathcal{H}_f \supset \mathcal{D}\cap\mathcal{H}^{N-k}+\mathcal{H}^{N-k+1}+\mathcal{H}_f \supset \mathcal{H}^{N-k+1}+\mathcal{H}_f\end{aligned} \tag{16}$$

for $k=1,2,\cdots,r$.

We may now use these sequences of subalgebras to obtain our initial coordinate representation of the situation described in § 1.4.

THEOREM 5. *In the situation of* § 1.4 *we meet the hypotheses of Theorem* 4 *by setting* $\beta(H_i)=X_i,\ 0\leqq i\leqq m,\ \mathcal{S}^{k+1}=\beta(\mathcal{H}^k)k=0,\cdots,N-1,\ \mathcal{N}=\beta(\mathcal{H}_f)$. *We obtain a diffeomorphism from M onto a graded vector space* (R^n,δ_t) *of degree N with the properties listed in Theorem* 4. *The graded vector space structure may be refined by adapting the coordinates to the homomorphic images under* β *of the sequences of subalgebras of* $\mathcal{H}$ *in* (14)-(16), *depending on whether N is even or odd, and whether or not there exists a real ad* H_0 *invariant polarization for* $\mathcal{H}$ *at f. This distinguishes the four cases listed above. In Cases* 1 *and* 3 *the resulting graded structure corresponds to that defined in equation* (8). *In Cases* 2 *and* 4 *the resulting graded structure corresponds to that defined in equation* (9). *In Case* 4, *however, we must add an extra subspace.*

Proof. The only hypotheses of Theorem 4 not trivially satisfied by the situation described in § 2.4 are the assumptions $\mathcal{R}^k\cap\mathcal{N}=\mathcal{R}^k\cap(\mathcal{S}^{k+1}+\mathcal{N})$. However, these are shown to be satisfied in Lemma 5. It is clear that in refining the coordinates to the homomorphic images under β of the sequences of subspaces in (14) and (15) we may define correspondingly refined graded vector spaces, which in Cases 1 and 3 is that defined in (8) and in Case 2 is that defined in (9), whilst retaining properties (i)-(iii) in Theorem 4.

In Case 4, which defines the sequences of subspaces in (16), there is an added inclusion

$$\mathcal{H}^{r}+\mathcal{H}_f \supset \mathcal{E}\cap\mathcal{H}^{r}+\mathcal{H}^{r+1}+\mathcal{H}_f \supset \mathcal{D}\cap\mathcal{H}^{r}+\mathcal{H}^{r+1}+\mathcal{H}_f \supset \mathcal{H}^{r+1}+\mathcal{H}_f.$$

By Lemma 6 we have (with respect to θ_f)

$$(\mathcal{E}\cap\mathcal{H}^{r}+\mathcal{H}^{r+1}+\mathcal{H}_f)^{\perp}=\mathcal{D}\cap\mathcal{H}^{r}+\mathcal{H}^{r+1}+\mathcal{H}_f.$$

Thus by standard symplectic algebra there exists a subspace $\mathcal{P}_L\subset H$ such that $\mathcal{P}_L^{\perp}=\mathcal{P}_L$ and

$$\mathcal{H}^{r}+\mathcal{H}_f \supset \mathcal{E}\cap\mathcal{H}^{r}+\mathcal{H}^{r+1}+\mathcal{H}_f \supset \mathcal{P}_L \supset \mathcal{D}\cap\mathcal{H}^{r}+\mathcal{H}^{r+1}+\mathcal{H}_f \supset \mathcal{H}^{r+1}+\mathcal{H}_f. \tag{17}$$

It is now clear that we can also refine the coordinates in Case 4, using the homomorphic images under β of the sequences of subspaces in (16) and (17), to give the graded vector space structure defined in equation (9), whilse also retaining properties (i)-(iii) of Theorem 4.

The dimensional constraints on the graded vector spaces defined in (8) and (9) are also valid in this case. To see this we simply apply Lemma 3 to the identities in Lemma 6, (whilst remembering that $\mathcal{H}^{N-k}+\mathcal{H}_f=\mathcal{H}^{k\perp}$ from Lemma 5), and use the non-degeneracy of the bilinear form ω on $T_{x_0}M$.

It remains to consider to what extent we may retain property (iv) of Theorem 4. By Lemmas 2 and 4 the subalgebras $\mathcal{H}_f$, $\mathcal{P}$, $\mathcal{D}$, and $\mathcal{E}$ may be invariant under $ad\, H_0$. Note also that $\mathcal{H}^k$ is invariant under $ad\, H_0$ also. It follows that all subalgebras encountered in the sequences of subalgebras (14), (15), (16) are therefore invariant under $ad\, H_0$. It is not true however that the subspace $\mathcal{P}_L$ in (17) is invariant under $ad\, H_0$. Indeed it is this noninvariance which causes us to seek complex polarizations $\mathcal{P}_c$ which are invariant.

Notice that each block A^k in the matrix representation of $-ad\,\beta(H_0)$ on $\mathcal{S}/\mathcal{N}$, where A^k is its representation on $\mathcal{R}^k+\mathcal{N}/\mathcal{N}$, $1\leqq k\leqq N$ by (iv) of Theorem 4, is also equal to the induced representation on $((\mathcal{S}^k+\mathcal{N})/\mathcal{N}/(\mathcal{S}^{k+1}+\mathcal{N})/\mathcal{N}$ since $\mathcal{R}^k\cap\mathcal{N}=\mathcal{R}^k\cap(\mathcal{S}^{k+1}+\mathcal{N})$, and hence is also equal to the induced representation of $-ad\, H_0$ on

$$\frac{(\mathcal{H}^{k-1}+\mathcal{H}_f)/\mathcal{H}_f}{(\mathcal{H}^k+\mathcal{H}_f)/\mathcal{H}_f}.$$

Since all of the extra subspaces introduced in (14)-(16), are "sandwiched" between subspaces $\mathcal{H}^{k+1}+\mathcal{H}_f$ for some k, and these are invariant under $-ad\, H_0$, it follows that A^k has a block triangular form. In all but Case 4 we have

$$A^k=\begin{pmatrix}A^k_{11} & 0\\ A^k_{21} & A^k_{22}\end{pmatrix}$$

for each k but in Case 4 and $k=r+1$, $N=2r+1$ corresponding to the subspaces in (17) we have

$$A^k=\begin{pmatrix}A^k_{11} & 0 & 0 & 0\\ A^k_{21} & A^k_{22} & A^k_{23} & 0\\ A^k_{21} & A^k_{32} & A^k_{33} & 0\\ A^k_{41} & A^k_{42} & A^k_{43} & A^k_{44}\end{pmatrix}.$$

□

4.3. In this subsection we work out further details concerning the coordinate representation of the situation in § 1.4 obtained in Theorem 5.

LEMMA 7. *Let $\Phi: M\to R^n$ be the diffeomorphism obtained in Theorem 5 from the symplectic manifold (M,ω) onto the graded vector space (R^n,δ_t). Then $\sigma=\Phi^{-1*}\omega$ defines a symplectic form on R^n, such that (R^n,δ_t,σ) is a graded symplectic vector space.*

Proof. In this lemma the initial graded vector space structure obtained in Theorem 5 supplies sufficient structure for the proof of the result. Using the notation at the beginning of § 4.1, we abbreviate the map Φ^{-1} by $(x^1,x^2,\cdots,x^N)\to\gamma^1(x^1)\circ\cdots\circ\gamma^N(x^N)=\phi(x)(x_0)$ where x^i represents each component of the coordinates $(x_{i1}),\cdots,(x_{in_i})$ and γ^i represents each of the flows $\gamma_{i1},\cdots,\gamma_{in_i}$. In this case each flow γ_{ij} corresponds to a Hamiltonian vector field $X_{h_{ij}}$ which belongs to $\mathcal{R}^i$ by Remark 1 in § 4.1; thus h_{ij} belongs to $\mathcal{O}^{i-1}$. Abbreviate $X_{h_{ij}}$ by X^i and h_{ij} by h^i.

It now follows that

$$\begin{aligned}\sigma_x\left(\frac{\partial}{\partial x^i},\frac{\partial}{\partial x^j}\right)&=((\Phi^{-1})^*\omega)_x\left(\frac{\partial}{\partial x^i},\frac{\partial}{\partial x^j}\right)\\&=\omega_{\Phi^{-1}(x)}\left((\Phi^{-1})_*\frac{\partial}{\partial x^i},\Phi_*^{-1}\frac{\partial}{\partial x^j}\right)\\&=\omega_{\phi(x)(x_0)}\left(\frac{\partial}{\partial x^i}\phi(x)(x_0),\frac{\partial}{\partial x^i}\phi(x)(x_0)\right).\end{aligned}$$

Now

$$\frac{\partial}{\partial x^i}\phi(x)(x_0) = \gamma^1(x^1)_* \cdots \gamma^{i-1}(x^{i-1})_* X^i(\gamma^i(x^i) \circ \cdots \circ \gamma^N(x^N)(x_0))$$

$$= \phi(x)_* \gamma^N(-x^N)_* \cdots \gamma^i(-x^i)_* X^i(\gamma^i(x^i) \circ \cdots \circ \gamma^N(x^N)(x_0)).$$

Since $p \to \phi(x)(p)$, $\phi(x): M \to M$ is a composition of symplectic maps $p \to \gamma_{ij}(x_{ij})(p)$ for each $x \in R^n$, it follows that $\phi(x)^*\omega = \omega$. Moreover by Jacobi's theorem, Abraham and Marsden [21], if $\psi: M \to M$ is a symplectic map and X_h is a Hamiltonian vector field with Hamiltonian h, then $\psi_*^{-1} X_h \circ \psi = X_{h \circ \psi}$. It follows that

$$\sigma_x\left(\frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j}\right) = \omega_{x_0}(X_{h^i \circ \gamma^i(x^i) \circ \cdots \circ \gamma^N(x^N)}, X_{h^j \circ \gamma^j(x^j) \circ \cdots \circ \gamma^N(x^N)}).$$

By definition of the Poisson bracket we may express this as

$$[h^i \circ \gamma^i(x^i) \circ \cdots \circ \gamma^N(x^N), h^j \circ \gamma^j(x^j) \circ \cdots \circ \gamma^N(x^N)](x_0).$$

Up to sign we may represent this as a sum of iterated Poisson brackets

$$\sum_{\substack{k_N, \cdots, k_i \\ l_N, \cdots, l_i}} [(ad\, h^N)^{k_N} \cdots (ad\, h^i)^{k_i}(h^i), (ad\, h^N)^{l_N} \cdots (ad\, h^j)^{l_j}(h^j)](x_0)$$

$$\times \frac{(x^N)^{k_N}}{k_N!} \cdots \frac{(x^i)^{k_i}}{k_i!} \frac{(x^N)^{l_N}}{l_N!} \cdots \frac{(x^j)^{l_j}}{l_j!}.$$

Since each $h^k \in \mathscr{H}$, a nilpotent Lie algebra this is a polynomial on $x^N, \cdots, x^{\min(i,j)}$.

We must verify that σ whose i,jth component is given above satisfies $\delta_t^* \sigma = t^{N+1}\sigma$, where the dilation δ_t is defined in our abbreviated terminology by

$$\delta_t(x^1, \cdots, x^N) = (tx^1, \cdots, t^N x^N).$$

This is equivalent to showing that the above polynomial expression lies in $\mathscr{O}^{N+1-i-j}$. However by construction each $h^j \in \mathscr{O}^{j-1}$, and Lemma 5 Assumptions (iii) and (iv) in § 1.4 imply that $\theta_f(\mathscr{O}^l, \mathscr{O}^k) = 0$ unless $l+k = N-1$ or $l-1+k-1 = N+1$, from which the required result follows. □

COROLLARY 1. *In the situation of Lemma 7 the components of the symplectic form σ on R^n evaluated at zero are given by*

$$\sigma_0\left(\frac{\partial}{\partial x_{il}}, \frac{\partial}{\partial x_{jk}}\right) = \omega_{x_0}(X_{il}, X_{jk})$$

for $1 \leqq i \leqq n_l$, $1 \leqq j \leqq n_k$, $1 \leqq k, l \leqq N$.

Proof. Evaluate the expression in (19) at $x = 0$ and replace the abbreviations for the nomenclature used at the beginning of § 4.1. □

LEMMA 8. *In the situation of Theorem 5 the coordinate system for the refined graded vector space may be chosen so that $\sigma = (\Phi^{-1^*}\omega)$ takes its canonical form at $x = 0$.*

Proof. By Corollary 1 and Remark 1 it is clear that special attention must be paid to the choice of the vector fields $X_{ij} \in \mathscr{R}^i$.

We use the following result which can be found in Irving [8], and is derived using standard technqiues of symplectic linear algebra. Let (V, ω) be a symplectic vector space, with a sequence of subspaces $\{V_k\}_{k=1}^{2q+1}$

$$V = V_1 \supset V_2 \supset \cdots \supset V_{q+1} \supset \cdots \supset V_{2q-1} \supset V_{2q} \supset \{0\} = V_{2q+1}$$

satisfying $V_{2q-k+2}^{\perp} = V_k$ for $k = 1, \cdots, 2q+1$.In particular $V_{q+1} = V_{q+1}^{\perp}$ is a Lagrangian subspace. Assume that $V_k = U_k \oplus V_{k+1}$ is a direct sum decomposition for $k = 1, \cdots, 2q$.

Then we may choose a basis for V consisting of vectors selected only from the spaces U_k such that the resulting coordinate system for V displays ω in the form

$$\left(\begin{array}{c|c} 0 & I \\ \hline -I & 0 \end{array}\right)$$

and

$$I = \begin{pmatrix} & & & I_{n_q} \\ & 0 & \iddots & \\ & I_{n_2} & & \\ & & 0 & \\ I_{n_1} & & & \end{pmatrix} \qquad n_k = \text{Dim } U_k,\ k = 1, \cdots, q.$$

The partitioning obviously corresponds to the subspaces V_k. By suitable labelling of the coordinates, as in (8) and (9), we obtain the canonical representation of a symplectic form.

In the situation considered here we apply Lemmas 3, 6 and 5 in turn to the sequences of subspaces in (14)-(16), adding the subspace $\mathscr{P}_L$ defined in (17) to the sequence in (16). This defines sequences of subspaces of $T_{x_0}M$ as above, in each of the cases one to four, which satisfy the required orthogonality conditions. Although the corresponding subspaces U_k are not uniquely defined, we have the expression

$$T_{x_0}M = \mathscr{R}^1(x_0) \oplus \mathscr{R}^2(x_0) \oplus \cdots \oplus \mathscr{R}^N(x_0).$$

In the situation above either V_k coincides with a direct sum $\mathscr{R}^j(x_0) \oplus \cdots \oplus \mathscr{R}^N(x_0) = \mathscr{S}^j(x_0)$ or $\mathscr{S}^j(x_0) \supsetneq V_k \supsetneq \mathscr{S}^{j+1}(x_0)$. Moreover, in this latter case we have $\mathscr{S}^j(x_0) \supset V_{k-1} \supset V_k \supset V_{k+1} \supset \mathscr{S}^{j+1}(x_0)$. It follows that each subspace U_k may be defined as a subspace of some $\mathscr{R}^j(x_0)$ subspace. We deduce that there exists a basis of $T_{x_0}M$ with each element contained in the intersection of some $\mathscr{R}^j(x_0)$ subspace and some V_k subspace, such that in the coordinate system defined by this basis ω_{x_0} is in its canonical form. We now select the vector fields X_{ij} so as to coincide with this basis when evaluated at x_0. This choice meets all the conditions imposed in Theorem 5, and by Corollary 1 ensures that σ evaluated at $x = 0$ is in canonical form. □

Combining these results we obtain the main result of this section.

THEOREM 6. *In the situation of* § 1.4 *there exists a symplectic diffeomorphism* Φ *from* M *onto a graded symplectic vector space* (R^n, δ_t, σ), $\Phi(x_0) = 0$, *with* δ_t *defined in* (8) *or* (9) *and* $\sigma(0)$ *in canonical form.*

Further, with respect to this graded structure $H_0 \circ \Phi^{-1} \in Q^{N+1}$ *and* $H_i \circ \Phi^{-1} \in Q^N$, $1 \leqq i \leqq m$.

Proof. We apply Theorem 5 to obtain an initial diffeomorphism Φ from M onto a graded vector space (R^n, δ_t) with δ_t defined in (8) or (9), and such that $\Phi_* \beta(H_0) \circ \Phi^{-1} \in P^0$, $1 \leqq i \leqq m$, $\Phi_* \beta(H_i) \circ \Phi^{-1} \in P^{-1}$, and $\Phi(x_0) = 0$. By Lemma 7 ${\Phi^{-1}}^* \omega = \sigma$ makes (R^n, δ_t, σ) into a graded symplectic vector space, with Φ a symplectic map. By Lemma 8 we modify our initial choice of Φ so that $\sigma(0)$ is in canonical form. Since Φ is symplectic $\Phi_* \beta(H_i) \circ \Phi^{-1} = \beta(H_i \circ \Phi^{-1})$, $0 \leqq i \leqq m$ with respect to σ. By Lemma 1 we deduce that $H_0 \circ \Phi^{-1} \in Q^{N+1}$, $H_i \circ \Phi^{-1} \in Q^N$. □

5. A Darboux–Weinstein theorem. In this section we are given a graded symplectic vector space of degree N, (R^n, δ_t, ω), where ω is a nonconstant form on R^n. We wish to show that there is a global diffeomorphism ϕ of R^n such that

(i) $\phi^* \bar{\omega} = \omega$ where $\bar{\omega}$ is the constant symplectic form on R^n which is equal to $\omega(0)$.

(ii) ϕ preserves the graded structure of (R^n, δ_t). That is ϕ commutes with the dilation δ_t, $\phi \circ \delta_t = \delta_t \circ \phi$.

We note that if X is a vector field on (R^n, δ_t) such that $X \in P^0$, then $\delta_{t*}X = X \circ \delta_t$. Therefore, if $(s, x) \to \psi(s)(x)$ is the flow of X we have

$$\delta_t \circ \psi(s) = \psi(s) \circ \delta_t.$$

Therefore we may synthesize diffeomorphisms ϕ as above from the flows of vector fields in P^0.

We need a technical result before showing that we can in fact achieve such a result.

LEMMA 9. *If $(s, x) \to X(s)(x)$ is a time varying vector field on a graded vector space (R^n, δ_t) of degree N such that for each s $X(s) \in P^0$ and for each $x \in R^n$ the map $s \to X(s)(x)$ is continuous then X is complete. That is its flow*

$$(s, u, x) \to \phi(s, u)(x) \qquad \phi(u, u)(x) = x$$

$$\frac{d\phi}{ds}(s, u)(x) = X(s)(\phi(s, u)(x))$$

is C^1 and defined on $R \times R \times R^n$. Moreover each map $x \to \phi(s, u)(x)$ commutes with δ_t.

Proof. The fact that $X(s) \in P^0$ implies that we may write $X(s)$ in the coordinates of $\mathbb{R}^n$ as

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \vdots \\ \dot{x}_N \end{bmatrix} = \begin{bmatrix} A_1(t)x_1 \\ A_2(t)x_2 + a_2(t, x_1) \\ \vdots \\ A_N(t)x_N + a_N(t, x_1 \cdots x_{N-1}) \end{bmatrix}$$

where $a_k(s, \delta_t x) = t^k a_k(s, x) = t^k a_k(s, x_1 \cdots x_{k-1})$. Each of the components of the matrices $A_k(t)$ and the coefficients of the polynomials $a_k(t, x)$ are continuous in t by assumption. This implies that the fundamental solution of $\dot{x}_k(s) = A_k(s)x_k(s)$, $x_k(u) = x_k^0$, is C^1, $x_k(s) = \Phi(s, u)x_k^0$, and defined for all $(s, u) \in \mathbb{R} \times \mathbb{R}$. This in turn implies by successive integration of the equations above that the vector field X is complete. That $\delta_t(\phi(s, u)(x)) = \phi(s, u)(\delta_t(x))$ follows as before from $\delta_{t*}X(s)(x) = X(s)(\delta_t(x))$. □

To prove the Darboux-Weinstein result we need the following construction, on a graded vector space of degree $N(R^n, \delta_t)$ with closed two form ω, satisfying $\delta_t^* \omega = t^{N+1}\omega$. We note that nondegeneracy is not required of ω.

If $Z(t)$ is the vector field $d\delta_s/ds|_{s=t}$ along δ_t then

$$\frac{d}{ds}(\delta_s^* \omega)_{s=t} = \delta_t^* L_{Z(t)}\omega = \delta_t^* i(Z(t))\, d\omega + d\delta_t^*(i(Z(t))\omega).$$

Since $d\omega = 0$, δ_0 is the zero map and δ_1 is the identity map we may integrate this equation with respect to t on $[0, 1]$ to obtain $\omega = dI(\omega)$ where

$$I(\omega) = \int_0^1 \delta_s^*(i(Z(s))\omega)\, ds.$$

We note that $\delta_s \circ \delta_t = \delta_t \circ \delta_s$ for $t, s > 0$, and hence $\delta_{t*}Z(s) = Z(s) \circ \delta_t$. Thus

$$\delta_t^* I(\omega) = \int_0^1 \delta_s^*(i(\delta_{t*}^{-1}Z(s) \circ \delta_t)\delta_t^* \omega)\, ds$$

$$= \int_0^1 \delta_s^*(i(Z(s))t^{N+1}\omega)\, ds = t^{N+1} I(\omega).$$

Thus $I(\omega)$ also satisfies $\delta_t^* I(\omega) = t^{N+1} I(\omega)$. We prove our version of the Darboux-Weinstein theorem in Weinstein [19].

THEOREM 7. *Given a graded symplectic vector space* (R^n, δ_t, ω) *with polynomic symplectic form* ω, *there exists a diffeomorphism* ϕ *of* R^n, *which preserves the graded structure and satisfies* $\phi^*\omega = \bar{\omega}$, *where* $\bar{\omega}$ *is the constant symplectic form defined by* $\omega(0)$. *Moreover the linear part of* ϕ *is the identity map.*

Proof. Let $\Omega = \omega - \bar{\omega}$ and $\Omega_t = \bar{\omega} + t\Omega$. Now $d\Omega = O$ and $\delta_t^*\Omega = t^{N+1}\Omega$. It follows by the preceding argument that $\sigma = I(\Omega)$ satisfies $d\sigma = \Omega$ and $\delta_t^*\sigma = t^{N+1}\sigma$. We also claim that for any t, Ω_t is a nondegenerate bilinear form on the whole of R^n. Consider the matrix representation of ω with respect to the coordinate system of the graded vector space. It therefore has a corresponding block structure corresponding to the decomposition $R^n = R^{n_1} \oplus \cdots \oplus R^{n_N}$. Identifying $\partial/\partial x^i$ with any vector field $\partial/\partial x_{ij}$ as in the proof of Lemma 7 we may write

$$\omega = \sum_{k=-(N-1)}^{N-1} \omega_k, \quad \text{where } \omega_k\left(\frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j}\right) = \delta_{i+j,N+1-k}\omega\left(\frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j}\right)$$

for $-(N-1) \leqq k \leqq (N-1)$.

It follows that the components α_k of ω_k are polynomials which satisfy

$$\alpha_k \circ \delta_t = t^k \alpha_k.$$

Since ω is polynomic in these coordinates, we must have $\omega_{-1} = \cdots = \omega_{-(N-1)} = 0$ and ω_0 constant. Moreover, since ω_k for $k > 0$ vanish at $x = 0$ we see that $\omega_0 = \bar{\omega}$.

It follows that $\Omega_t = \bar{\omega} + \sum_{k=1}^{N-1} t\omega_k$ where $\bar{\omega}$ has the block form

$$\begin{pmatrix} & & & A_N \\ & 0 & \iddots & \\ & A_2 & & 0 \\ A_1 & & & \end{pmatrix}$$

and ω_k has the block form

$$\begin{pmatrix} 0 & & B_{N-k}(x) & 0 \\ & \ddots & & \ddots \\ B_1(x) & 0 & & \\ \vdots & \ddots & & \\ 0 & & & 0 \end{pmatrix}.$$

Since $\bar{\omega}$ is invertible it follows that Ω_t is also invertible for any t and any $x \in R^n$. Thus Ω_t defines an isomorphism $\Omega_t^\# : T_x R^n \to T_x R^n$.

We define a time varying vector field on R^n by $Y(t) = -\Omega_t^{\#-1}(\sigma)$. Clearly $\Omega_t^{\#-1}$ is continuous (even linear) in t. We claim that $Y(t) \in P^0$ for each t, also. However, by definition $\Omega_t(Y(t), Z) = -\sigma(Z)$ and $\delta_s^*\Omega_t = s^{N+1}\Omega_t$, $\delta_s^*\sigma = s^{N+1}\sigma$. Thus it is clear that $\delta_{s*}Y(t) = Y(t) \circ \delta_s$ which simply states that $Y(t) \in P^0$ for each t as claimed.

We may now apply Lemma 9 to see that $Y(t)$ is complete. Let $(t, x) \to \phi(t, x)$ denote the map $(t, x) \to \phi(t, 0)(x)$ where ϕ is the flow of Y. It follows that

$$\begin{aligned} \frac{d}{dt}\phi_t^*\Omega_t &= \phi_t^* \frac{d\Omega_s}{ds}\bigg|_{s=t} + \phi_t^* L_{Y(t)}\Omega_t \\ &= \phi_t^*\Omega + \phi_t^* d(i(Y(t))\Omega_t) + \phi_t^* i(Y(t))\, d\Omega_t \end{aligned}$$

$$= \phi_t^* \Omega - \phi_t^* \, d\sigma \quad \text{since } d\Omega_t = 0$$

$$= 0 \quad \text{since } \Omega = d\sigma.$$

Therefore by integrating with respect to t for $t \in [0, 1]$ we obtain $\phi_1^* \Omega_1 = \Omega_0$. Thus the required diffeomorphism ϕ is just ϕ_1.

Now $\Omega = \omega - \bar{\omega} = \sum_{k=1}^{N-1} \omega_k$, where the components of ω_k belong to Q^k. Since $\Omega = d\sigma$ we may write $\sigma = \sum_{k=1}^{N-1} \sigma_k$ where the components of σ belong to Q^{k+1}. In particular, σ contains no components with linear terms. On the other hand, $\Omega_t^{\#^{-1}} = \bar{\omega}^{\#^{-1}} + t\alpha$ for some matrix α, with polynomial coefficients. It follows that $Y(t) = -\Omega_t^{\#^{-1}}(\sigma)$ is a vector field with polynomial coefficients containing no linear, or constant terms. This shows that the linear part of the map $x \to \phi(1, 0)(x)$, is the identity map as desired. □

6. Summary.

6.1. In this final section we draw together the results of the previous sections, and give some examples.

By combining Theorems 6 and 7 we obtain the following abstract result.

THEOREM 8. *Given a symplectic manifold* (M, ω), *a specific point* $x_0 \in M$, *a solvable Lie algebra* $\mathcal{H}$ *of functions on* M *generated by* $H_1 \cdots H_m$, *satisfying the assumptions* (i)-(iv) *of* § 1.4, *then there exists a diffeomorphsim* $\Phi : M \to R^n$ *onto a graded symplectic vector space* (R^n, δ_t, σ) *such that* $\Phi(x_0) = 0$, $\Phi^* \sigma = \omega$, σ *is in its canonical representation, and with respect to the graded structure*

$$H_0 \circ \Phi^{-1} \in Q^{N+1}, \quad H_i \circ \Phi^{-1} \in Q^N, \quad 1 \leqq i \leqq m.$$

We may also give the following control theoretic version of this result.

THEOREM 9. *Given an homogeneous Volterra series* (2) *which is realizable by a Hamiltonian system* (5), *there exists another Hamiltonian realization on a graded symplectic vector space, with the symplectic form exhibited in canonical form, which satisfies all of the conditions* (i)-(iv) *to Theorem* 1.

This result completely answers the problem posed in § 1.3. The partial converse of Theorem 9 is also true and mimics Theorem 2. By employing Theorem 3 we can also ensure linearity in the "p" variables.

THEOREM 10. *Given a Hamiltonian system* (5) *on a graded symplectic vector space* $(R^n, \delta_t, \Omega_c)$ *of degree* N, *defined by* $H_0 \in Q^{N+1}$ *and* $H_i \in Q^N$, $1 \leqq i \leqq m$, *then the input-output map is a homogeneous Volterra series* (2). *By symplectic change of coordinates onto another graded symplectic vector space* $(R^n, \delta_t', \Omega_c)$ *we can assume that all functions* H_i, $1 \leqq i \leqq m$ *are affine in the "*p*" coordinates and polynomial in the "*q*" variables.*

Linearity in the "p" coordinates for the function H_0 is discussed in § 2.

Of course it is this representation with linearity in the "p" coordinates which is obtained in Theorems 8 and 9. Moreover, the discussion about linearity in the "p" coordinates for the function H_0 is concerned eventually with the four cases described in § 4.2. Indeed the discussion concerns most importantly the quadratic terms in the H_0 function, or equivalently the linear terms in the corresponding Hamiltonian vector field X_{H_0}. This is discussed at length in Theorem 5, and results are also valid for the representation obtained in Theorem 6. In the Darboux-Weinstein Theorem 7 it is shown that the final coordinate change is the identity in its linear component which in turn shows that this structure for the linear part of X_{H_0}, or quadratic part of H_0 is still valid in the final representations presented in Theorems 8 and 9. The interested reader should have no difficulty in writing down explicit forms for each of the four cases described in § 4.2. See also Irving [8] for more details concerning these systems.

We conjecture that similar results are also true in the nonhomogeneous case, where we consider arbitrary finite Volterra series, or equivalently dispense with Assumption (iv) in § 1.4. Indeed, most of the results in § 4 can be carried out successfully using inclusions (11)-(13). We do, however, lose linearity on the "p" variables. It has not been possible to obtain a corresponding global version of the Darboux-Weinstein theorem however.

6.2. In this subsection we present some examples of accessible Hamiltonian systems all with respect to the canonical symplectic structure.

Example 1. $R^4 = R^2 \oplus R^2$, is a graded vector space of degree $N = 2$.

$$H_0 = p_2 q_1 + q_2 q_1^2, \qquad H_1 = p_1,$$
$$\dot{q}_1 = u, \qquad \dot{q}_2 = q_1,$$
$$\dot{p}_1 = -p_2 - 2q_2 q_1, \qquad \dot{p}_2 = -q_1^2,$$
$$y = p_1.$$

In this example there exists an $ad\, H_0$ invariant polarization for $\mathcal{H}$.

Example 2. $R^6 = R^2 \oplus R^2 \oplus R^2$, is a graded vector space of degree $N = 3$,

$$H_1 = p_1, \qquad H_0 = p_2 q_1 + p_3 q_1^2 + q_2 q_1^3$$
$$\dot{q}_1 = u, \quad \dot{q}_2 = q_1, \quad \dot{q}_3 = q_1^2,$$
$$\dot{p}_3 = -q_2^2, \quad \dot{p}_2 = -q_1^3 - 2q_2 q_3, \quad \dot{p}_1 = -2p_3 q_1 - p_2 - 3q_2 q_1^2,$$
$$y = p_1.$$

In this example there exists an $ad\, H_0$ invariant polarization for H.

Example 3. R^2 is a graded vector space of degree $N = 1$.

$$H_1 = p_1, \qquad H_0 = \tfrac{1}{2}(p_1^2 + q_1^2),$$
$$\dot{q}_1 = p_1 + u, \qquad \dot{p}_1 = -q_1,$$
$$y = p_1.$$

In this example there is no real $ad\, H_0$ invariant polarization for $\mathcal{H}$.

Example 4. $R^4 = R \oplus R^2 \oplus R$ is a graded vector space of degree

$$N = 3, \quad H_1 = p_1, \quad H_0 = p_2 q_1^2 + \tfrac{1}{2}(p_2^2 + q_2^2),$$
$$\dot{q}_1 = u, \quad \dot{q}_2 = q_1^2 + p_2, \quad \dot{p}_2 = -q_2, \quad \dot{p}_1 = -2p_2 q_1,$$
$$y = p_1.$$

In this example there is no real $ad\, H_0$ invariant polarization for $\mathcal{H}$.

REFERENCES

[1] L. AUSLANDER AND B. KOSTANT, *Polarizations and unitary representations of solvable Lie groups*, Inventiones Math., 14 (1977), pp. 255-354.

[2] R. W. BROCKETT, *Volterra series and geometric control theory*, Automatica, 12 (1976), pp. 167-176.

[3] K. T. CHEN, *Decompositions of differential equations*, Maths Annals. 146 (1962), pp. 263-278.

[4] P. E. CROUCH, *Dynamical realizations of finite Volterra series*, this Journal, 19 (1981), pp. 177-202.

[5] P. E. CROUCH AND M. IRVING, *On the finite Volterra series which admit Hamiltonian realizations*, to appear in Mathematical Systems Theory 1984.

[6] J. B. GONCALVES, *Nonlinear controllability and observability with applications to gradient systems*, Ph.D. Thesis, Univ. Warwick, Coventry, 1981.

[7] R. W. GOODMANN, *Nilpotent Lie groups: structure and applications to analysis*, Lecture Notes in Mathematics 562, Springer-Verlag, Berlin, 1976.
[8] M. IRVING, *Hamiltonian systems with nilpotent structures*, Ph.D. Thesis, Univ. Warwick, Coventry, 1983.
[9] A. A. KIRILLOV, *Unitary representations of nilpotent Lie groups*, Russian Math. Surveys, 17 (1962), pp. 53-104.
[10] B. KOSTANT, *Quantization and unitary representations*, C. T. Taam, ed., Lecture Notes in Mathematics 170, Springer-Verlag, Berlin (1970), pp. 87-208.
[11] A. J. KRENER, *A decomosition theory for differential systems*, this Journal, 15 (1977), pp. 813-829.
[12] C. M. LESIAK AND A. J. KRENER, *The existence and uniqueness of Volterra series for nonlinear systems*, IEEE Trans. Automat. Control, AC-23 (1970), pp. 1090-1095.
[13] R. S. PALAIS, *A global formulation of the Lie theory of transitive groups*, Memoirs of the A.M.S. No. 22, 1957.
[14] L. PUKANSKY, *Leçons sur la représentation des groupes*, Soc. Math. de France, Monographie Dunon, Paris, 1967.
[15] J. M. SOURIAU, *Structure des systèmes dynamiques*, Maîtrises de Mathématiques, Dunod, Paris, 1970.
[16] H. J. SUSSMANN, *Existence and uniqueness of minimal realizations of nonlinear systems*, Math. Systems Theory, 10 (1977), pp. 263-284.
[17] A. J. VAN DER SCHAFT, *Controllability and observability for affine nonlinear systems*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 490-492.
[18] N. R. WALLACH, *Symplectic Geometry and Fourier Analysis*, Mathematical Sciences Press, Brookline, MS, 1977.
[19] A. WEINSTEIN, *Symplectic manifolds and their Lagrangian submanifolds*, Adv. Mathematics, 6 (1971), pp. 329-346.
[20] PH. B. ZWART AND W. M. BOOTHBY, *On compact homogeneous symplectic manifolds*, Annales de l'Institut Fourier, de l'Université de Grenoble, 30 (1980), pp. 129-157.
[21] R. ABRAHAM AND J. E. MARSDEN, *Foundations of Mechanics*, Benjamin, Cummings, 1978.
[22] A. J. VAN DER SCHAFT, *Linearization of Hamiltonian and gradient systems*, IMA J. Math. Control Inform. (1984).
[23] ———, *System theoretic descriptions of physical systems*, Doctoral Dissertation, Univ. Groningen, Mathematical Centre Tracts, Mathematical Centre, Amsterdam, 1983.
[24] R. W. BROCKETT, *Control theory and analytical mechanics*, in the 1976 Ames Research Centre (NASA) Conference on Geometric Control Theory, C. Martin and R. Hermann, eds., in Lie Groups: History Frontiers and Applications, Mathematical Science Press, Brookline, MA, 1977.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 3, May 1986

005

MINIMUM VARIANCE CONTROL OF DISCRETE TIME MULTIVARIABLE ARMAX SYSTEMS*

U. SHAKED† AND P. R. KUMAR‡

Abstract. We consider multivariable ARMAX stochastic systems. These systems can incorporate the following complicating features: general delay structures, nonminimum phase transfer functions, different dimensions for input and output vectors. We obtain the control laws which minimize the variance of the output process while maintaining system stability.

Key words. minimum variance control, vector ARMAX systems, nonminimum phase systems

1. Introduction. We consider multivariable linear stochastic systems in an ARMAX format:

$$A(z)y(t) = z^d B(z)u(t) + C(z)w(t). \tag{1}$$

Here z is the *backward* shift operator: $zy(t) := y(t-1)$. $y(t) \in \mathbb{R}^m$ is the output, $u(t) \in \mathbb{R}^l$ is the input and $w(t) \in \mathbb{R}^m$ is a white noise process, i.e. it is wide sense stationary, $Ew(t) = 0$ and covariance $Ew(t)w^T(s) = Q\delta_{ts}$.

$$A(z) = I + \sum_{i=1}^{n} A_i z^i. \tag{2i}$$

$$B(z) = B_0 + \sum_{i=1}^{n} B_i z^i, \qquad B_0 \neq 0, \quad B(z) \text{ is of full rank.} \tag{2ii}$$

$$C(z) = C_0 + \sum_{i=1}^{n} C_i z^i, \qquad C^{-1}(z) \text{ is analytic inside the closed unit disc.} \tag{2iii}$$

(2iv) d, the delay, is an integer with $d \geqq 1$.

We shall define as "admissible", control laws which are of the form $u(t) = M(z)y(t)$ where

(3i) $M(z)$ is a matrix of rational functions;

(3ii) $M(z)$ is analytic at $z = 0$.

The condition (3ii) restricts us to the set of nonanticipative control laws, while (3i) is imposed merely for convenience.

We shall further say that an admissible control law $u(t) = M(z)y(t)$ is "stabilizing" if the four transfer functions

$$\begin{aligned} &M(z)[I - z^d A^{-1}(z)B(z)M(z)]^{-1}, \qquad [I - z^d A^{-1}(z)B(z)M(z)]^{-1}, \\ &[I - z^d M(z)A^{-1}(z)B(z)]^{-1}, \qquad z^d A^{-1}(z)B(z)[I - z^d M(z)A^{-1}(z)B(z)]^{-1} \end{aligned} \tag{4}$$

are all analytic inside the closed unit disc. The restrictions in (4) are imposed so that the resulting closed-loop system is internally stable.

* Received by the editors May 29, 1984, and in revised form January 25, 1985.

† Department of Electronics Systems, Tel-Aviv University, Tel Aviv, Israel.

‡ Department of Electrical and Computer Engineering and the Coordinated Science Laboratory, University of Illinois, Urbana, Illinois 61801. The research of the second author has been supported by the National Science Foundation under grant ECS-8304435 and grant ECS-85-06628 and the U.S. Army Research Office under contract DAAG29-84-K-0005 (administered through the Laboratory for Information and Decision Systems at the Massachusetts Institute of Technology).

Our goal in this paper is to find a control law, from among the set of all admissible stabilizing control laws, which minimizes the variance $Ey^T(t)y(t)$ of the output process in steady state.

For single-input, single-output (i.e., $m = l = 1$) minimum phase systems, the problem has been solved by Astrom [1]. The minimum variance control law is shown to be

$$u(t) = -\frac{-G(z)}{B(z)F(z)}\, y(t) \tag{5i}$$

where $F(z)$, a polynomial of degree $d-1$, and $G(z)$, a polynomial, satisfy

$$C(z) = A(z)F(z) + z^d G(z). \tag{5ii}$$

If the system is of nonminimum phase, then while the above control law still minimizes the variance of the output process from among the set of all admissible control laws, it is *not* however stabilizing. To satisfy stability, one must "sacrifice" some variance. This constrained optimization problem of obtaining a control law which minimizes the output variance over the set of all admissible, stabilizing control laws, for single-input, single-output systems has been solved by Peterka [2]. It is shown to be

$$u(t) = -\frac{S(z)}{R(z)}\, y(t) \tag{6i}$$

where $R(z)$, a polynomial of degree $(n+d-1)$, and $S(z)$, a polynomial, satisfy

$$B^*(z)C(z) = A(z)R(z) + z^d B(z)S(z). \tag{6ii}$$

Here, $B^*(z)$ is the minimum phase spectral factor of $B(z)B(z^{-1})$.

In the multi-input, multi-output case, Borison [3] has considered the situation where (i) the number of inputs is equal to the number of outputs, (5ii) B_0 is invertible and (iii) $B(z)$ is of minimum phase, i.e. $\det B(z) \neq 0$ for $0 < |z| \leqq 1$. Under these conditions, the optimal solution is given by multivariable analog of (5i, ii). This treatment is not fully general from several points of view. Firstly, conditions (i) and (iii) are restrictive. Secondly, the restriction that B_0 is invertible, condition (ii), means that by defining a new control $\bar{u}(t) := B_0 u(t)$, we really have a system where for each output variable there is one special input variable which influences that output variable after other input variables have ceased to influence it. Moreover, the different output variables will be influenced by their special input variables with the same delay. This simplifies the problem considerably and in fact one outgrowth of this restriction is that the control law really minimizes, separately, the variance of each output variable, or equivalently, the same control law simultaneously minimizes $Ey^T(t)Ry(t)$ for all $R \geqq 0$. We shall see that this situation is not true in general.

In another treatment of the multi-input, multi-output case, Goodwin, Ramadge, and Caines [4] assume that $A(z) = (1 + \alpha_1 z + \cdots + \alpha_n z^n)I$ where $\alpha_1, \cdots, \alpha_n$ are scalars. Stability of the solution is not considered, but use is made of the solution only when $d = 1$, the number of inputs is equal to the number of outputs, B_0 is invertible, and the system is of minimum phase, i.e. $\det B(z) \neq 0$ for $0 < |z| \leqq 1$, in which situation there are no problems. Recently Dugard, Goodwin, and Xianya [6] have studied the minimum variance problem through an examination of the role of the interactor matrix, and have obtained the minimum variance control law when the interactor matrix is diagonal.

We also refer the reader to Bayoumi and El Bagoury [5] for some errors in previous attempts to deal with the problem of minimum variance control of multi-variable systems.

In this paper, our goal is to treat all the complications caused by (i) B_0 possibly singular, i.e. general delay structures, (ii) nonminimum phase systems, i.e. det $B(z)$ possibly vanishing in $0<|z|<1$ and (iii) rectangular systems where the number of inputs is different from the number of outputs. Throughout, we address the problem of minimizing $Ey^T(t)y(t)$ while maintaining system stability.

If one wishes to minimize $Ey^T(t)Ry(t)$ for some positive definite R, then this is easily accomplished by defining $\bar{y}(t) := R^{1/2}y(t)$, $\bar{A}(z) := R^{1/2}A(z)R^{-1/2}$, $\bar{B}(z) := R^{1/2}B(z)$, $\bar{C}(z) := R^{1/2}C(z)$ and considering the system $\bar{A}(z)\bar{y}(t) = z^d\bar{B}(z)u(t) + \bar{C}(z)w(t)$, which satisfies assumptions (2i-iv).

Our treatment proceeds in the order of increasing generality. In § 2 we treat systems with general delay structures, with the solution given by Theorems 2.1, 2.2 and 2.3. In § 3 we treat nonminimum phase systems, with the solution given in Theorem 3.1 and finally in § 4 we treat rectangular systems, with the solution provided in Theorem 4.1.

2. Nonuniform delay systems. In this section we obtain the admissible, stabilizing, minimum variance control law for the multivariable ARMAX system (1), when it has a general delay structure. For this reason we allow det $B(z)$ to have zeros at the origin, because such zeros correspond to nonuniform transmission delays in different input-output channels.

Except for such zeros at the origin, we assume that the system is of a minimum phase, i.e., det $B(z) \neq 0$ for $0<|z|<1$. The system is also assumed to have the same number of inputs and outputs, i.e. it is square.

The complete solution for this problem is furnished by the following three Theorems.

THEOREM 2.1. *Suppose there exist $F(z)$ and $G(z)$ which satisfy*:

(7i) $F(z) = \sum_{i=0}^{d+p-1} F_i z^i$ *for some p, and F_0 is invertible.*

(7ii) $G(z)$ *is a matrix of rational functions which are analytic at $z=0$.*

(7iii) $\lim_{z\to 0} z^d F^T(z^{-1})A^{-1}(z)B(z) = 0.$

(7iv) $C(z) = A(z)F(z) + z^d B(z)G(z).$

Then, the admissible, stabilizing control law which minimizes the variance $Ey^T(t)y(t)$ of the output, is

$$u(t) = -G(z)F^{-1}(z)y(t).$$

The resulting minimum variance is

$$Ey^T(t)y(t) = \operatorname{tr} \sum_{i=0}^{d+p-1} F_i^T F_i Q.$$

THEOREM 2.2. *Define the following*:

(8i) *Let $\sum_{i=0}^{\infty} D_i z^i$ be a power series expansion of $A^{-1}(z)B(z)$.*

(8ii) *Let p be the largest power of z^{-1} in $B^{-1}(z)A(z)$.*

(8iii) *Let $E_0, E_1, \cdots, E_p$ be matrices satisfying*

$$B^{-1}(z)A(z) = E_p z^{-p} + E_{p-1}z^{-p+1} + \cdots + E_0 + o(1).$$

(8iv) *Let*

$$\mathcal{W}_m^n := \begin{bmatrix} 0 & \cdots & 0 & D_m \\ \vdots & & D_m & D_{m+1} \\ 0 & \cdots & \vdots & \vdots \\ D_m & \cdots & D_{n-1} & D_n \end{bmatrix} \quad \text{and} \quad \mathcal{E}_m^n := \begin{bmatrix} E_n & 0 & \cdots & 0 \\ E_{n-1} & E_n & & \vdots \\ \vdots & \vdots & \ddots & 0 \\ E_m & E_{m+1} & \cdots & E_n \end{bmatrix}.$$

Then, the matrix

$$[\mathcal{W}_0^{p-1}, \mathcal{E}_1^{pT}]$$

has full rank.

THEOREM 2.3. *Define the following*:

(9i) *Define* $F_0, \cdots, F_{d-1}$ *recursively by* $F_0 := C_0$ *and*

$$F_k := C_k - \sum_{i=1}^{k} A_i F_{k-i} \quad \text{for } k = 1, \cdots, d-1.$$

(9ii) *Define* $H_d, H_{d+1}, \cdots$ *as the coefficient matrices in the power series expansion*

$$A^{-1}(z)C(z) - \sum_{i=0}^{d-1} F_i z^i =: \sum_{i=d}^{\infty} H_i z^i.$$

(9iii) *Let* K *and* J *be matrices which satisfy the linear system of equations*

$$[\mathcal{W}_0^{p-1}, \mathcal{E}_1^{pT}][K^T, J^T]^T = [H_d^T, \cdots, H_{d+p-1}^T]^T.$$

(9iv) *Define* $F_d, \cdots, F_{d+p-1}$ *by* $[F_d^T, \cdots, F_{d+p-1}^T] := \mathcal{E}_1^{pT} J$.

(9v) *Define* $F(z) := \sum_{i=0}^{d+p-1} F_i z^i$ *and* $G(z) := z^{-d} B^{-1}(z)[C(z) - A(z)F(z)]$.

Then, $F(z)$ *and* $G(z)$ *satisfy* (7i-iv).

The significance of the three Theorems 2.1, 2.2 and 2.3 is the following. Theorem 2.1 gives sufficient conditions for the solution $u(t) = -G(z)F^{-1}(z)y(t)$ to be optimal. Theorem 2.2 asserts that a certain matrix is of full rank. Theorem 2.3 uses the solution of a system of linear equations, guaranteed to exist by Theorem 2.2, to construct $F(z)$ and $G(z)$ which satisfy the sufficient conditions of Theorem 2.1. Thus, we have a constructive procedure for obtaining an admissible, stabilizing, minimum variance control law.

One useful property of the minimum variance control law is that it does *not* depend on the noise covariance Q. Thus, the *same* control law is optimal irrespective of the noise covariance.

As we have mentioned earlier at the end of § 1, the above theorems can be employed to solve the problem of minimizing $Ey^T(t)Ry(t)$ for any positive definite R. However, in general, the solution will depend on R. This means, in particular, that the control law of Theorems 2.1, 2.2 and 2.3 does not separately minimize the variance of each output variable. This differentiates the case $\det B_0 \neq 0$, considered in [3], from the general delay structures considered here.

The minimum variance $\operatorname{tr} \sum_{i=0}^{d+p-1} F_i^T F_i Q$ can be decomposed into two parts. $\operatorname{tr} \sum_{i=d}^{d+p-1} F_i^T F_i Q$ can be regarded as the increase in variance resulting from the singularity of B_0, while the remaining part $\sum_{i=0}^{d-1} F_i^T F_i Q$ is the variance due to the delay of d time units. In the case considered in [3], only the latter part is present.

The proofs of Theorems 2.1, 2.2 and 2.3 follow immediately from Lemmas 2.4–2.10 below.

LEMMA 2.4. *Suppose $F(z)$ is a matrix of polynomials, which, together with a certain $G(z)$ satisfies* (7ii, iii *and* iv). *Let $u(t)=M(z)y(t)$ be any admissible control law which is applied to the system* (1). *Then, the output $y(t)$ of the closed loop system can be decomposed as $y(t)=y_1(t)+y_2(t)$ with*

$$y_1(t)=F(z)w(t), \qquad y_2(t)=z^dA^{-1}(z)B(z)[G(z)+T(z)A^{-1}(z)C(z)]w(t)$$

where

$$T(z):=M(z)[I-z^dA^{-1}(z)B(z)M(z)]^{-1}.$$

Furthermore, the two components $y_1(t)$ and $y_2(t)$ are uncorrelated.

Proof. The closed loop system is clearly $Ay=z^dBMy+Cw$, and so $y=(I-z^dA^{-1}BM)^{-1}A^{-1}Cw$ and $u=TA^{-1}Cw$. Substituting for u, we therefore get $Ay=z^dBTA^{-1}Cw+Cw$. Using (7iv) for C gives the required decomposition for the closed-loop output y. Throughout this paper we shall evaluate variances of processes by contour integration along the unit circle. This is possible, since by our stability assumption (4), the control laws considered give a stable closed loop system. To see that the two components are uncorrelated, note first that

$$\operatorname{cor}(Fw, z^dA^{-1}B[G+TA^{-1}C]w)$$
$$=\frac{\operatorname{tr}}{2\pi i}\oint F^T(z^{-1})z^dA^{-1}(z)B(z)[G(z)+T(z)A^{-1}(z)C(z)]Q\frac{dz}{z}$$

where, here and in the sequel, the contour is a circle centered at the origin and with unit radius. Note that $F^T(z^{-1})$ has a singularity only at the origin. Now $G(z)$ is analytic at the origin, by assumption. Also, because $M(z)$ is analytic at the origin, so is $T(z)$, and therefore also $T(z)A^{-1}(z)C(z)$. Utilizing (7iii), we see that the integrand vanishes at the origin. Moreover, the integrand also has no singularities elsewhere in the unit disc since the control is stabilizing, and so the above integral vanishes. □

LEMMA 2.5. *Suppose that $F(z)$ and $G(z)$ satisfy* (7i–iv). *Then, the control law which minimizes $Ey^T(t)y(t)$ over the set of all admissible control laws is $u(t)=M(z)y(t)$, where*

$$M(z)=-G(z)F^{-1}(z)$$

and the resulting minimum variance is

$$Ey^T(t)y(t)=\operatorname{var}(F(z)w(t))=\operatorname{tr}\sum_{i=0}^{p+d-1}F_i^TF_iQ.$$

Proof. From Lemma 2.4 it follows that for an admissible choice of M,

$$Ey^T(t)y(t)=\operatorname{var}(F(z)w(t))$$
$$+\frac{\operatorname{tr}}{2\pi_i}\oint[G(z^{-1})+T(z^{-1})A^{-1}(z^{-1})C(z^{-1})]^TB^T(z^{-1})A^{-T}(z^{-1}) \tag{10}$$
$$\cdot A^{-1}(z)B(z)[G(z)+T(z)A^{-1}(z)C(z)]Q\frac{dz}{z}.$$

Since $F(z)$ does not depend on the choice of $M(z)$, the best that one can hope to do, if one wishes to minimize the variance, is to choose $M(z)$ so that the integral

on the right-hand side above is zero. One way to do this is to choose $M(z)$ so as to make $G(z)+T(z)A^{-1}(z)C(z)=0$, i.e. $T(z)=-G(z)C^{-1}(z)A(z)$. Since $T=M[I-z^dA^{-1}BM]^{-1}$, M would have to be chosen so that $T^{-1}=M^{-1}-z^dA^{-1}B$, i.e.

$$\begin{aligned} M &= [(I+z^dA^{-1}BT)T^{-1}]^{-1} \\ &= T(I+z^dA^{-1}BT)^{-1} \\ &= -GC^{-1}A(I-z^dA^{-1}BGC^{-1}A)^{-1} \\ &= -GC^{-1}A[I-A^{-1}(C-AF)C^{-1}A]^{-1} \\ &= -GF^{-1}. \end{aligned}$$

It remains to be seen whether this choice of M is admissible. Clearly it is a matrix of rational functions and so (3i) is satisfied. So we need to only check that (3ii), i.e. nonanticipativity, is satisfied. Now $G(z)$ is analytic at the origin, by assumption, and also $F^{-1}(0)=F_0^{-1}=C_0^{-1}$ exists by assumption, showing that $M(z)$ is analytic at the origin. □

LEMMA 2.6. *Suppose $F(z)$ and $G(z)$ satisfy* (7i, ii, iv). *Then, the control law*

$$u(t)=-G(z)F^{-1}(z)y(t)$$

is stabilizing.

Proof. To determine that the control law is stabilizing, we need to check that the four transfer functions in (4) are all analytic inside the closed unit disc, with M given by $M=-GF^{-1}$. Simple calculation using (7iv) shows that

(11i) $$M[I-z^dA^{-1}BM]^{-1}=-GC^{-1}A=-z^{-d}B^{-1}(C-AF)C^{-1}A,$$

(11ii) $$[I-z^dA^{-1}BM]^{-1}=FC^{-1}A,$$

(11iii) $$[I-z^dMA^{-1}B]^{-1}=B^{-1}AFC^{-1}B=I+[-z^{-d}B^{-1}(C-AF)C^{-1}A][z^dA^{-1}B],$$

(11iv) $$z^dA^{-1}B[I-z^dMA^{-1}B]^{-1}=z^dFC^{-1}B.$$

B^{-1} is analytic inside the closed unit disc, except possibly at the origin, by assumption. $(C-AF)$ is a polynomial by (7i). Also C^{-1} is analytic inside the closed unit disc by (2iii). Hence $z^{-d}B^{-1}(C-AF)C^{-1}A$ is analytic inside the closed unit disc, except perhaps at the origin. However $z^{-d}B^{-1}(C-AF)C^{-1}A=GC^{-1}A$ and since G is analytic at the origin, so is $GC^{-1}A$. Hence (11i) is analytic inside the closed unit disc. (11ii) and (11iv) are both analytic inside the closed unit disc since $C^{-1}(z)$ is so and A, B, F are all matrices of polynomials. Examining (11iii), $B^{-1}AFC^{-1}B$ is analytic inside the closed unit disc, except perhaps at the origin. However $z^dA^{-1}B$ is analytic at the origin, and (11i) has also been shown to be so. Hence $B^{-1}AFC^{-1}B=I+[z^{-d}B^{-1}(C-AF)C^{-1}A][z^dA^{-1}B]$ is also analytic at the origin, thus showing that (11iii) is analytic inside the closed unit disc. □

It may be noted that at this stage Theorem 2.1 has been proved. Now we need to establish Theorems 2.2 and 2.3

LEMMA 2.7. *Let*

(12) $$\tilde{\mathscr{E}}_0^p=\begin{bmatrix} \tilde{E}_p & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & 0 \\ \tilde{E}_0 & \cdots & \cdots & \tilde{E}_p \end{bmatrix}$$

for some matrices $\tilde{E}_0, \cdots, \tilde{E}_p$. If $\tilde{\mathscr{E}}_0^p \mathscr{W}_0^p=\operatorname{diag}(0, \cdots, 0, I)$ then there exists a square

matrix N of the form

$$N=\begin{bmatrix} I & & 0 & 0 \\ \alpha_1 & & & \vdots \\ \vdots & \ddots & & \vdots \\ \vdots & & & 0 \\ \alpha_p & \cdots & \alpha_1 & I \end{bmatrix} \tag{13}$$

such that $\tilde{\mathscr{E}}_0^p = N\mathscr{E}_0^p$.

Proof. Since $\tilde{\mathscr{E}}_0^p \mathscr{W}_0^p = \operatorname{diag}(0, \cdots, 0, I)$ it follows that

$$(\tilde{E}_p z^{-p} + \tilde{E}_{p-1} z^{-p+1} + \cdots + \tilde{E}_0)(D_0 + D_1 z + \cdots) = I + O(z)$$

where $O(z) = \alpha_1 z + \alpha_2 z^2 + \cdots$ for some matrices $\alpha_1, \alpha_2, \cdots$. Hence

$$\begin{aligned}(\tilde{E}_p z^{-p} + \cdots + \tilde{E}_0) &= (I + O(z))(D_0 + D_1 z + \cdots)^{-1} \\ &= (I + O(z))(E_p z^{-p} + E_{p-1} z^{-p+1} + \cdots + E_0 + o(1)).\end{aligned}$$

Equating coefficients of identical powers of z^{-1}, we get

$$\tilde{E}_p = E_p, \quad \tilde{E}_{p-i} = E_{p-i} + \sum_{k=1}^{i} \alpha_k E_{p+k-i} \quad \text{for } i = 1, \cdots, p.$$

Hence the suggested N suffices.

LEMMA 2.8.

$$N(\mathscr{W}_0^{(p-1)T}) = R(\mathscr{E}_1^{pT}).$$

Here $N(\cdot)$ denotes the null space and $R(\cdot)$ the range space.

Proof. Consider $(\beta_1^T, \cdots, \beta_p^T)^T \in N(\mathscr{W}_0^{(p-1)T})$. Suppose to the contrary that $(\beta_1^T, \cdots, \beta_p^T)^T \notin R(\mathscr{E}_1^{pT})$. Since $D_0 \neq 0$ we can find a row vector β_0^T so that $(\beta_0^T, \cdots, \beta_p^T)[D_0^T, \cdots, D_p^T]^T \neq 0$. Since $(\beta_1^T, \cdots, \beta_p^T)^T \notin R(\mathscr{E}_1^{pT})$, it follows that $(\beta_1^T, \cdots, \beta_p^T)^T \notin R([E_1, \cdots, E_p]^T)$. Hence $(\beta_0^T, \cdots, \beta_p^T)^T \notin R([E_0, \cdots, E_p]^T)$. Since $[E_0, \cdots, E_p][D_0^T, \cdots, D_p^T]^T = I$, as is easily checked, it follows that by choosing $(m-1)$ rows from $[E_0, \cdots, E_p]$ (if I above is $m \times m$) and the row $(\beta_0^T, \cdots, \beta_p^T)$ we can build a matrix $[\bar{E}_0, \bar{E}_1, \cdots, \bar{E}_p]$ with m rows and such that $[\bar{E}_0, \cdots, \bar{E}_p][D_0^T, \cdots, D_p^T]^T$ is of full rank. Premultiplying by an appropriate non-singular matrix we can obtain $[\tilde{E}_0, \cdots, \tilde{E}_p]$ such that $[\tilde{E}_0, \cdots, \tilde{E}_p][D_0^T, \cdots, D_p^T]^T = I$ and the rows of $[\tilde{E}_0, \cdots, \tilde{E}_p]$ span the row-space of $[\bar{E}_0, \cdots, \bar{E}_p]$. Now let $\tilde{\mathscr{E}}_0^p$ be defined from $\tilde{E}_0, \cdots, \tilde{E}_p$ as in the statement of Lemma 2.7. It is easily checked that $\tilde{\mathscr{E}}_0^p \mathscr{W}_0^p = \operatorname{diag}(0, \cdots, 0, I)$. Lemma 2.7 now applies and shows that the rows of $[\tilde{E}_0, \cdots, \tilde{E}_p]$ are linear combinations of the rows of $\mathscr{E}_0^p$. But then the rows of $[\tilde{E}_1, \cdots, \tilde{E}_p]$ are linear combinations of the rows of $\mathscr{E}_1^p$, which contradicts our assumption that $(\beta_1^T, \cdots, \beta_p^T)^T \notin R(\mathscr{E}_1^{pT})$. This shows that $N(\mathscr{W}_0^{(p-1)T}) \subseteq R(\mathscr{E}_1^{pT})$. The reverse containment $R(\mathscr{E}_1^{pT}) \subseteq N(\mathscr{W}_0^{(p-1)T})$ follows trivially from the relationship $\mathscr{E}_0^p \mathscr{W}_0^p = \operatorname{diag}(0, \cdots, 0, I)$. □

LEMMA 2.9. $[\mathscr{W}_0^{p-1}, \mathscr{E}_1^{pT}]$ *is a full rank matrix.*

Proof. Suppose $\rho^T[\mathscr{W}_0^{p-1}, \mathscr{E}_1^{pT}]=0$ for some vector ρ. Since $\rho \in N(\mathscr{W}_0^{(p-1)T})$ it follows by Lemma 2.8 that $\rho = \mathscr{E}_1^{pT}\gamma$ for some γ. But $\rho^T\mathscr{E}_1^{pT}\gamma = 0$, and so $\gamma^T\mathscr{E}_1^p\mathscr{E}_1^{pT}\gamma = 0$. Hence $\rho = \mathscr{E}_1^{pT}\gamma = 0$. □

Thus we have also proved Theorem 2.2. Now we complete the proof of Theorem 2.3.

LEMMA 2.10. *If* $F(z)$ *and* $G(z)$ *are defined as in Theorem* 2.3, *then* (7i-iv) *of Theorem* 2.1 *are satisfied.*

Proof. (7i) is trivial since $F(0) = F_0 = C_0$ is invertible by assumption. (7iv) follows from the definition of $G(z)$. So we need to check only (7ii) and (7iii). Now

$$(7\text{iii}) \Leftrightarrow \lim_{z\to 0} z^d[F_0^T + F_1^T z^{-1} + \cdots + F_{d+p-1}^T z^{-d-p+1}][D_0 + D_1 z + \cdots + D_{p-1}z^{p-1}] = 0$$

$$\Leftrightarrow \lim_{z\to 0} [F_d^T + F_{d+1}^T z^{-1} + \cdots + F_{d+p-1}^T z^{-p+1}][D_0 + D_1 z + \cdots + D_{p-1}z^{p-1}] = 0$$

$$\Leftrightarrow \text{coefficients of nonpositive powers of } z \text{ vanish in}$$

$$[F_d^T + \cdots + F_{d+p-1}^T z^{-p+1}][D_0 + D_1 z + \cdots + D_{p-1}z^{p-1}]$$

$$\Leftrightarrow [F_d^T, \cdots, F_{d+p-1}^T]^T \in N(\mathscr{W}_0^{(p-1)T})$$

$$\Leftrightarrow [F_d^T, \cdots, F_{d+p-1}^T] \in R(\mathscr{E}_1^{pT})$$

$$\Leftrightarrow [F_d^T, \cdots, F_{d+p-1}^T] = \mathscr{E}_1^{pT}J \text{ for some matrix } J.$$

Similarly

$$(7\text{iv}) \Leftrightarrow z^{-d}B^{-1}(z)[C(z) - A(z)F(z)] = O(1)$$

$$\Leftrightarrow z^{-d}B^{-1}(z)A(z)[A^{-1}(z)C(z) - F(z)] = O(1)$$

$$\Leftrightarrow z^{-d}B^{-1}(z)A(z)[A^{-1}(z)C(z) - F_0 - F_1 z - \cdots - F_{d+p-1}z^{d+p-1}] = O(1)$$

$$\Leftrightarrow z^{-d}B^{-1}(z)A(z)[H_d z^d + H_{d+1}z^{d+1} + \cdots - F_d z^d - F_{d+1}z^{d+1} \cdots - F_{d+p-1}z^{d+p-1}]$$

$$= O(1),$$

$$\Leftrightarrow B^{-1}(z)A(z)[(H_d - F_d) + \cdots + (H_{d+p-1} - F_{d+p-1})z^{p-1} + O(z^p)] = O(1)$$

$$\Leftrightarrow \text{coefficients of strictly negative powers of } z \text{ vanish in}$$

$$[E_p z^{-p} + \cdots + E_1 z^{-1} + O(1)][(H_d - F_d) + \cdots + (H_{d+p-1} - F_{d+p-1})z^{p-1} + O(z^p)]$$

$$\Leftrightarrow [(H_d - F_d)^T, \cdots, (H_{d+p-1} - F_{d+p-1})^T]^T \in N(\mathscr{E}_1^p)$$

$$\Leftrightarrow [(H_d - F_d)^T, \cdots, (H_{d+p-1} - F_{d+p-1})^T]^T \in R(\mathscr{W}_0^{p-1})$$

$$\Leftrightarrow [(H_d - F_d)^T, \cdots, (H_{d+p-1} - F_{d+p-1})^T]^T \in \mathscr{W}_0^{p-1}K \text{ for some matrix } K.$$

Thus if $[\mathscr{W}_0^{p-1}, \mathscr{E}_1^{pT}][K^T, J^T]^T = [H_d^T, \cdots, H_{d+p-1}^T]^T$ and $\mathscr{E}_1^{pT}J = [F_d^T, \cdots, F_{d+p-1}^T]$, as we have assumed, then both (7ii) and (7iii) are satisfied. □

The proofs of Theorems 2.1, 2.2 and 2.3 are now complete.

3. Square nonminimum phase systems. We now turn to the problem of minimizing the variance over the set of admissible, stabilizing control laws for systems which have nonminimum phase transfer functions besides those caused by pure delays.

Thus we consider systems for which $\det B(z)$ may vanish in $\{z: 0 < |z| < 1\}$ besides possible vanishing in $\{z: z = 0 \text{ or } |z| > 1\}$. We do not allow $\det B(z)$ to vanish in $\{z: |z| = 1\}$ since we have imposed the requirement in (4) that our closed-loop systems

should be *strictly* stable as opposed to just stable, i.e. we have required analyticity of the four transfer functions in (4) in the *closed* unit disc and not just the *open* unit disc. If we are willing to admit such a relaxation, then our solution is valid even for det $B(z)$ vanishing on the unit circle $\{z: |z|=1\}$.

In this section, we also assume that the number of inputs is equal to the number of outputs, i.e. the system is square with $m=l$ in (1).

By Lemma 2.5 we see that we have already solved the problem of obtaining the admissible control law which minimizes the variance of the output, and the control law which does this is just the control law of Theorems 2.1, 2.2 and 2.3. However, this control law is *not* stabilizing, i.e. it does not satisfy (4), when det $B(z)$ vanishes in $\{z: 0<|z|\leqq 1\}$. The reason is that Lemma 2.6 is no more valid, as can be seen from an examination of (11).

For single-input, single-output systems, Peterka [2] has solved the problem of obtaining the control law which minimizes the output variance over the class of all admissible, stabilizing control laws. We now solve this problem for the multivariable case.

We will obtain the solution by reducing the problem to the type considered in the previous section. Accordingly we will denote the $F(z)$ and $G(z)$ generated by Theorem 2.3 by $F(A(\cdot), B(\cdot), C(\cdot), d)(z)$ and $G(A(\cdot), B(\cdot), C(\cdot), d)(z)$ in order to explicity display the functional arguments on which they depend. We note here that the algorithms of Theorems 2.2 and 2.3 can be employed even when $d=0$ to generate F and G.

THEOREM 3.1. *We assume that $A^{-1}(z)$ and $B^{-1}(z)$ have no poles in common inside the closed unit disc, $A^{-1}(z)$ and $B^{-1}(z^{-1})$ have no poles in common inside the closed unit disc and $A^{-1}(z)$ and $A^{-1}(z^{-1})$ have no poles in common. In the above and what follows, by a zero of $X(z)$ we shall mean a singularity of $X^{-1}(z)$, and by a pole of $X(z)$ we mean a singularity of $X(z)$.*

(14i) *Let $\Delta(z)$ be a spectral factor which satisfies*

$$\Delta^T(z^{-1})\Delta(z) = B^T(z^{-1})A^{-T}(z^{-1})A^{-1}(z)B(z)$$

and is such that its poles are those of $A^{-1}(z)B(z)$, while its nonzero zeros are outside the closed unit disc images of the nonzero zeros of $A^{-1}(z)B(z)$. By an "outside the closed unit disc image of z", we mean η such that $\eta=z$ if $|z|>1$ and $\eta=z^{-1}$ if $|z|<1$.

(14ii) *Let $\alpha(z)$ and $\beta(z)$ be matrices of polynomials such that*

$$\alpha^{-1}(z)\beta(z) = \Delta(z)$$

is a left coprime representation of $\Delta(z)$, and such that the zeros of $\beta(z)$ are the zeros of $\Delta(z)$, while the poles of $\alpha^{-1}(z)$ are the poles of $\Delta(z)$.

(14iii) *Let $\theta(z) := \alpha^{-1}(z)\beta(z)B^{-1}(z)[C(z)-A(z)F(z)]z^{-d}$, where*

$$F(z) := F(A(\cdot), B(\cdot), C(\cdot), d)(z) \text{ and}$$

$$G(z) := G(A(\cdot), B(\cdot), C(\cdot), d)(z).$$

(14iv) *Let $\theta_+(z)$ and $\theta_-(z)$ satisfying $\theta(z)=\theta_+(z)+\theta_-(z)$ be such that $\theta_+(z)$ is the sum of all the partial fraction terms of $\theta(z)$ which have poles either outside the closed unit disc (including infinity) or coinciding with the poles of $A^{-1}(z)$ inside the closed unit disc, and constant terms, if any.*

(14v) *Let* $\gamma(z)$ *be a polynomial matrix such that*

$$\theta_+(z) = \alpha^{-1}(z)\gamma(z).$$

(*The existence of such a polynomial matrix* $\gamma(z)$ *will be proved.*)

(14vi) *Let* $\tilde{F}(z) := F(\alpha(\cdot), \beta(\cdot), \gamma(\cdot), 0)(z)$ *and*

$$\tilde{G}(z) := G(\alpha(\cdot), \beta(\cdot), \gamma(\cdot), 0)(z).$$

Then, the control law which minimizes $Ey^T(t)y(t)$ *over the class of all admissible, stabilizing control laws is given by*

$$u(t) = -\tilde{G}(z)[F(z) + z^d A^{-1}(z)B(z)(G(z) - \tilde{G}(z))]^{-1} y(t). \tag{15}$$

The resulting minimum variance is

$$\begin{aligned} Ey^T(t)y(t) &= \operatorname{tr} \sum_j F_j^T F_j Q + \operatorname{tr} \sum_j \tilde{F}_j^T \tilde{F}_j Q \\ &+ \frac{\operatorname{tr}}{2\pi i} \oint \{\alpha^{-1}(z)\beta(z)[G(z) - \tilde{G}(z)] - \tilde{F}(z)\} \\ &\quad \cdot \{\alpha^{-1}(z^{-1})\beta(z^{-1})[G(z^{-1}) - \tilde{G}(z^{-1})] - \tilde{F}(z^{-1})\}^T Q \frac{dz}{z} \end{aligned} \tag{16}$$

where

$$F(z) =: \sum_j F_j z^j \quad \text{and} \quad \tilde{F}(z) =: \sum_j \tilde{F}_j z^j.$$

Proof. Let $u(t) = M(z)y(t)$. From (10), it follows that

$$\begin{aligned} Ey^T(t)y(t) &= \operatorname{tr} \sum_j F_j^T F_j Q \\ &+ \frac{\operatorname{tr}}{2\pi i} \oint B^T(z^{-1})A^{-T}(z^{-1})A^{-1}(z)B(z) \\ &\quad \cdot [G(z) + T(z)A^{-1}(z)C(z)] \\ &\quad \cdot Q[G(z^{-1}) + T(z^{-1})A^{-1}(z^{-1})C(z^{-1})]^T \frac{dz}{z}. \end{aligned}$$

Since $\beta^T(z^{-1})\alpha^{-T}(z^{-1})\alpha^{-1}(z)\beta(z) = B^T(z^{-1})A^{-T}(z^{-1})A^{-1}(z)B(z)$ from (14i, ii), it follows that

$$\begin{aligned} Ey^T(t)y(t) &= \operatorname{tr} \sum_j F_j^T F_j Q + \frac{\operatorname{tr}}{2\pi i} \oint \beta^T(z^{-1})\alpha^{-T}(z^{-1}) \\ &\quad \cdot [\alpha^{-1}(z)\beta(z)G(z) + \alpha^{-1}(z)\beta(z)T(z)A^{-1}(z)C(z)] \\ &\quad \cdot Q[G(z^{-1}) + T(z^{-1})A^{-1}(z^{-1})C(z^{-1})]^T \frac{dz}{z}. \end{aligned}$$

Substituting $G(z)=B^{-1}(z)[C(z)-A(z)F(z)]z^{-d}$, and using (14iii), (14iv) gives

$$Ey^T(t)y(t)=\operatorname{tr}\sum_j F_j^T F_j Q + \frac{\operatorname{tr}}{2\pi i}\oint [\theta_+(z)+\theta_-(z)+\alpha^{-1}(z)\beta(z)T(z)A^{-1}(z)C(z)] \cdot Q[\theta_+(z^{-1})+\theta_-(z^{-1})+\alpha^{-1}(z^{-1})\beta(z^{-1}) \cdot T(z^{-1})A^{-1}(z^{-1})C(z^{-1})]^T \frac{dz}{z}.$$

Since $y_2=z^dA^{-1}B(G+TA^{-1}C)w$, in the notation of Lemma 2.4, our assumption that the control law is stabilizing shows that $A^{-1}B(G+TA^{-1}C)$ has no singularities inside the closed unit disc, and in particular that it has no singularities inside the closed unit disc coinciding with those of $A^{-1}(z)$. Hence the residues of $[\theta_+(z)+\alpha^{-1}(z)\beta(z)T(z)A^{-1}(z)C(z)]$ evaulated at the zeros of $A(z)$ or $\alpha(z)$ which are inside the closed unit disc are zero. Since therefore $[\theta_+(z)+\alpha^{-1}(z)\beta(z)T(z)A^{-1}(z)C(z)]$ and $(\theta_-^T(z^{-1}))/z$ are both analytic inside the closed unit disc, where we have also used the fact that $\theta_-(z)$ is a matrix of strictly proper rational functions, we deduce that the cross term

$$\frac{\operatorname{tr}}{2\pi i}\oint [\theta_+(z)+\alpha^{-1}(z)\beta(z)T(z)A^{-1}(z)C(z)]Q\theta_-^T(z^{-1})\frac{dz}{z}$$

vanishes. Hence

$$Ey^T(t)y(t)=\operatorname{tr}\sum_j F_j^T F_j Q+\frac{\operatorname{tr}}{2\pi i}\oint \theta_-(z)Q\theta_-^T(z^{-1})\frac{dz}{z} + \frac{\operatorname{tr}}{2\pi i}\oint [\theta_+(z)+\alpha^{-1}(z)\beta(z)T(z)A^{-1}(z)C(z)] \cdot Q[\theta_+(z^{-1})+\alpha^{-1}(z^{-1})\beta(z^{-1}) \cdot T(z^{-1})A^{-1}(z^{-1})C(z^{-1})]^T \frac{dz}{z}. \tag{17}$$

The first two terms in the right-hand side of (17) do not depend on the choice of $T(z)$ and, therefore, on the choice of $M(z)$. Hence to minimize $Ey^T(t)y(t)$, we need to only minimize

$$\frac{\operatorname{tr}}{2\pi i}\oint [\theta_+(z)+\alpha^{-1}(z)\beta(z)T(z)A^{-1}(z)C(z)] \cdot Q[\theta_+(z^{-1})+\alpha^{-1}(z^{-1})\beta(z^{-1})T(z^{-1})A^{-1}(z^{-1})C(z^{-1})]^T \frac{dz}{z}. \tag{18}$$

Now let us examine $\theta_+(z)$. From (14i, ii) we see that

$$\alpha^{-1}(z)\beta(z)B^{-1}(z)=\alpha^T(z^{-1})\beta^{-T}(z^{-1})B^T(z^{-1})A^{-T}(z^{-1})A^{-1}(z). \tag{19}$$

An examination of the right-hand side of (19) shows that the only poles of (19) which do not coincide with those of $A^{-1}(z)$ are either at the origin or coincide with the poles of $\beta^{-T}(z^{-1})$, and so all the poles of the left-hand side of (19) which do not coincide with the poles of $A^{-1}(z)$ are inside the closed unit disc. Substituting (19) in the

expression for $\theta(z)$ in (14iii), we obtain

$$\theta(z)=\alpha^T(z^{-1})\beta^{-T}(z^{-1})B^T(z^{-1})A^{-T}(z^{-1})A^{-1}(z)[C(z)-A(z)F(z)]z^{-d}.$$

Utilizing the definition of $\theta_+(z)$, we see therefore that

$$\theta_+(z)=\alpha^{-1}(z)\gamma(z)$$

for some polynomial matrix $\gamma(z)$. This proves the existence of $\gamma(z)$ claimed in (14v). Now substituting for $\theta_+(z)$ in (18) shows that to minimize $Ey^T(t)y(t)$, we need to minimize

$$\frac{\operatorname{tr}}{2\pi i}\oint[\alpha^{-1}(z)\gamma(z)+\alpha^{-1}(z)\beta(z)T(z)A^{-1}(z)C(z)]$$

$$\cdot\, Q[\alpha^{-1}(z^{-1})\gamma(z^{-1})+\alpha^{-1}(z^{-1})\beta(z^{-1})T(z^{-1})A^{-1}(z^{-1})C(z^{-1})]^T\frac{dz}{z}.$$

Define $S(z):=T(z)A^{-1}(z)C(z)$ and our problem now is how to choose $S(z)$, analytic at the origin, so as to minimize

$$\frac{\operatorname{tr}}{2\pi i}\oint[\alpha^{-1}(z)\gamma(z)+\alpha^{-1}(z)\beta(z)S(z)]$$

$$\cdot\, Q[\alpha^{-1}(z^{-1})\gamma(z^{-1})+\alpha^{-1}(z^{-1})\beta(z^{-1})S(z^{-1})]^T\frac{dz}{z}. \tag{20}$$

But this resembles the problem of § 2, where since we had

$$\begin{aligned}y(t)&=[A^{-1}(z)C(z)+z^dA^{-1}(z)B(z)T(z)A^{-1}(z)C(z)]w(t)\\&=[A^{-1}(z)C(z)+z^dA^{-1}(z)B(z)S(z)]w(t)\end{aligned}$$

we had to choose $S(z)$, analytic at the origin, so as to minimize

$$\frac{\operatorname{tr}}{2\pi i}\oint[A^{-1}(z)C(z)+z^dA^{-1}(z)B(z)S(z)]$$

$$\cdot\, Q[A^{-1}(z^{-1})C(z^{-1})+z^{-d}A^{-1}(z^{-1})B(z^{-1})S(z^{-1})]^T\frac{dz}{z}. \tag{21}$$

Making the obvious identifications between (20) and (21), we can apply the results of § 2 and see that the optimal choice for $S(z)$ is

$$S(z)=-\tilde{G}(z) \tag{22}$$

where $\tilde{G}(z)$ is as in (14vi). Furthermore the minimum value of (20) is

$$\sum_j \tilde{F}_j^T\tilde{F}_jQ. \tag{23}$$

Since (23) is the minimum value of the third term in the right-hand side of (17), it follows by substituting in (17) that the resulting variance is

$$Ey^T(t)y(t)=\operatorname{tr}\sum_j F_j^TF_jQ+\operatorname{tr}\sum_j\tilde{F}_j^T\tilde{F}_jQ+\frac{\operatorname{tr}}{2\pi i}\oint\theta_-(z)Q\theta_-^T(z^{-1})\frac{dz}{z}. \tag{24}$$

Since $C(z)=A(z)F(z)+z^dB(z)G(z)$ and $\gamma(z)=\alpha(z)\tilde{F}(z)+\beta(z)\tilde{G}(z)$, which follow from the definitions of F, G, $\tilde{F}$ and $\tilde{G}$ in (14iii, vi) we obtain $G(z)=B^{-1}(z)[C(z)-A(z)F(z)]z^{-d}$ and $\alpha^{-1}(z)\gamma(z)=\tilde{F}(z)+\alpha^{-1}(z)\beta(z)\tilde{G}(z)$. Substituting

these two expressions in the definitions of $\theta(z)$ and $\theta_+(z)$ in (14iii, v), we get

$$\begin{aligned}\theta_-(z) &= \theta(z) - \theta_+(z) \\ &= \alpha^{-1}(z)\beta(z)G(z) - \tilde{F}(z) - \alpha^{-1}(z)\beta(z)\tilde{G}(z) \\ &= \alpha^{-1}(z)\beta(z)[G(z) - \tilde{G}(z)] - \tilde{F}(z).\end{aligned} \tag{25}$$

Substituting (25) in (24) gives the expression (16) claimed as the minimum variance. We still need to determine that the choice of $S(z)$ in (22) corresponds to (15) and also that it is stabilizing. Since $S(z) = T(z)A^{-1}(z)C(z)$, we obtain that the choice of $T(z)$ is $T(z) = -\tilde{G}(z)C^{-1}(z)A(z)$ and since $T = M[I - z^d A^{-1}BM]^{-1}$ it follows that

$$\begin{aligned}M(z) &= T(z)[I + z^d A^{-1}(z)B(z)T(z)]^{-1} \\ &= T(z)[I - z^d A^{-1}(z)B(z)\tilde{G}(z)C^{-1}(z)A(z)]^{-1} \\ &= -\tilde{G}(z)[A^{-1}(z)C(z) - z^d A^{-1}(z)B(z)\tilde{G}(z)]^{-1} \\ &= -\tilde{G}(z)[F(z) + z^d A^{-1}(z)B(z)(G(z) - \tilde{G}(z))]^{-1}\end{aligned}$$

which coincides with the control law of (15). It remains to be shown that this choice of $M(z)$ is stabilizing, i.e. it satisfies (4). Simple calculations show that two of the transfer functions in (4) are

$$\begin{aligned}M[I - z^d A^{-1}BM]^{-1} &= -\beta^{-1}(\gamma - \alpha\tilde{F})C^{-1}A, \\ [I - z^d MA^{-1}B]^{-1} &= I - z^d\tilde{G}C^{-1}A = I - z^d\beta^{-1}(\gamma - \alpha\tilde{F})C^{-1}B\end{aligned} \tag{26}$$

which are both analytic inside the closed unit disc, since β^{-1} and C^{-1} are. The third transfer function in (4) is $[I - z^d A^{-1}(z)B(z)M(z)]^{-1}$ which can, by simple calculation, be seen to be equal to $I - z^d A^{-1}(z)B(z)\tilde{G}(z)C^{-1}(z)A(z)$, which in turn is $I - z^d A^{-1}(z)B(z)\beta^{-1}(z)[\gamma(z) - \alpha(z)\tilde{F}(z)]C^{-1}(z)A(z)$. Except for the term $A^{-1}(z)$, all other quantities are analytic inside the closed unit disc, and so if this transfer function has any singularities inside the closed unit disc, they must coincide with those of $A^{-1}(z)$. However, we also have

$$\begin{aligned}[I - z^d A^{-1}(z)B(z)M(z)]^{-1} &= [F(z) + z^d A^{-1}(z)B(z)(G(z) - \tilde{G}(z))]C^{-1}(z)A(z) \\ &= \{F(z) + z^d A^{-1}(z)B(z)\beta^{-1}(z)\alpha(z)[\theta_-(z) - \tilde{F}(z)]\} \\ &\quad \cdot C^{-1}(z)A(z)\end{aligned}$$

which, if it has any singularities inside the closed unit disc coinciding with those of $A^{-1}(z)$, can only be singularities of $A^{-1}(z)B(z)\beta^{-1}(z)\alpha(z)$ inside the closed unit disc coinciding with those of $A^{-1}(z)$. However, by (14i, ii), we have $A^{-1}(z)B(z)\beta^{-1}(z)\alpha(z) = A^T(z^{-1})B^{-T}(z^{-1})\beta^T(z^{-1})\alpha^{-T}(z^{-1})$. The only poles of the right-hand side inside the closed unit disc are either at the origin or coincident with the poles of $\alpha^{-1}(z^{-1})$ or $B^{-1}(z^{-1})$. By our assumptions, there can however be no poles of $A^{-1}(z)$ in any of these locations, showing that $[I - z^d A^{-1}BM]^{-1}$ is analytic inside the closed unit disc. The last transfer function of (14) we need to check is $z^d A^{-1}B[I - z^d MA^{-1}B]^{-1}$. Since (26), which is a factor, has no poles inside the closed unit disc, it follows that if there are poles of $z^d A^{-1}B[I - z^d MA^{-1}B]^{-1}$ inside the closed unit disc, they must be poles of A^{-1}. Simple calculation shows that $z^d A^{-1}B[I - z^d MA^{-1}B]^{-1} = z^d FC^{-1}B + z^d A^{-1}B\beta^{-1}\alpha(\theta_- + \tilde{F})C^{-1}B$. The first term is analytic inside the closed unit disc, and so is $(\theta_- + \tilde{F})C^{-1}B$. Hence we only need to show that $A^{-1}(z)B(z)\beta^{-1}(z)\alpha(z)$ has no poles inside the closed unit disc which coincide with those of $A^{-1}(z)$. But we have already done this. □

If the system is of minimum phase, i.e. $B^{-1}(z)$ is analytic in $\{z: 0<|z|\leqq 1\}$, then $\alpha = A$, $\beta = B$ and so $\theta = A^{-1}[C - AF]z^{-d}$, thus showing that $\gamma = [C - AF]z^{-d}$. Hence $\tilde{F} = 0$ and $\tilde{G} = G$. Thus the control law (15) above reduces to what it is in the minimum phase case of Theorem 2.1. Moreover the minimum variance (16) also reduces to what it is in Theorem 2.1.

The additional cost of stably controlling a nonminimum phase system is therefore

$$\operatorname{tr} \sum_j \tilde{F}_j^T \tilde{F}_j Q + \frac{\operatorname{tr}}{2\pi i} \oint \{\alpha^{-1}(z)\beta(z)[G(z) - \tilde{G}(z)] - \tilde{F}(z)\} \cdot Q\{\alpha^{-1}(z^{-1})\beta(z^{-1})[G(z^{-1}) - \tilde{G}(z^{-1})] - \tilde{F}(z^{-1})\}^T \frac{dz}{z}.$$

This is the "sacrifice" in variance that must be made to obtain a stable system. If one just wants to minimize the variance without paying attention to stability, then this sacrifice need not be made.

One useful property of the control law (15) is that it does not depend on the noise covariance $Ew(t)w^T(t)$. Thus, the same control law is optimal irrespective of the covariance $Ew(t)w^T(t)$.

4. Rectangular systems. Now we consider rectangular systems, i.e. systems where the number of inputs is not equal to the number of outputs.

If the system has more inputs than outputs, then the previous results can still be used if we replace $B^{-1}(z)$ by $B^{\#}(z)$, any right inverse of $B(z)$. The proofs proceed as before.

So we turn our attention to systems where the number of inputs is less than the number of outputs. Before describing the solution, we first discuss some pitfalls. One way of proceeding, it might appear, is to make the system "square" by adding fictitious inputs with small "gains" ε which are then driven to zero. This can, however, result in matrices $M(z)$ and $T(z)$ which become unbounded as $\varepsilon \to 0$. Another way of making the system square is to add fictitious inputs which have *delays* which are then driven to infinity. However, the resulting solution for $F(z)$ will be a power series, at best.

We therefore adopt the more fruitful approach of the following Theorem. As in previous sections, we assume that the system has no zeros exactly on the unit circle $\{z: |z| = 1\}$, or more precisely, $B^T(z^{-1})A^{-T}(z^{-1})A^{-1}(z)B(z)$ has no zeros on the unit circle $\{z: |z| = 1\}$.

THEOREM 4.1. *We assume that $A^{-1}(z)$ and $A^{-1}(z^{-1})$ have no poles in common and also that for every pole t_k of $A^{-1}(z)$ inside the closed unit disc, its residue R_k in the partial fraction expansion of $A^{-1}(z)B(z)$ satisfies the condition* $\lim_{z \to t_k} B^T(z^{-1})A^{-T}(z^{-1})R_k \neq 0$.

(27i) *Let $\delta(z) := A^{-1}(z)B(z)$.*

(27ii) *Let $\Delta(z) = \alpha^{-1}(z)\beta(z)$ be a square minimum phase spectral factor satisfying $\Delta^T(z^{-1})\Delta(z) = \delta^T(z^{-1})\delta(z)$ and such that the nonzero zeros of the polynomial matrix $\beta(z)$ are outside the unit circle images of the nonzero zeros of $\delta^T(z^{-1})\delta(z)$ while the poles of the polynomial matrix $\alpha^{-1}(z)$ are those of $\delta(z)$.*

(27iii) *Define $\tilde{\theta}(z) := \Delta^{-T}(z^{-1})\delta^T(z^{-1})A^{-1}(z)C(z)$ and decompose $\tilde{\theta}(z)$ as $\tilde{\theta}(z) =: \tilde{\theta}_+(z) + \tilde{\theta}_-(z)$ where $\tilde{\theta}_-(z)$ consists of those partial fraction terms with poles which are inside the unit circle and not coinciding with those of $A^{-1}(z)$.*

(27iv) *Let $\tilde{\theta}_+(z) = \tilde{\alpha}^{-1}(z)\tilde{\gamma}(z)$ where $\tilde{\alpha}(z)$ is a square polynomial matrix with zeros corresponding to those of $A(z)$ and $\tilde{\gamma}(z)$ is a rectangular matrix of polynomials with more columns than rows.*

(27v) *Let* $\tilde{F}(z) =: F(\tilde{\alpha}(\cdot), \beta(\cdot), \tilde{\gamma}(\cdot), d)(z)$ *and* $\tilde{G}(z) := G(\tilde{\alpha}(\cdot), \beta(\cdot), \tilde{\gamma}(\cdot), d)(z)$. *Then, the control law which minimizes the output variance* $Ey^T(t)y(t)$ *over the set of all admissible stabilizing control laws is*

$$u(t) = -\tilde{G}(z)[C(z) - z^d B(z)\tilde{G}(z)]^{-1}A(z)y(t).$$

Proof. Let $\tilde{\delta}(z)$ be a full rank left annihilator of $\delta(z)$. Clearly

$$\begin{bmatrix} \tilde{\delta}(z) \\ [\delta^T(z^{-1})\delta(z)]^{-1}\delta^T(z^{-1}) \end{bmatrix} [\tilde{\delta}^T(z^{-1})[\tilde{\delta}(z)\tilde{\delta}^T(z^{-1})]^{-1}, \delta(z)] = I$$

and so each of the matrices on the left-hand side of the above is the inverse of the other. Multiplying the two matrices above in the reverse order gives

$$\tilde{\delta}^T(z^{-1})[\tilde{\delta}(z)\tilde{\delta}^T(z^{-1})]^{-1}\tilde{\delta}(z) + \delta(z)[\delta^T(z^{-1})\delta(z)]^{-1}\delta^T(z^{-1}) = I.$$

Hence, for any admissible $u(t) = M(z)y(t)$, we can decompose $y(t) = y_1(t) + y_2(t)$ where

$$\begin{aligned} y_1(t) &= \tilde{\delta}^T(z^{-1})[\tilde{\delta}(z)\tilde{\delta}^T(z^{-1})]^{-1}\tilde{\delta}(z)A^{-1}(z)C(z)w(t), \quad \text{and} \\ y_2(t) &= \delta(z)\{[\delta^T(z^{-1})\delta(z)]^{-1}\delta^T(z^{-1})A^{-1}(z)C(z) + z^d T(z)A^{-1}(z)C(z)\}w(t) \end{aligned} \tag{28}$$

where $T(z)$ is as in Lemma 2.4. By integrating along the unit circle, it can be seen that $Ey_1^T(t)y_2(t) = 0$. So $Ey^T(t)y(t) = Ey_1^T(t)y_1(t) + Ey_2^T(t)y_2(t)$. The first term on the right-hand side does not depend on the choice of $M(z)$. Hence, to minimize $Ey^T(t)y(t)$ we need to minimize only $Ey_2^T(t)y_2(t)$. Now

$$\begin{aligned} Ey_2^T(t)y_2(t) = \frac{\text{tr}}{2\pi i}\oint \delta^T(z^{-1})\delta(z)\{[\delta^T(z^{-1})\delta(z)]^{-1}\delta^T(z^{-1}) + z^d T(z)\} \\ \cdot A^{-1}(z)C(z)QC^T(z^{-1})A^{-T}(z^{-1}) \\ \cdot \{[\delta^T(z)\delta(z^{-1})]^{-1}\delta^T(z) + z^{-d}T(z^{-1})\}^T \frac{dz}{z} \end{aligned}$$

and so, using $\Delta^T(z^{-1})\Delta(z) = \delta^T(z^{-1})\delta(z)$, we get

$$\begin{aligned} Ey_2^T(t)y_2(t) &= \frac{\text{tr}}{2\pi i}\oint [\Delta^{-T}(z^{-1})\delta^T(z^{-1}) + z^d\Delta(z)T(z)] \\ &\quad \cdot A^{-1}(z)C(z)QC^T(z^{-1})A^{-T}(z^{-1}) \\ &\quad \cdot [\Delta^{-T}(z)\delta^T(z) + z^{-d}\Delta(z^{-1})T(z^{-1})]^T \frac{dz}{z} \\ &= \frac{\text{tr}}{2\pi i}\oint [\tilde{\theta}_-(z) + \tilde{\theta}_+(z) + z^d\alpha^{-1}(z)\beta(z)T(z)A^{-1}(z)C(z)] \\ &\quad \cdot Q[\tilde{\theta}_-(z^{-1}) + \tilde{\theta}_+(z^{-1}) + z^{-d}\alpha^{-1}(z^{-1})\beta(z^{-1}) \\ &\quad \cdot T(z^{-1})A^{-1}(z^{-1})C(z^{-1})]^T \frac{dz}{z}. \end{aligned}$$

As in Theorem 3.1, the cross term

$$\frac{\text{tr}}{2\pi i}\oint [\tilde{\theta}_+(z) + z^d\alpha^{-1}(z)\beta(z)T(z)A^{-1}(z)C(z)]Q\tilde{\theta}_-^T(z^{-1})\frac{dz}{z}$$

vanishes. Also, the term $(\text{tr}/2\pi i)\oint \tilde{\theta}_-(z)Q\tilde{\theta}_-^T(z^{-1})dz/z$ can be ignored since it does

not depend on $M(z)$. Hence, the problem becomes one of minimizing

$$\begin{aligned}(29)\quad &\frac{\mathrm{tr}}{2\pi i}\oint [\tilde{\theta}_+(z)+z^d\alpha^{-1}(z)\beta(z)T(z)A^{-1}(z)C(z)]\\ &\qquad\cdot Q[\tilde{\theta}_+(z^{-1})+z^{-d}\alpha^{-1}(z^{-1})\beta(z^{-1})T(z^{-1})A^{-1}(z^{-1})C(z^{-1})]^T\frac{dz}{z}\\ &=\frac{\mathrm{tr}}{2\pi i}\oint [\tilde{\alpha}^{-1}(z)\tilde{\gamma}(z)+z^d\alpha^{-1}(z)\beta(z)S(z)]\\ &\qquad\cdot Q[\tilde{\alpha}^{-1}(z^{-1})\tilde{\gamma}(z^{-1})+z^{-d}\alpha^{-1}(z^{-1})\beta(z^{-1})S(z^{-1})]^T\frac{dz}{z}.\end{aligned}$$

The slight difference, because $\alpha \neq \tilde{\alpha}$, between the problem of minimizing (29) and the problem of minimizing (20) is unimportant, and the rest of the proof proceeds as in the proof of Theorem 3.1. □

The variance of (28) represents an additional cost due to the nonsquareness of the system. The cost term $(\mathrm{tr}/2\pi i)\oint \tilde{\theta}_-(z)Q\tilde{\theta}_-^T(z^{-1})(dz/z)$ is also affected by the nonsquareness and will be positive even if the system is of minimum phase.

Acknowledgments. The second author would like to thank G. Verghese for several useful discussions. He would also like to thank S. Mitter for his friendly hospitality during his visit to MIT.

REFERENCES

[1] K. J. ÅSTRÖM, *Introduction to Stochastic Control Theory*, Academic Press, New York, 1970.
[2] V. PETERKA, *On steady state minimum variance control strategy*, Kybernetika, 8 (1972), pp. 218-231.
[3] V. BORISON, *Self-tuning regulators for a class of multivariable systems*, Automatica, 15 (1979), pp. 209-215.
[4] G. GOODWIN, P. RAMADGE AND P. CAINES, *Discrete time stochastic adaptive control*, this Journal, 19 (1981), pp. 829-853.
[5] M. M. BAYOUMI AND M. A. EL BAGOURY, *Comments on self-tuning adaptive control of cement raw material blending*, Automatica, 15 (1979), pp. 693-694.
[6] L. DUGARD, G. C. GOODWIN AND X. XIANYA, *The role of the interactor matrix in multivariable stochastic adaptive control*, Automatica, 20 (1984), pp. 701-709.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 3, May 1986

006

EXISTENCE THEOREMS FOR OPTIMAL CONTROL AND CALCULUS OF VARIATIONS PROBLEMS WHERE THE STATES CAN JUMP*

J. M. MURRAY†

Abstract. We examine optimal control and calculus of variations problems where the states can be discontinuous. In the control theory setting this is caused by allowing impulse controls. Conditions under which optimal solutions will exist for these problems are determined after we derive suitable objective functionals that can handle the jump terms.

Key words. control theory, calculus of variations, existence, impulse controls, jumps

1. Introduction. In this paper the question of existence of an optimal solution for optimal control and calculus of variations problems is investigated. For the usual class of problems where the state vector x must be absolutely continuous, current existence theory requires that the Hamiltonian $H(t, x, p)$ nowhere has the value of $+\infty$. This is equivalent to $P(t, x)$, the effective domain of $H(t, x, \cdot)$, being the whole space $\mathbb{R}^n$. In many economic applications this is undesirable since the elements of $P(t, x)$ have the interpretation of price vectors, which may be subject to constraints; the case $P(t, x) \equiv \mathbb{R}^n_+$, the nonnegative orthant, is often encountered.

Dealing with these problems requires replacing the original space of absolutely continuous functions with the space of functions of bounded variation for the state vectors x. Jumps in the state can occur when the costate vector p strikes the boundary of $P(t, x)$. In the control theory setting these jumps arise from using impulse controls. To incorporate the cost of any jumps in the state and the use of impulses, the objective functional must be extended, and in such a way so that it reduces to the usual case when jumps and impulses are not present. Rockafellar [13] has derived such an extension and obtained existence results for his problem but with the assumption that the Lagrangian $L(t, x, v)$ is convex in (x, v) for each t. This implies that $P(t, x)$ is independent of x.

In this paper, calculus of variations and optimal control problems are extended to allow jumps in the state and impulse controls and existence theorems are derived for these problems. These results differ from Rockafellar's in that the Lagrangian $L(t, x, v)$ is merely convex in v for each (t, x) as is the case for existence theorems for problems where the state must be absolutely continuous.

Other results in this area can be found in [2], [6], [7], [14], [15], and [16].

2. Notation and definitions. We will only consider calculus of variations and optimal control problems over a fixed time interval $[t_0, t_1]$. To simplify notation we will denote this interval by T.

Let $\mathcal{A}_1(D)$ be the space of absolutely continuous functions over the interval D. By $\mathcal{A}_m(D)$ we will mean the space of vector-valued functions $f = (f_1, \cdots, f_m)$ where $f_i \in \mathcal{A}_1(D)$, $i = 1, \cdots, m$. Since we mostly deal with the interval T and the case where n is the dimension of the state space, we will often shorten $\mathcal{A}_n(T)$ to $\mathcal{A}$, with the time interval and the number of components understood to be the above.

* Received by the editors March 6, 1984, and in revised form November 5, 1984.

† Department of Applied Mathematics, University of New South Wales, Kensington, 2033, New South Wales, Australia. This paper formed part of the author's doctoral dissertation completed at the University of Washington, Seattle, Washington, under the supervision of Professor R. T. Rockafellar. That work was partially supported by the Air Force Office of Scientific Research, under grant F49620-82-K-0012.

Let $\mathscr{C}_1(D)$ be the space of continuous functions on the interval D. We will also use $\mathscr{C}_m(D)$ and $\mathscr{C}$.

For $D=[s_0, s_1]$ let $\mathscr{NB}_1(D)$ be the space of functions of bounded variation that are right continuous on $(s_0, s_1]$.

Two functions of bounded variation, f and g, will be called *equivalent* if $f=g$ except on a countable subset of (s_0, s_1). Let $\mathscr{B}_1(D)$ be the space of equivalence classes of functions of bounded variation on D. Belonging to each equivalence class is a unique element of $\mathscr{NB}_1(D)$, and corresponding to this function is a unique Borel measure. This measure then determines the particular equivalence class of $\mathscr{B}_1(D)$. As above, $\mathscr{B}_m(D)$ will denote the space of vector valued functions $f=(f_1, \cdots, f_m)$ where each f_i belongs to $\mathscr{B}_1(D)$ and for ease of notation we will write $\mathscr{B}_n(T)$ as $\mathscr{B}$.

Let $\mathscr{M}(D)$ be the space of one-dimensional, nonnegative regular Borel measures on D. Where D is obvious from the context we will simply write $\mathscr{M}$.

Let $g \in \mathscr{B}_m(D)$, where $D=[s_0, s_1]$. Then $dg(t)=\dot{g}(t)\,dt+\gamma(t)\,d\theta(t)$ for some $\theta \in \mathscr{M}$, and Borel measurable function γ, and we can define

$$\|g\|_{m,V} \triangleq |g(s_0)|+\int_D |\dot{g}(t)|\,dt+\int_D |\gamma(t)|\,d\theta(t)$$

whose value is independent of the choice of θ and γ.

In what follows, B_d^m denotes the m-dimensional closed ball with radius d and centred at the origin. As above, we will often omit the superscript m when it equals n the dimension of the state space, and simply write B_d.

Let $L_1(D)$ be the class of Lebesgue measurable functions on D. By $L_m(D)$ we will mean the space of vector-valued functions $f=(f_1, \cdots, f_m)$ where $f_i \in L_1(D)$, $i=1, \cdots, m$. Where the interval involved is obvious we will often write L_m instead of $L_m(D)$.

The *effective domain* of the function f is the set $\operatorname{dom} f \equiv \{y: f(y)<\infty\}$.

When the minimum of a function g may not be attained we write $\inf_y g(y)$ in place of $\min_y g(y)$, and, for the latter statement, there is the implicit understanding that the minimum is attained.

The generalized problem of Bolza in calculus of variations has the following form

$$\underset{x \in \mathscr{A}}{\text{minimize}}\ J(x)=l(x(t_0), x(t_1))+\int_T L(t, x(t), \dot{x}(t))\,dt \tag{2.1}$$

where l and L have values in $\mathbb{R} \cup \{+\infty\}$.

An example where $J(x)$ has a finite infimum over all $x \in \mathscr{A}$ but no such x attains the infimum, is the following:

Example 1. minimize $J(x)=l(x(0), x(1))+\int_0^1 (1-t)|\dot{x}(t)|\,dt$ where

$$l(a, b)=\begin{cases} 0 & \text{if } a=0,\ b=\pi, \\ +\infty & \text{otherwise.} \end{cases}$$

This has a minimizing sequence $\{\bar{x}_n\}$ where

$$\bar{x}_n(t)=\begin{cases} 0, & i \in \left[0, 1-\dfrac{1}{n}\right], \\ n\left(t-1+\dfrac{1}{n}\right)\pi, & t \in \left(1-\dfrac{1}{n}, 1\right], \end{cases}$$

but no minimum in the space $\mathscr{A}$ since the limit function is discontinuous at $t=1$. In

order to include such problems in the general framework, we wish to extend problem (2.1) to states x that have discontinuities, which we can perhaps consider as arising from the use of impulse controls in the control theory setting. The larger class of functions we will use will be $\mathcal{B}$. If we so wanted, we could manage this by using a limiting process of the kind evident in the example. However, it is our purpose here to actually derive a new objective functional for $\mathcal{B}$ which is equivalent to (2.1) when the two overlap. In [13] where the Lagrangian, $L(t, x, v)$, was convex in (x, v) for each t, a suitable extension of (2.1) turned out to be

$$(2.2) \qquad \underset{x\in\mathcal{B}}{\text{minimize}}\ I(x) = l(x(t_0), x(t_1)) + \int_T L(t, x(t), \dot{x}(t))\, dt + \int_T \bar{r}(t, \xi(t))\, d\theta(t)$$

where $dx(t) = \dot{x}(t)\, dt + \xi(t)\, d\theta(t)$, $\theta \in \mathcal{M}$ and $\bar{r}$ is the recession function for L (this is defined later).

The suitability of the extension (2.2) of (2.1) arises from the ability to derive existence results for it and its compatibility with (2.1) when $x \in \mathcal{A}$. Also it agrees with the limiting process mentioned above and one has under certain conditions $\inf_{\mathcal{A}} J = \min_{\mathcal{B}} I$.

It is our task in this paper to obtain existence theorems for extensions of the problem (2.1) when L is merely convex in v for each (t, x). These will then give some idea on how to properly formulate problems when the states can jump.

Our desire to consider the case when L is only convex in v is partly in an attempt to mirror existence theory when the states must belong to $\mathcal{A}$. In some cases it turns out that we can use the same extension (2.2) when the convexity condition is relaxed. However at other times we cannot and must resort to a more complicated formula that reflects the underlying situation. We will separate the two situations and show when the simpler formula may be used and, in that case, the other formula actually reduces to it.

The need for something subtler than (2.2) can be seen from the following.

Example 2. Let $x, v \in \mathbb{R}^2$ and define the matrix function R by

$$R(x) = \begin{pmatrix} \cos|x| & \sin|x| \\ -\sin|x| & \cos|x| \end{pmatrix}.$$

Let

$$L(t, x, v) = \begin{cases} (1-t)|v| & \text{if } (0,1)\cdot R(x)\cdot \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \leqq 0, \\ +\infty & \text{otherwise.} \end{cases}$$

This L is not convex in (x, v), but, is convex in v for each (t, x). Consider the problem

$$\underset{x\in\mathcal{A}}{\text{minimize}}\ J(x) = \int_0^1 L(t, x(t), \dot{x}(t))\, dt$$

subject to $x(0) = (0, 0)^T$, $x(1) = (0, \pi)^T$. Obviously, for any $x \in \mathcal{A}$, $J(x) > 0$. From the form of L it would seem that we would want to delay our departure from the origin as long as possible and hence it might be profitable to attempt to use a generalization of the sequence $\{\bar{x}_n\}$ of Example 1. However this turns out to be prohibitively expensive since

$$\dot{\bar{x}}_n(t) = \begin{pmatrix} 0 \\ n\pi \end{pmatrix}$$

is not feasible for $t \in (1-1/n, 1-1/2n)$. Consider, instead, the more roundabout sequence $\{\tilde{x}_n\}$ where

$$\tilde{x}_n(t) = \begin{cases} \begin{pmatrix} 0 \\ 0 \end{pmatrix}, & t \in \left[0, 1-\frac{1}{n}\right], \\ \begin{pmatrix} n\pi\left(t-1+\frac{1}{n}\right) \\ 0 \end{pmatrix}, & t \in \left(1-\frac{1}{n}, 1-\frac{1}{2n}\right], \\ \begin{pmatrix} \pi/2 \\ 4n\pi\left(t-1+\frac{1}{n}\right)-2\pi \end{pmatrix}, & t \in \left(1-\frac{1}{2n}, 1-\frac{1}{4n}\right], \\ \begin{pmatrix} 2n\pi(1-t) \\ \pi \end{pmatrix}, & t \in \left(1-\frac{1}{4n}, 1\right]. \end{cases}$$

Then it can be verified that each $\tilde{x}_n$ is feasible and

$$J(\tilde{x}_n) = \frac{31\pi}{16n} \quad \text{which implies } \{\tilde{x}_n\} \text{ is a minimizing sequence.}$$

The important point here is that although $\{\bar{x}_n\}$ and $\{\tilde{x}_n\}$ have the same discontinuous limit, the cost that we want to attribute to the limit depends on how we arrive at that limit. This is the reason why the expression (2.2) does not apply here: the $\bar{r}$ term is not capable of handling the many different ways we can jump from one point to another. In the totally convex case it is cheapest to jump "in a straight line" following the pattern of $\{\bar{x}_n\}$ but this is often not so when L is only convex in v as this example demonstrates. We must look for a more discerning extension of J.

In the latter sections of this paper, we will use the existence result obtained for calculus of variations problems to determine when a solution exists for certain generalized problems in optimal control.

We will assume that the following hold throughout this paper.

General assumptions. $L: T \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a Lebesgue normal integrand. This is equivalent to the epigraph of L,

$$\text{epi } L(t, \cdot, \cdot) = \{(x, v, \alpha) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}: \alpha \geqq L(t, x, v)\}$$

being closed and depending Lebesgue measurably on t, in the sense that for each closed $D \subset \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}$, the set

$$\{t \in T: D \cap \text{epi } L(t, \cdot, \cdot) \neq \varnothing\}$$

is Lebesgue measurable. Normality implies that $L(t, x(t), v(t))$ is Lebesgue measurable in t whenever $(x(t), v(t))$ is and that $L(t, \cdot, \cdot)$ is lower semicontinuous for each $t \in T$. These results may be found in [12].

$$L(t, x, \cdot) \text{ is convex for each } (t, x),$$

$$l: \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\} \text{ is lower semicontinuous.}$$

DEFINITIONS. The *state constraint* multifunction $X: T \rightrightarrows \mathbb{R}^n$ is given by $X(t) = \{x: \exists v \text{ with } L(t, x, v) < \infty\}$. Since $X(t)$ may not be the whole space, we include problems with constraints.

The *Hamiltonian* $H: T \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ is defined by $H(t, x, p) = \sup_v \{p \cdot v - L(t, x, v)\}$.

The *adjoint state constraint* multifunction $P: T \times \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is given by

$$P(t, x) = \{p: H(t, x, p) < \infty\} = \operatorname{dom} H(t, x, \cdot).$$

The *recession function* for L is

$$r: T \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\},$$

$$r(t, x, \xi) = \lim_{\lambda \to +\infty} \frac{L(t, x, v + \lambda\xi) - L(t, x, v)}{\lambda}$$

where $v \in \operatorname{dom} L(t, x, \cdot)$. The function r is independent of v since $L(t, x, \cdot)$ is convex (see [8, Thm. 8.5]). Using [8, Thm. 13.3], we can also write

$$r(t, x, \xi) = \sup\{p \cdot \xi: p \in P(t, x)\} = \sup\{p \cdot \xi: p \in \operatorname{cl} P(t, x)\}.$$

If $r(t, x_1, \xi) = r(t, x_2, \xi)$ for all $x_1, x_2 \in X(t)$ then we can define $\bar{r}(t, \xi)$ to be this common value. This is the function that appears in (2.2).

A multifunction $S: T \rightrightarrows \mathbb{R}^n$ is upper semicontinuous if whenever K is a compact subset of $\mathbb{R}^n$, the set $\{t \in T: K \cap S(t) \neq \varnothing\}$ is closed.

A multifunction $S: T \rightrightarrows \mathbb{R}^n$ is *lower semicontinuous* if the set $\{t \in T: U \cap S(t) \neq \varnothing\}$ is open relative to T for every open $U \subset \mathbb{R}^n$.

A multifunction $S: T \rightrightarrows \mathbb{R}^n$ is *fully lower semicontinuous* if S is lower semicontinuous and one has $x_0 \in \operatorname{cl} S(\tau)$ whenever there are neighbourhoods U and V of x_0 and τ such that the set $\{t \in V: S(t) \supset U\}$ is dense in V. (This definition is taken from [9, p. 457].)

For the two cases, where (2.2) is valid and where it is not, we will have slightly different sets of hypotheses. For the former, they will be labelled (*A*), and for the latter, (*B*).

Assumptions (A).

Assumption (A1). There exists a multifunction $\bar{P}: T \rightrightarrows \mathbb{R}^n$ such that for all $x \in X(t)$, we have

$$\operatorname{cl} P(t, x) = \bar{P}(t) \quad \text{for all } t \in T.$$

If this holds, the recession function r will be independent of $x \in X(t)$, so we can define the function $\bar{r}$ such that

$$\bar{r}(t, \xi) = r(t, x, \xi) \quad \text{where } x \text{ is any element of } X(t).$$

This is the function that appears in (2.2). In that case, r was independent of x because L was totally convex, that is $L(t, \cdot, \cdot)$ was convex for all t.

Assumption (A2). $\bar{P}$ is fully lower semicontinuous on T and $\operatorname{int} \bar{P}(t) \neq \varnothing$ for any $t \in T$. Here $\operatorname{int} D$ denotes the interior of the set D.

Assumption (A3). The multifunction $X: T \rightrightarrows \mathbb{R}^n$ is upper semicontinuous and $X(t)$ is closed for each $t \in T$.

Assumption (A4). For each $x \in \mathscr{B}$ such that $x(t) \in X(t)$ a.e. one has

$$\int_V |H(t, x(t), p)|\, dt < \infty$$

whenever V is an open subset of T and p is a point of $\mathbb{R}^n$ having a neighbourhood U such that $U \subset \bar{P}(t)$, $\forall t \in V$.

This assumption is needed for the proof of Theorem 5 of [9], which we require. It is used to prove the following proposition. Define for $x \in \mathscr{B}$,

$$E_x = \left\{p \in \mathscr{C}: \int_T H(t, x(t), p(t))\, dt < \infty\right\}.$$

PROPOSITION 1. *Assume* (A1), (A2) *and* (A4) *hold. Then, if* $x(t) \in X(t)$ *a.e., one has*

$$\operatorname{int} E_x = \{p \in \mathscr{C}: p(t) \in \operatorname{int} \bar{P}(t), \forall t \in T\} \neq \varnothing.$$

(To simplify notation, we will call the right-hand side of the above equation, $\mathscr{E}$.)

Proof. Fix an $x \in \mathscr{B}$ such that $x(t) \in X(t)$ a.e. and let $f(t, \cdot) = H(t, x(t), \cdot)$. Then f is a normal convex integrand since it is the conjugate of $L(t, x(t), \cdot)$ (see [12]). The required result then follows from [9, Thm. 5]. Q.E.D.

Assumption (A5). For each $M \geqq 0$ and function $p \in \mathscr{E}$, there exists an integrable function $\Gamma: T \to \mathbb{R}$ such that whenever $x \in \mathscr{B}$ satisfies $x(t) \in X(t)$ a.e. and $\|x\|_V \leqq M$, then

$$H(t, x(t), p(t)) \leqq \Gamma(t) \quad \text{a.e.}$$

Note that the following simpler, but more restrictive assumption would imply both (A4) and (A5):

For each $M \geqq 0$ there exists a summable function $\Gamma: T \to \mathbb{R}$ such that

$$|H(t, x, p)| \leqq \Gamma(t)$$

when $x \in X(t)$, $p \in \bar{P}(t)$, $|x| \leqq M$ and $|p| \leqq M$.

Assumptions (B).

Assumption (B1).

$$0 \in \operatorname{cl} P(t, x) \quad \text{for all } (t, x).$$

This is equivalent to saying that $r(t, x, \xi) \geqq 0$ for all (t, x, ξ) and it implies $q(t, a, a) = 0$ so that the only contributions to the summation in (2.3) are from the atoms of dx, that is, where $x(t^-)$ and $x(t^+)$ differ. Since a function of bounded variation only has countably many atoms, the summation will be over a countable number of t values.

We will need the following bounded forms of P and r. Let $K > 0$ and define

$$P_K(t, x) = P(t, x) \cap B_K,$$

$$r_K(t, x, \xi) = \sup\{p \cdot \xi: p \in P_K(t, x)\}.$$

Of course, $r_K \leqq r$ (B_K is the closed ball of radius K).

Assumption (B2). For $x \in X(t)$, $\operatorname{int} P(t, x) \neq \varnothing$ for any $t \in T$.

Assumption (B3). For any K, M and ε all greater than zero, there exists a $\gamma_{K\varepsilon} > 0$ such that for all $x \in B_M$, $t \in T$, $\tilde{\gamma} \geqq \gamma_{K\varepsilon}$ and $\xi \in \mathbb{R}^n$ with $|\xi| \geqq 1$ we have $L(t, x, \tilde{\gamma}\xi)/\tilde{\gamma} \geqq r_K(t, x, \xi) - \varepsilon|\xi|$.

Assumption (B4). For any K, $M > 0$, $x \in X(t)$, $x' \in X(t')$ with $x, x' \in B_M$ there exist γ_3 and γ_4 such that

$$|r_K(t', x', \xi) - r_K(t, x, \xi)| \leqq \gamma_3|\xi| \cdot |t - t'| + \gamma_4|\xi| \cdot |x - x'|$$

for all $\xi \in \mathbb{R}^n$.

Assumption (B5). The multifunction $X: T \rightrightarrows \mathbb{R}^n$ is upper semicontinuous and $X(t)$ is closed for each $t \in T$.

Assumption (B6). For each $x \in \mathscr{A}$ such that $x(t) \in X(t)$ we have

$$\int_V |H(t, x(t), p)|\, dt < \infty$$

whenever V is an open subset of T and p is a point of $\mathbb{R}^n$ having a neighbourhood U such that $U \subset P(t, x(t))$, for all $t \in V$.

Assumption (B7). For each $M \geqq 0$ and set $D \subset \{x \in \mathscr{A}: \|x\|_V \leqq M \text{ and } x(t) \in X(t) \text{ for all } t \in T\}$, if there exists a function $\bar{p} \in \mathscr{C}$ such that

$$\bar{p}(t) \in \operatorname{int} \bigcap_{x \in D} P(t, x(t)),$$

then there exists an integrable function Γ, depending on M, D and $\bar{p}$ such that

$$H(t, x(t), \bar{p}(t)) \leqq \Gamma(t) \quad \text{a.e.}$$

for all $x \in D$. Assumption (B7), (or (A5) for its class of problems), is required in the proof of Proposition 2. Notice that assumptions (A) and (B) overlap, which is not surprising since one problem, and its proof, is a generalization of the other.

There is much correspondence between the two sets of assumptions. Assumption (B4) says that the multifunction $P(t, x)$ has some sort of Lipschitz property and when combined with (B2) would imply the weaker assumptions (A2) for $\bar{P}$. Assumptions (B5) and (A3) are the same, whereas (B6) and (B7) are generalizations of (A4) and (A5) respectively. Apart from (B1) whose inclusion has already been explained the only assumption in (B) that is not in (A) is (B3). This is some sort of uniformity condition on the Lagrangian. We know, because of convexity, that for fixed (t, x) the Lagrangian will, for large v, essentially behave like its recession function, but in the proof of the theorem we need to ensure this property, not only for a fixed (t, x), but for a whole range of them.

If assumptions (A) are satisfied, which imply the recession function is independent of x, then we will use the functional I as expressed in (2.2). However if the recession function is not independent of x for our problem, then we must resort to the functional Φ defined in the following manner.

Let $x \in \mathscr{B}$ and

$$dx(t) = \dot{x}(t)\, dt + \xi_s(t)\, d\theta_s(t) + \xi_a(t)\, d\theta_a(t)$$

where $\xi_s(t)\, d\theta_s(t)$ represents the smooth or nonatomic singular part of $dx(t)$ and $\xi_a(t)\, d\theta_a(t)$ represents the atomic singular part. Then, for the general problem where the adjoint state constraint $P(t, x)$, and hence the recession function r, can depend on x we will take for the extended functional the following

$$\begin{aligned} \Phi(x) = l(x(t_0), x(t_1)) &+ \int_T L(t, x(t), \dot{x}(t))\, dt \\ &+ \int_T r(t, x(t), \xi_s(t))\, d\theta_s(t) + \sum_{t \in T} q(t, x(t^-), x(t^+)) \end{aligned} \tag{2.3}$$

where $x \in \mathscr{B}$ and

$$q(t, a, b) = \inf_{y \in \mathscr{A}[0,1]} \left\{ \int_0^1 r(t, y(s), \dot{y}(s))\, ds: y(0) = a, y(1) = b \right\}. \tag{2.4}$$

It can be shown [5, Prop. II.2] that Φ reduces to the functional I in (2.2), when r is independent of $x \in X(t)$, and it is obviously equivalent to J of (2.1) when x belongs to $\mathscr{A}$. The problem associated with (2.3) is

$$\min_{x \in \mathscr{B}} \Phi(x). \tag{2.5}$$

To show that we can find an optimal solution for the problems (2.2) and (2.5), we will prove that under certain conditions the level sets

$$\begin{aligned} &\{x \in \mathcal{B}\colon I(x) \leqq \alpha\}, \\ &\{x \in \mathcal{B}\colon \Phi(x) \leqq \alpha\} \end{aligned} \tag{2.6}$$

where $\alpha \in \mathbb{R}$, are compact in an appropriate topology. The topology that we will be using is the weak* topology on $\mathcal{B}$. A neighbourhood of $\bar{x} \in \mathcal{B}$ in this topology is by definition any subset of $\mathcal{B}$ which contains a set of the form

$$V(\bar{x}, F, \delta) = \left\{ x \in \mathcal{B}\colon |\bar{x}(t_0) - x(t_0)| + \left| \int_T p(t)[d\bar{x}(t) - dx(t)] \right| < \delta, p \in F \right\}$$

where $\delta > 0$ and F is a finite subset of $\mathcal{C}$.

This topology is based on the pairing

$$\langle x, (a, p) \rangle = x(t_0) \cdot a + \int_T p(t)\, dx(t)$$

between $\mathcal{B}$ and $\mathbb{R}^n \times \mathcal{C}$. The latter is a Banach space under the norm $\|(a, p)\| = \max\{|a|, \|p\|\}$ and in this way $\mathcal{B}$ can be identified with the dual Banach space $(\mathbb{R}^n \times \mathcal{C})^*$.

Notation. If $\bar{x}$ is the limit of the generalized sequence x_ν in the weak* topology on $\mathcal{B}$, we shall write it as $x_\nu \to^{w^*} \bar{x}$.

Theorem 1 will be the basis for our general existence theorem. It shows that the functional Φ is weak* lower semicontinuous over sets that are bounded by the norm $\|\cdot\|_V$. We will later place other conditions on the problem which ensure that the level sets (2.6) are bounded with this norm. Then the level sets will be compact so that a minimum will exist, if the problem is feasible.

Theorem 1A provides the same sort of results for the functional I and problem (2.2).

DEFINITIONS. Let $M \geqq 0$. Then define

$$\Phi_M(x) = l(x(t_0), x(t_1)) + \int_T L(t, x(t), \dot{x}(t))\, dt + \int_T r(t, x(t), \xi_s(t))\, d\theta_s(t) + Q_M(x)$$

where

$$Q_M(x) = \inf_{\{m_i\}} \sum_{t_i \in T} \inf_y \left\{ \int_0^1 r(t_i, y(s), \dot{y}(s))\, ds\colon y(0) = x(t_i^-), \; y(1) = x(t_i^+), y \in \mathcal{A}[0, 1], \int_0^1 |\dot{y}(s)|\, ds \leqq m_i \right\}$$

where the $\{t_i\}$ consist of the atoms of dx and

$$\sum_i m_i = M - \left[|x(t_0)| + \int_T |\dot{x}(t)|\, dt + \int_T |\xi_s(t)|\, d\theta_s(t) \right]$$

where $m_i \geqq 0$ for all i. The first infimum in the definition of $Q_M(x)$ is over all such $\{m_i\}$. Notice that if $\|x\|_V > M$ then $Q_M(x)$ and hence $\Phi_M(x)$ are positively infinite.

THEOREM 1. *Let assumptions* (B) *hold. Also assume there exist constants* $\gamma_1, \gamma_2 \in \mathbb{R}$ *such that* $L(t, x, v) \geqq -\gamma_1|x| + \gamma_2$. *Then for all real numbers* α *and* M, *the set*

$$S_{\alpha,M} = \{x \in \mathcal{B}\colon \Phi_M(x) \leqq \alpha\} \tag{2.7}$$

is compact in the weak topology of* $\mathcal{B}$.

THEOREM 1A. *Let assumptions* (A) *hold. Then for all real numbers* α *and* M, *the set*

$$\tilde{S}_{\alpha,M} = \{x \in \mathcal{B}: I(x) \leqq \alpha, \|x\|_V \leqq M\}$$

is compact in the weak topology of* $\mathcal{B}$.

We will suffice ourselves here in proving Theorem 1. The proof of Theorem 1A (see [5]) follows similar, but simpler steps which we will later outline.

It will turn out useful to have an alternate form of Φ_M. This involves "fitting together" the various pieces of x and the y to form a new state vector $\tilde{x}$ which is absolutely continuous. This follows the approach used by Rishel in [5] to obtain necessary conditions. We will now construct this alternate form.

Fix a $\theta \in \mathcal{M}$. Define the function $\hat{\psi}: T \to \mathbb{R}$ by

$$\begin{aligned} &\hat{\psi}(t_0) = t_0, \\ &\hat{\psi}(\tau) = \int_{[t_0,\tau]} dt + \int_{[t_0,\tau]} d\theta(t) \quad \text{for } t_0 < \tau \leqq t_1. \end{aligned} \tag{2.8}$$

Let $\eta_0 = \hat{\psi}(t_0) = t_0$ and $\eta_1 = \hat{\psi}(t_1)$.

Define the multifunction ψ by

$$\begin{aligned} &\psi(t_0) = [\hat{\psi}(t_0), \hat{\psi}(t_0^+)], \\ &\psi(t) = [\hat{\psi}(t^-), \hat{\psi}(t)] \quad \text{for } t_0 < t \leqq t_1. \end{aligned}$$

Since $\hat{\psi}$ is strictly increasing, the inverse multifunction of ψ is actually a function, which we shall call ζ

$$\zeta: [\eta_0, \eta_1] \to \mathbb{R}.$$

We also have for $\eta_\alpha < \eta_\beta$,

$$0 \leqq \zeta(\eta_\beta) - \zeta(\eta_\alpha) \leqq \eta_\beta - \eta_\alpha.$$

So ζ is absolutely continuous and

$$0 \leqq \frac{d\zeta}{d\eta} \leqq 1.$$

Define $h: [\eta_0, \eta_1] \to \mathbb{R}$ by

$$h(\eta) = 1 - \frac{d\zeta}{d\eta}(\eta).$$

Then if $t \in T$ and $\bar{\eta}$ is such that $t = \zeta(\bar{\eta})$, we have

$$t = \zeta(\bar{\eta}) = \eta_0 + \int_{\eta_0}^{\bar{\eta}} \frac{d\zeta}{d\eta}(\eta)\, d\eta = \eta_0 + \int_{\eta_0}^{\bar{\eta}} (1 - h(\eta))\, d\eta. \tag{2.9}$$

From (2.8), (2.9) and the definition of ζ, we deduce that $(d\zeta/d\eta)(\eta) \in \{0, 1\}$, and likewise $h(\eta)$, $\forall \eta \in [\eta_0, \eta_1]$. If $t \in T$ corresponds to an $\bar{\eta} \in [\eta_0, \eta_1]$ via ζ, and $d\zeta(\bar{\eta})/d\eta = 0$, then t belongs to the support of θ almost always.

Now suppose that we are given an $x \in \mathcal{B}$ with

$$dx(t) = \dot{x}(t)\, dt + \xi(t)\, d\theta(t), \qquad \theta \in \mathcal{M}.$$

Although the choice of θ is not unique, we can choose a suitable one relative to x and fix it. This then determines a ζ and h.

DEFINITION.

$$\mathscr{A}_\theta(x)=\{\tilde{x}\in\mathscr{A}[\eta_0,\eta_1]:\ \tilde{x}(\eta_0)=x(t_0),\ \tilde{x}(\eta_1)=x(t_1) \text{ and } \tilde{x}(\eta)=x(\zeta(\eta)) \text{ a.e. when } h(\eta)=0\}.$$

Let us consider the following functional for $\tilde{x}\in\mathscr{A}_\theta(x)$

$$\Phi_\theta(\tilde{x},x)=l(\tilde{x}(\eta_0),\tilde{x}(\eta_1))+\int_{\eta_0}^{\eta_1}L\left(\zeta(\eta),\tilde{x}(\eta),\frac{d\tilde{x}}{d\eta}(\eta)\right)(1-h(\eta))\,d\eta + \int_{\eta_0}^{\eta_1}r\left(\zeta(\eta),\tilde{x}(\eta),\frac{d\tilde{x}}{d\eta}(\eta)\right)h(\eta)\,d\eta. \tag{2.10}$$

This can be rewritten as

$$\Phi_\theta(\tilde{x},x)=l(x(t_0),x(t_1))+\int_T L(t,x(t),\dot{x}(t))\,dt + \int_{\eta_0}^{\eta_1}r\left(\zeta(\eta),\tilde{x}(\eta),\frac{d\tilde{x}}{d\eta}(\eta)\right)h(\eta)\,d\eta.$$

DEFINITION.

$$\mathscr{A}_{M\theta}(x)=\{\tilde{x}\in\mathscr{A}_\theta(x):\ \|\tilde{x}\|_V\leqq M\}.$$

Using this definition and the positive homogeneity of $r(t,x,\cdot)$, we find that

$$\begin{aligned}\Phi_M(x)&=\inf_{\tilde{x}\in\mathscr{A}_{M\theta}(x)}\Phi_\theta(\tilde{x},x)\\&=l(x(t_0),x(t_1))+\int_T L(t,x(t)\dot{x}(t))\,dt\\&\quad+\inf_{\tilde{x}\in\mathscr{A}_{M\theta}(x)}\int_{\eta_0}^{\eta_1}r\left(\zeta(\eta),\tilde{x}(\eta),\frac{d\tilde{x}}{d\eta}(\eta)\right)h(\eta)\,d\eta.\end{aligned} \tag{2.11}$$

From the above we see that the choice of θ has no real effect on the value of Φ_M provided of course

$$dx(t)=\dot{x}(t)\,dt+\xi(t)\,d\theta(t).$$

So the size of $[\eta_0,\eta_1]$ can be chosen for any purpose we have in mind, subject to the constraint $[t_0,t_1]\subset[\eta_0,\eta_1]$. The formula in (2.11) is the alternate form for Φ_M.

PROPOSITION 2. *Let assumption* (B7) *hold. Let* $M\geqq 0$ *and* $D\subset\{x\in\mathscr{A}:\ \|x\|_V\leqq M$ *and* $x(t)\in X(t)$ *for all* $t\in T\}$. *For each* t *let* $\mathscr{D}(t)=\{x(t):\ x\in D\}$. *If there exists a* $p\in\mathscr{C}$ *such that* $p(t)\in\operatorname{int}\bar{P}(t)\equiv\operatorname{int}\bigcap_{y\in\mathscr{D}(t)}P(t,y)$ *then* $H(t,\cdot,p(t))$ *is upper semicontinuous on* $\mathscr{D}(t)$.

Proof. Write $L_0=-H$ and $K_0(t,x,p,v)=L(t,x,v)-pv$. Then $L_0(t,x,p)=\inf K_0(t,x,p,v)$ where $K_0(t,\cdot,\cdot,\cdot)$ is lower semicontinuous. Using the equivalence theorem of [11] it suffices to show that for fixed $t\in T$, $\alpha\in\mathbb{R}$ and $N\geqq 0$ the set

$$s=\{v:\ x\in\mathscr{D}(t) \text{ with } |x|\leqq N \text{ and } L(t,x,v)-pv\leqq\alpha\}$$

is bounded.

Since $p(t)\subset\operatorname{int}\bar{P}(t)$, $\exists\delta>0$ and $p_i\in\operatorname{int}\bar{P}(t)$, $i=1,\cdots,k$ such that

$$p(t)+B_\delta\subset\operatorname{co}\{p_1,\cdots,p_k\}\subset\operatorname{int}\bar{P}(t).$$

By (B7), $\exists \Gamma_i$ such that

$$H(t, x, p_i) \leqq \Gamma_i(t), \qquad i = 1, \cdots, k, \forall x \in \mathscr{D}(t).$$

By the convexity of $H(t, x, p)$ in p, we have

$$H(t, x, \tilde{p}) \leqq \Gamma(t) = \max_{i=1,\cdots,k} \Gamma_i(t) \quad \forall \tilde{p} \in p(t) + B_\delta, \forall x \in \mathscr{D}(t).$$

Then on $\mathscr{D}(t)$,

$$\tilde{p}v - L(t, x, v) \leqq \Gamma(t) \quad \forall \tilde{p} \in p(t) + B_\delta.$$

Hence

$$L(t, x, v) \geqq \sup_{\tilde{p} \in p(t)+B_\delta} \{\tilde{p}v - \Gamma(t)\} = p(t)v + \delta|v| - \Gamma(t).$$

Therefore on $\mathscr{D}(t)$

$$\alpha \geqq L(t, x, v) - p(t)v \geqq p(t)v + \delta|v| - \Gamma(t) - p(t)v = \delta|v| - \Gamma(t)$$

$$\Rightarrow |v| \leqq \frac{\alpha + \Gamma(t)}{\delta}$$

and S is bounded. Q.E.D.

PROPOSITION 3. *Let assumptions* (B5) *and* (B7) *hold. Define*

$$\psi(x, p) = \int_S H(t, x(t), p(t))\, dt$$

where S is a compact subset of T. Suppose $\bar{x} \in \mathscr{A}(S)$, $x_j \to \bar{x}$ uniformly on S and

$$\{x_j\}_{j=1}^{\infty} \subset \{x \in \mathscr{A}(S): \|x\|_V \leqq M, x(t) \in X(t), t \in S\}.$$

Let $\bar{p} \in \mathscr{C}(S)$. If $\bar{p}$ is such that $\exists \varepsilon > 0$ and k with

$$\bar{p}(t) + B_\varepsilon \subset P(t, x_j(t)) \quad \forall j \geqq k,$$

then

$$\limsup_{j \to \infty} \psi(x_j, \bar{p}) \leqq \psi(\bar{x}, \bar{p}).$$

Proof. Using Proposition 2 and Fatou's lemma, we have

$$\limsup \psi(x_j, \bar{p}) \leqq \int_S \limsup_{j \to \infty} H(t, x_j(t), \bar{p}(t))\, dt$$

$$\leqq \int_S H(t, \bar{x}(t), \bar{p}(t))\, dt = \psi(\bar{x}, \bar{p}). \qquad \text{Q.E.D.}$$

3. Proof of Theorem 1. Fix $\alpha \in \mathbb{R}$ and $M > 0$. To avoid the trivial, we will assume that $S_{\alpha,M}$ is nonempty.

Since $S_{\alpha,M} \subset \{x \in \mathscr{B}: \|x\|_V \leqq M\}$ and this last set is bounded in the strong topology, to show weak* compactness we only need to show that $S_{\alpha,M}$ is closed in the weak* topology. This task is made simpler by realizing that the boundedness of $S_{\alpha,M}$ causes it to be metrizable so it suffices to look at sequences to prove the weak* closure of $S_{\alpha,M}$.

Let $\{x_j\}_{j=1}^{\infty} \subset S_{\alpha,M}$ and $x_j \to^{w^*} \bar{x}$, where $\bar{x} \in \mathscr{B}$. By [4, Thm. 33, p. 291], $\|\bar{x}\|_V \leqq M$. Let

$$dx_j(t) = \dot{x}_j(t)\, dt + \xi_j(t)\, d\theta_j(t), \qquad j = 1, \cdots$$

and

$$d\bar{x}(t) = \dot{\bar{x}}(t)\, dt + \bar{\xi}(t)\, d\bar{\theta}(t).$$

Choose the θ_j and $\bar{\theta}$ so that the corresponding reparametrizations in the alternate form of Φ_M, (2.11), have the same endpoints η_0 and η_1. Since each of the θ_j and $\bar{\theta}$ has support whose Lebesgue measure is zero in T, then for any $\delta > 0$ we can find a relatively open set $A \subset T$ such that

$$m(A) < \delta \quad \text{and} \quad A \supset \bigcup_{j=1}^{\infty} \operatorname{supp}\{\theta_j\} \cup \operatorname{supp}\{\bar{\theta}\}$$

where $\operatorname{supp}\{\theta_j\}$ = the support set of θ_j and m is the Lebesgue measure.

Our mode of attack will be to show, separately, that on $T \backslash A$ and A, Φ_M is approximately weak* lower semicontinuous with respect to this sequence. As we make A smaller, the approximation improves until we obtain the required results.

Firstly, we will investigate the behaviour of Φ_M on a set similar to A. Fix $\varepsilon > 0$. For each j, choose an $\tilde{x}_j \in \mathscr{A}_{M\theta_j}(x_j)$ such that

$$|\Phi_M(x_j) - \tilde{\Phi}(\tilde{x}_j)| < \varepsilon \tag{3.1}$$

where $\tilde{\Phi}(\tilde{x}_j) = \Phi_{\theta_j}(\tilde{x}_j, x_j)$ (see (2.10)).

Although each of the $\tilde{x}_j$ are absolutely continuous, it may happen that as $j \to \infty$ they will approach a function that is not absolutely continuous, which is exactly the type of behaviour we are trying to model. To see how this limit behaves, we will reparametrize the problem yet again so that the derivatives involved are bounded and the limit functions will be absolutely continuous. In this endeavour, we will follow the approach of Warga [16].

For each j, define

$$\rho_j(\eta) = \max\left\{1, \left|\frac{d\tilde{x}_{j1}}{d\eta}(\eta)\right|, \cdots, \left|\frac{d\tilde{x}_{jn}}{d\eta}(\eta)\right|\right\}.$$

Because each of the $\tilde{x}_{ji}$, $i = 1, \cdots, n$ are integrable over $[\eta_0, \eta_1]$, ρ_j is integrable for each j. For each j, define

$$\alpha_j(\eta) = \eta_0 + \int_{\eta_0}^{\eta} \rho_j(s)\, ds.$$

Since $\rho_j(\eta) \geqq 1$ for all $\eta \in [\eta_0, \eta_1]$, α_j is strictly increasing. As α_j is also continuous, it has an inverse ω_j and

$$\omega_j(\tau) = \tau_0 + \int_{\tau_0}^{\tau} \beta_j(s)\, ds, \qquad \tau \in [\tau_0, \tau_{j1}]$$

where

$$0 \leqq \beta_j(\tau) \leqq \frac{1}{\rho_j(\omega_j(\tau))} \leqq 1$$

and $\tau_0 = \eta_0$, $\tau_{j1} = \alpha_j(\eta_1) \leqq M + (\eta_1 - \eta_0) \equiv \tau_1$. Let

$$\chi_j(\tau) = \tilde{x}_j(\omega_j(\tau)).$$

Then for $i = 1, \cdots, n$

$$0 \leqq \left|\frac{d\chi_{ji}}{d\tau}(\tau)\right| = \left|\frac{d\tilde{x}_{ji}}{d\eta}(\omega_j(\tau))\right| \beta_j(\tau) = \left|\frac{d\tilde{x}_{ji}}{d\eta}(\omega_j(\tau))\right| \Big/ \rho_j(\omega_j(\tau)) \leqq 1.$$

Let

$$z_j(\tau) = \zeta_j(\omega_j(\tau)),$$

where ζ_j corresponds to the metatime for x_j (see (2.9) and its derivation).

Then

$$0 \leqq \frac{dz_j}{d\tau}(\tau) = \frac{d\zeta_j}{d\eta}(\omega_j(\tau))\beta_j(\tau) \leqq \beta_j(\tau) \leqq 1.$$

Since $d\zeta_j/d\eta \in \{0, 1\}$, we also have $(dz_j/d\tau)(\tau) \in \{0, \beta_j(\tau)\}$. To ease the burden of notation, let us denote "$d/d\tau$" by "$'$". So, for example $dz_j/d\tau$ is replaced by $z_j'(\tau)$.

Although the triple (z_j, χ_j, ω_j) is defined on the interval $[\tau_0, \alpha_j(\eta_1)]$, we can extend it to $[\tau_0, \tau_1]$ by defining

$$\begin{aligned}(z_j(\tau), \chi_j(\tau), \omega_j(\tau) &= (z_j(\alpha_j(\eta_1)), \chi_j(\alpha_j(\eta_1)), \omega_j(\alpha_j(\eta_1)) \\ &= (t_1, \tilde{x}_j(\eta_1), \eta_1) \quad \text{when } \tau \in (\alpha_j(\eta_1), \tau_1].\end{aligned}$$

By the Arzela-Ascoli theorem, the sequence $\{(z_j, \chi_j, \omega_j)\}_{j=1}^{\infty}$ has a subsequence, which we shall again label in the same manner, that converges uniformly on $[\tau_0, \tau_1]$ to a triple $(\bar{z}, \bar{\chi}, \bar{\omega}) \in \mathscr{A}_{n+2}[\tau_0, \tau_1]$ that has the same bounds on its derivatives.

Our objective functional in the η variable was

$$\begin{aligned}\tilde{\Phi}(\tilde{x}_j) = l(\tilde{x}_j(\eta_0), \tilde{x}_j(\eta_1)) + \int_{\eta_0}^{\eta_1} \Bigg[& L\left(\zeta_j(\eta), \tilde{x}_j(\eta), \frac{d\tilde{x}_j}{d\eta}(\eta)\right)\frac{d\zeta_j}{d\eta}(\eta) \\ & + r\left(\zeta_j(\eta), \tilde{x}_j(\eta), \frac{d\tilde{x}_j}{d\eta}(\eta)\right)\left(1 - \frac{d\zeta_j}{d\eta}(\eta)\right)\Bigg] d\eta.\end{aligned}$$

Transforming the problem into the τ variable, we obtain

$$\begin{aligned}\hat{\Phi}(z_j, \chi_j, \omega_j) = {} & l(\chi_j(\tau_0), \chi_j(\tau_1)) \\ & + \int_{\tau_0}^{\tau_1} \Bigg[L\left(z_j(\tau), \chi_j(\tau), \frac{1}{\beta_j(\tau)}\chi_j'(\tau)\right)\frac{1}{\beta_j(\tau)} z_j'(\tau) \\ & + r\left(z_j(\tau), \chi_j(\tau), \frac{1}{\beta_j(\tau)}\chi_j'(\tau)\right)\left(1 - \frac{1}{\beta_j(\tau)} z_j'(\tau)\right)\Bigg]\beta_j(\tau)\, d\tau.\end{aligned}$$

Let

$$g_j(\tau) = \begin{cases} 0 & \text{when } z_j'(\tau) = 0, \\ 1 & \text{otherwise.} \end{cases}$$

Then

$$g_j(\tau) = z_j'(\tau)/\beta_j(\tau)$$

and

$$\begin{aligned}\hat{\Phi}_j \equiv {} & \hat{\Phi}(z_j, \chi_j, \omega_j) \\ = {} & l(\chi_j(\tau_0), \chi_j(\tau_1)) + \int_{\tau_0}^{\tau_1} \Bigg[L\left(z_j(\tau), \chi_j(\tau), \frac{1}{\beta_j(\tau)}\chi_j'(\tau)\right) g_j(\tau)\beta_j(\tau) \\ & + r(z_j(\tau), \chi_j(\tau), \chi_j'(\tau))(1 - g_j(\tau))\Bigg] d\tau.\end{aligned}$$

Define

$$D = \{\tau: \bar{g}(\tau) = 0\} \quad \text{and} \quad D_0 = \{\tau \in D: \bar{\beta}(\tau) = 0\}$$

where $\bar{\omega}'(\tau) = \bar{\beta}(\tau)$ and

$$\bar{g}(\tau) = \begin{cases} 0 & \text{when } \bar{z}'(\tau) = 0, \\ 1 & \text{otherwise.} \end{cases}$$

Since $\bar{\omega}$ is nondecreasing on $[\tau_0, \tau_1]$, there exist a countable number of distinct intervals D_i, $i = 1, 2, \cdots$ such that $m(D_0 \backslash \cup_{i=1}^{\infty} D_i) = 0$. As $\omega_j \to \bar{\omega}$ uniformly on $[\tau_0, \tau_1]$, for every $\varepsilon' > 0$ $\exists k$ such that $\forall j \geqq k$

$$|\omega_j(\tau) - \bar{\omega}(\tau)| < \varepsilon'. \tag{3.2}$$

There also exists an integer $\bar{n}$ such that, for some $\bar{\varepsilon}'$, $m(\cup_{i=1}^{\bar{n}} D_i) \geqq m(D_0) - \bar{\varepsilon}'$ and on each of the D_i we will have, using (3.2)

$$\int_{D_i} \beta_j(\tau)\, d\tau < 2\varepsilon', \qquad j \geqq k.$$

Therefore, for $j \geqq k$

$$\int_{D_0} |\beta_j(\tau)|\, d\tau = \int_{D_0} \beta_j(\tau)\, d\tau < \bar{\varepsilon}' + 2\bar{n}\varepsilon'.$$

Fixing $\bar{\varepsilon}'$ and hence $\bar{n}$, we can make ε' as small as we like and show that

$$\lim_{j \to \infty} \int_{D_0} |\beta_j(\tau)|\, d\tau < \bar{\varepsilon}'.$$

As $\bar{\varepsilon}'$ is arbitrary, we see that

$$\beta_j \xrightarrow{L_1^1} \bar{\beta} \quad \text{on } D_0.$$

So there is a subsequence which we shall still call $\{\beta_j\}_{j=1}^{\infty}$ such that

$$\beta_j(\tau) \to \bar{\beta}(\tau) \quad \text{a.e. on } D_0.$$

By Egorov's theorem,

$$\beta_j \to \bar{\beta} \quad \text{almost uniformly on } D_0.$$

Similarly $g_j \to \bar{g}$ almost uniformly on $D \backslash D_0$ (or more precisely, a subsequence of $\{g_j\}_{j=1}^{\infty}$) so for any $\varepsilon' > 0$, $\delta' > 0$, $\exists k$ (since the particular value of k is of no interest to us we will not worry that the same symbol k appears above) and a $D_{\varepsilon'}$ such that $\forall j \geqq k$

$$\beta_j(\tau) < \delta' \quad \text{or} \quad g_j(\tau) = 0 \quad \text{on } D_{\varepsilon'},$$
$$m(D \backslash D_{\varepsilon'}) < \varepsilon'.$$

Hence for $j \geqq k$, we have by the above and the assumption on L in the statement of the theorem, that the functional $\hat{\Phi}_j$ evaluated over D is,

$$\hat{\Phi}_j|_D \geqq \int_{D_{\varepsilon'}} \Big[L\Big(z_j(\tau), \chi_j(\tau), \frac{1}{\beta_j(\tau)} \chi_j'(\tau)\Big) \beta_j(\tau) g_j(\tau) + r(z_j(\tau), \chi_j(\tau), \chi_j'(\tau))(1 - g_j(\tau)) \Big]\, d\tau + (\gamma_2 - \gamma_1 M)\varepsilon' \tag{3.3}$$

where we have assumed, to simplify the expression, that $\tau_0, \tau_1 \notin D$ so that l does not appear.

Choosing a K and ε'', we obtain a $\gamma_{K\varepsilon''}$ from assumption (B3) such that

$$L\left(z_j, \chi_j, \frac{|\chi_j'|}{\beta_j}\cdot\frac{\chi_j'}{|\chi_j'|}\right)\frac{B_j}{|\chi_j'|} \geqq r_K\left(z_j, \chi_j, \frac{\chi_j'}{|\chi_j'|}\right) - \varepsilon'' \tag{3.4}$$

whenever $|\chi_j'|/\beta_j \geqq \gamma_{K\varepsilon''}$ which occurs for $j \geqq k$ on $D_{\varepsilon'}$ if

$$|\chi_j'| \geqq \delta'\gamma_{K\varepsilon''}.$$

We can rewrite (3.4) as,

$$L\left(z_j, \chi_j, \frac{\chi_j'}{\beta_j}\right)\beta_j \geqq r_K(z_j, \chi_j, \chi_j') - \varepsilon''|\chi_j'|$$

when $|\chi_j'| \geqq \delta'\gamma_{K\varepsilon''}$ on $D_{\varepsilon'}$. Therefore (4.3) gives

$$\begin{aligned}\hat{\Phi}_j|_D \geqq \int_{D_{\varepsilon'}} & [r_K(z_j(\tau), \chi_j(\tau), \chi_j'(\tau))g_j(\tau) \\ & + r(z_j(\tau), \chi_j(\tau), \chi_j'(\tau))(1-g_j(\tau))]\, d\tau \\ & + \delta'(\gamma_2 - \gamma_1 M)m(D_{\varepsilon'}) - \varepsilon'' M - K\delta'\gamma_{K\varepsilon''}m(D_{\varepsilon'}) + (\gamma_2 - \gamma_1 M)\varepsilon'\end{aligned}$$

(where the second last term appears because when $|\chi_j'| < \delta'\gamma_{K\varepsilon''}$ we have $r_K \leqq K|\chi_j'| < K\delta'\gamma_{K\varepsilon''}$. For ease of notation, let the last line of the above inequality be denoted by ν).

$$\geqq \int_{D_{\varepsilon'}} r_K(z_j(\tau), \chi_j(\tau), \chi_j'(\tau))\, d\tau + \nu.$$

Now X is a closed, upper semicontinuous multifunction by assumption (B5), so $\bar{\chi}(\tau) \in X(\bar{z}(\tau))$ because (z_j, χ_j) converges uniformly to $(\bar{z}, \bar{\chi})$, and $\chi_j(\tau) \in X(z_j(\tau))$. By assumption (B4), for any $\varepsilon''' > 0$, $\exists k$ such that $\forall j \geqq k$,

$$\hat{\Phi}_j|_D \geqq \int_{D_{\varepsilon'}} r_K(\bar{z}(\tau), \bar{\chi}(\tau), \chi_j'(\tau))\, d\tau - \varepsilon''' M(\gamma_3 + \gamma_4) + \nu. \tag{3.5}$$

By the properties of the χ_j' and χ_j and Theorem 3.1 of [1] we see that χ_j' converges weakly to $\bar{\chi}'$. So by Mazur's theorem [1, Thm. 3.2] there exists a sequence of convex combinations of the χ_j' that converges almost everywhere to $\bar{\chi}'$, that is

$$\sum_{i=1}^{\bar{k}(j)} \lambda_{ij}\chi'_{s(j)+i}(\tau) \to \bar{\chi}'(\tau) \quad \text{a.e.}$$

where $\sum_{i=1}^{\bar{k}(j)} \lambda_{ij} = 1$, $\lambda_{ij} \geqq 0$ and $s(j) + \bar{k}(j) < s(j+1) + 1$. Therefore, by (3.5) and Fatou's lemma

$$\begin{aligned}\liminf_j & \sum_{i=1}^{\bar{k}(j)} \lambda_{ij}\hat{\Phi}_{s(j)+i}|_D \\ & \geqq \int_{D_{\varepsilon'}} \liminf_j \sum_{i=1}^{\bar{k}(j)} \lambda_{ij} r_K(\bar{z}(\tau), \bar{\chi}(\tau), \chi'_{s(j)+i}(\tau))\, d\tau - \varepsilon''' M(\gamma_3 + \gamma_4) + \nu \\ & \geqq \int_{D_{\varepsilon'}} \liminf_j r_K\left(\bar{z}(\tau), \bar{\chi}(\tau), \sum_{i=1}^{\bar{k}(j)} \lambda_{ij}\chi'_{s(j)+i}(\tau)\right) d\tau \\ & \quad - \varepsilon''' M(\gamma_3 + \gamma_4) + \nu \quad \text{(since } r_K(t, x, \cdot) \text{ is convex)} \\ & \geqq \int_{D_{\varepsilon'}} r_K(\bar{z}(\tau), \bar{\chi}(\tau), \bar{\chi}'(\tau))\, d\tau \\ & \quad - \varepsilon''' M(\gamma_3 + \gamma_4) + \nu \quad \text{(since } r_K(t, x, \cdot) \text{ is lower semicontinuous).}\end{aligned}$$

Therefore

$$\liminf_j \hat{\Phi}_j|_D \geqq \int_{D_{\varepsilon'}} r_K(\bar{z}(\tau), \bar{\chi}(\tau), \bar{\chi}'(\tau))\, d\tau - \varepsilon''' M(\gamma_3+\gamma_4) + \nu.$$

Let $\bar{\varepsilon} > 0$. Then for our fixed K and ε' such that

$$-(\gamma_2 - \gamma_1 M)\varepsilon' < \frac{\bar{\varepsilon}}{5}$$

we can choose, in order,

$$\varepsilon''' M(\gamma_3+\gamma_4) < \frac{\bar{\varepsilon}}{5},$$

$$\varepsilon'' M < \frac{\bar{\varepsilon}}{5},$$

$$K\delta'\gamma_{K\varepsilon''} m(D_{\varepsilon'}) < \frac{\bar{\varepsilon}}{5},$$

$$-\delta'(\gamma_2 - \gamma_1 M) m(D_{\varepsilon'}) < \frac{\bar{\varepsilon}}{5}.$$

Hence,

$$\liminf_j \hat{\Phi}_j|_D \geqq \int_{D_{\varepsilon'}} r_K(\bar{z}(\tau), \bar{\chi}(\tau), \bar{\chi}'(\tau))\, d\tau - \bar{\varepsilon}.$$

Since K and ε' were arbitrary,

$$\begin{aligned}\liminf_j \hat{\Phi}(z_j, \chi_j, \alpha_j^{-1})|_D &\geqq \int_D r(\bar{z}(\tau), \bar{\chi}(\tau), \bar{\chi}'(\tau))\, d\tau - \bar{\varepsilon} \\ &= \hat{\Phi}(\bar{z}, \bar{\chi}, \bar{\alpha}^{-1})|_D - \bar{\varepsilon}.\end{aligned}$$

Therefore,

$$\liminf_j \hat{\Phi}(z_j, \chi_j, \alpha_j^{-1})|_D \geqq \hat{\Phi}(\bar{z}, \bar{\chi}, \bar{\alpha}^{-1})|_D. \tag{3.6}$$

Let us define an x^* such that

$$x^*(\bar{z}(\tau)) = \bar{\chi}(\tau), \qquad \tau \notin D.$$

Then x^* (or more precisely, its equivalence class) belongs to $\mathcal{B}$. What we wish to show is that x^* and $\bar{x}$ are equivalent and that from $\bar{\chi}$ we can construct an $\tilde{x} \in \mathcal{A}_{M\bar{\theta}}(\bar{x})$ so that (3.6) implies a corresponding result for the $\Phi(x_j)$ and $\Phi(\bar{x})$. Firstly let us show the equivalence of x^* and $\bar{x}$. The function $\bar{z}$ represents the original time variable t. We want to show that $x^*(\bar{z}(\tau)) = \bar{x}(\bar{z}(\tau))$ for almost all $\tau \notin D$ since the times $z(\tau)$, $\tau \in D$, correspond to the singular part of dx^*.

Fix $\varepsilon > 0$; then there exists an open set $A \subset T$ with $m(A) < \varepsilon$ such that $A \supset \bigcup_{j=1}^{\infty} z_j(D_j) \cup \bar{z}(D)$ where $D_j = \{\tau\colon z_j'(\tau) = 0\}$. Since $(z_j, \chi_j) \to (\bar{z}, \bar{\chi})$ uniformly, for every $\delta > 0$ $\exists k$ such that for $j \geqq k$

$$|z_j(\tau) - \bar{z}(\tau)| < \delta \quad \text{and} \quad |\chi_j(\tau) - \bar{\chi}(\tau)| < \delta \quad \forall \tau.$$

But on $T\backslash A$, $\bar{z}^{-1}$ and z_j^{-1}, $j = 1, 2, \cdots$ exist since we have $\bar{z}'(\tau) = 1$ when $\bar{z}(\tau) \in T\backslash A$ and $z_j'(\tau) = 1$ when $z_j(\tau) \in T\backslash A$. We will then have for $j \geqq k$ and $t \in T\backslash A$

$$|z_j^{-1}(t) - \bar{z}^{-1}(t)| < \delta. \tag{3.7}$$

Since $\bar{\chi}$ is uniformly continuous, given $\varepsilon' > 0$, we can choose the above δ sufficiently small such that

$$|\bar{\chi}(z_j^{-1}(t)) - \bar{\chi}(\bar{z}^{-1}(t))| < \varepsilon', \qquad t \in T \backslash A. \tag{3.8}$$

Therefore given $\varepsilon' > 0$ and a δ sufficiently small so that (3.7) and (3.8) are satisfied, there will exist a k such that for $j \geqq k$ and a.a. $t \in T \backslash A$

$$|x_j(t) - x^*(t)| < |\chi_j(z_j^{-1}(t)) - \bar{\chi}(z_j^{-1}(t))| + |\bar{\chi}(z_j^{-1}(t)) - \bar{\chi}(\bar{z}^{-1}(t))| < \delta + \varepsilon'.$$

In other words $x_j(t) \to x^*(t)$ a.e. on $T \backslash A$. Since A can be made arbitrarily small, we have that $x_j(t) \to x^*(t)$ a.e. on T. But $x_j(t) \to \bar{x}(t)$ a.e. so $\bar{x}(t) = x^*(t)$ a.e. and they belong to the same equivalence class.

Now we can use $\bar{\chi}$ to construct an $\bar{\bar{x}} \in A_{M\bar{\theta}}(\bar{x})$. It may have happened that over some interval $[\alpha, \beta]$ in D that $\bar{\chi}(\alpha) = \bar{\chi}(\beta)$. If this is the case, we can remove that portion of $\bar{\chi}$ and replace it with $\bar{\chi} = \bar{\chi}(\alpha)$ a constant value on $[\alpha, \beta]$ and make the corresponding objective functional no worse, and possibly better since on D the integrand is r and $r(\bar{z}, \bar{\chi}, 0) = 0$. This is equivalent to removing that portion of $[\tau_0, \tau_1]$ that corresponds to $[\alpha, \beta]$ and joining up the endpoints $\bar{\chi}(\alpha)$ and $\bar{\chi}(\beta)$. So suppose we have gone through this last procedure and we can find no subinterval of D that has equal $\bar{\chi}$ endpoints.

From $\bar{x}$ we can generate a function $\hat{\psi}$ which effectively produces a new time variable η. For every subinterval $D^i = [\tau_{i1}, \tau_{i2}]$ of D we can produce a corresponding interval $\bar{\omega}(D^i)$. Each D^i also corresponds to a time $\bar{z}(D^i)$ in T. An $\bar{\bar{x}} \in A_{M\bar{\theta}}(x)$ that will suffice is one that on the η interval $\nu_i = [\hat{\psi}(\bar{z}(D^i)^-), \hat{\psi}(\bar{z}(D^i)^+)]$ has

$$\bar{\bar{x}}(\eta) = \bar{\chi}\left(\tau_i^- + \frac{m(D^i)}{m(\nu_i)}(\eta - \hat{\psi}(\bar{z}(D^i)^-))\right)$$

where $\hat{\psi}$ is given by (2.8) for a $\bar{\theta}$ such that

$$d\bar{x}(t) = \dot{x}(t)\, dt + \bar{\xi}(t) d\bar{\theta}(t).$$

$m(\nu_i)$ will not be zero in the above because we have removed all subintervals $[\alpha, \beta]$ of D where $\bar{\chi}(\alpha) = \bar{\chi}(\beta)$. The rest of $\bar{\bar{x}}$ is given uniquely by $\bar{x}$ and hence $\bar{\chi}$. The positive homogeneity of r and the fact that the recession function of r is r will ensure that

$$\tilde{\Phi}(\bar{\bar{x}})|_{\bar{\omega}(D)} \leqq \hat{\Phi}(\bar{\chi})|_D.$$

Then (3.6) and the above imply

$$\liminf_j \tilde{\Phi}(\bar{\bar{x}}_j)|_{\bar{\omega}(D)} \geqq \tilde{\Phi}(\bar{\bar{x}})|_{\bar{\omega}(D)}.$$

As before, for any $\bar{\varepsilon} > 0$ we can find a set $A \supset \bar{z}(D)$ such that $m(A) < \bar{\varepsilon}$. By the assumption on L in the statement of the Theorem,

$$\int_A L(t, x_j(t), \dot{x}_j(t))\, dt \geqq \bar{\varepsilon}(\gamma_2 - \gamma_1 M).$$

If the following holds:

$$\int_T L(t, \bar{x}(t), \dot{\bar{x}}(t))\, dt < \infty, \tag{3.9}$$

then using (3.1) and the above, we find

$$\begin{aligned}\liminf_j \Phi(x_j)|_A &\geqq \liminf_j \tilde{\Phi}(\tilde{x}_j)|_{\bar{\omega}(D)} - \varepsilon + \bar{\varepsilon}(\gamma_2 - \gamma_1 M)\\ &\geqq \tilde{\Phi}(\bar{\tilde{x}})|_{\bar{\omega}(D)} - \varepsilon + \bar{\varepsilon}(\gamma_2 - \gamma_1 M)\\ &\geqq \Phi(\bar{x})|_A - \varepsilon + \bar{\varepsilon}(\gamma_2 - \gamma_1 M) - \int_A L(t, \bar{x}(t), \dot{\bar{x}}(t))\, dt.\end{aligned}$$

So if (3.9) holds, then for any $\varepsilon' > 0$ we can find an A containing the singular parts of $d\bar{x}$ and the dx_j, $j = 1, 2, \cdots$ such that

$$\liminf_j \Phi(x_j)|_A \geqq \Phi(\bar{x})|_A - \varepsilon'. \tag{3.10}$$

To finish the proof, we need to display similar behaviour on $T\backslash A$, and to verify (3.9).

Suppose (3.9) does not hold. Then we can replace it in the following with some arbitrarily large number. Then for some comparable number N, we can use the lower bound assumed on L in the statement of the theorem to show that

$$\liminf \Phi(x_j)|_{T\backslash A} \geqq N$$

in contradiction to $\{x_j\}_{j=1}^{\infty} \subset S_{\alpha,M}$. So we can assume that (3.9) does hold.

Let $\hat{x}$ be an arbitrary element of $\{x_j\}_{j=1}^{\infty} \cup \{\bar{x}\}$. On $T\backslash A$, $\hat{x}$ is absolutely continuous. Define

$$\hat{P}(t) = P(t, \hat{x}(t)) \quad \forall t \in T\backslash A.$$

Using [9, Thm. 5], we can write (where we have omitted the l term for convenience. Its lower semicontinuity ensures the corresponding results apply when it is included)

$$\begin{aligned}\Phi(\hat{x})|_{T\backslash A} &= \int_{T\backslash A} L(t, \hat{x}(t), \dot{\hat{x}}(t))\, dt\\ &= \sup_{p \in \mathscr{C}_{\hat{x}}} \left\{ \int_{T\backslash A} p(t)\dot{\hat{x}}(t)\, dt - \int_{T\backslash A} H(t, \hat{x}(t), p(t))\, dt \right\} \qquad (3.11)\\ &\equiv \sup_{p \in \mathscr{C}_{\hat{x}}} \phi(\hat{x}, p)|_{T\backslash A}\end{aligned}$$

where

$$\mathscr{C}_{\hat{x}} = \{p \in \mathscr{C}: p(t) \in \operatorname{int} \hat{P}(t), \forall t \in T\backslash A\}.$$

In light of (3.11) and the lower bound on $L(\cdot, \cdot, \cdot)$ for $|x| \leqq M$, we can find a $\bar{p} \in \mathscr{C}_{\bar{x}}$ such that

$$|\Phi(\bar{x})|_{T\backslash A} - \phi(\bar{x}, \bar{p})|_{T\backslash A}| < \varepsilon.$$

Suppose $x_j \to \bar{x}$ uniformly on $T\backslash A$ (this need not be the case, but we will show later that it is approximately true). By assumption (B4)

$$P_K(t, x') \subset P_K(t, x) + B_{\gamma_4|x-x'|}.$$

For K sufficiently large,

$$\bar{p}(t) \in \operatorname{int} P_K(t, \bar{x}(t)) \quad \forall t \in T\backslash A.$$

So $\exists \bar{\delta} > 0$ such that

$$\bar{p}(t) + B_{\bar{\delta}} \subset \operatorname{int} P_K(t, \bar{x}(t)) \quad \forall t \in T\backslash A.$$

Since $x_j \to \bar{x}$ uniformly on $T\backslash A$, there exists a k such that $\forall j \geqq k$

$$|x_j(t) - \bar{x}(t)| \leqq \frac{\bar{\delta}}{2\gamma_4} \quad \forall t \in T\backslash A.$$

Hence for $j \geqq k$, $\bar{p}(t) + B_{\bar{\delta}/4} \subset P(t, x_j(t))$. By Proposition 3

$$\liminf_j \phi(x_j, \bar{p})|_{T\backslash A} \geqq \phi(\bar{x}, \bar{p})|_{T\backslash A}.$$

But

$$\liminf_j \Phi(x_j)|_{T\backslash A} \geqq \liminf (\phi(x_j, \bar{p})|_{T\backslash A} \geqq \phi(\bar{x}, \bar{p})|_{T\backslash A} \geqq \Phi(\bar{x})|_{T\backslash A} - \varepsilon.$$

That is,

$$\liminf_j \Phi(x_j)|_{T\backslash A} \geqq \Phi(\bar{x})|_{T\backslash A} - \varepsilon \tag{3.12}$$

which is analogous to (3.10). However, in the above, we assumed that $x_j \to \bar{x}$ uniformly on $T\backslash A$. In general this is not implied by sequential weak* convergence, but we may do the following.

As we stated earlier, there exists a subsequence which we shall also call $\{x_j\}$ such that

$$x_j(t) \to \bar{x}(t) \quad \text{a.e.}$$

By Egorov's theorem, for given $\hat{\varepsilon} > 0$, there exists a relatively open set $V \subset T\backslash A$ such that $m(V) < \hat{\varepsilon}$ and

$$x_j \to \bar{x} \quad \text{uniformly on } (T\backslash A)\backslash V.$$

By choosing $\hat{\varepsilon}$ sufficiently small we have for $\varepsilon' > 0$

$$\Phi(\bar{x})|_V < \varepsilon' \quad \text{and} \quad \Phi(x_j)|_V \geqq \hat{\varepsilon}(\gamma_2 - \gamma_1 M).$$

These become negligible for small enough $\hat{\varepsilon}$, and we are left with $x_j \to \bar{x}$ uniformly on $(T\backslash A)\backslash V$, where we apply the previous reasoning. Thus (3.12) holds.

Combining (3.12) and (3.10) and realizing that ε is arbitrary, we obtain

$$\liminf_j \Phi(x_j) \geqq \Phi(\bar{x})$$

and the proof of Theorem 1 is complete. Q.E.D.

Remark. If we let $x_j = \bar{x}$ for all j and we choose a minimizing sequence of $\tilde{\Phi}$, (see (3.1)), we find that the above proof implies that the infimum in (2.11), and hence the infima in the definition of Φ_M, are attained.

Let a function $g: T \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be such that $g(t, x, \cdot)$ is nonnegative, convex, positively homogeneous and lower semicontinuous for all (t, x). Then we can define the set

$$G(t, x) = \{\zeta: \zeta \cdot \xi \leqq g(t, x, \xi) \text{ for all } \xi \in \mathbb{R}^n\}$$

and following the guidance of r_k and P_K, we can also define g_K and G_K.

Let us define the functional Φ_g on $\mathcal{B}$ by

$$\Phi_g(x) = l(x(t_0), x(t_1)) + \int_T L(t, x(t), \dot{x}(t))\, dt$$

$$+ \int_T g(t, x(t), \xi_s(t))\, d\theta_s(t) + \sum_{t \in T} q_g(t, x(t^-), x(t^+))$$

where

$$q_g(t, a, b) = \inf\left\{\int_0^1 g(t, y(s), \dot{y}(s))\, ds\colon y(0) = a,\ y(1) = b \text{ and } y \in \mathcal{A}[0, 1]\right\}.$$

We can also define Φ_{gM} analogously to Φ_M.

COROLLARY 1. *Let g have the properties stated above and let the assumptions of Theorem* 1 *be met but with g_K in place of r_K in assumption* (B4). *If $g \leqq r$ then*

$$Z_{\alpha,M} = \{x \in \mathcal{B}\colon \Phi_{g,M} \leqq \alpha\}$$

is compact in the weak topology on $\mathcal{B}$.*

Proof. This follows by making the appropriate substitutions in the proof of Theorem 1. Q.E.D.

What happens if we define a functional Φ_g as above but where $g(t, x, \xi) > r(t, x, \xi)$ for some (t, x, ξ)? If this were the case, we could not be certain that an optimal solution exists, for it may happen that if $\{x_j\} \subset \mathcal{A}$ and $x_j \to^{w^*} \bar{x}$ with $\bar{x} \in \mathcal{B}\backslash\mathcal{A}$, then

$$\lim_j \Phi_g(x_j) = \Phi(\bar{x}) < \Phi_g(\bar{x}).$$

So Theorem 1 and Corollary 1 show that the functional Φ with its terms containing the recession function for L gives an upper-limit on the well defined extensions of the original functional over $\mathcal{A}$.

Outline of proof of Theorem 1*A.* The major simplifying factor in the proof of Theorem 1A is that the "straight line" is the cheapest way to go between two points although this fact is more of an outcome of the theorem rather than the basis for its proof. Not only does this simplify the form of the functional, but in the proof, one does not have to divide the interval into two components A and $T\backslash A$, one only has to follow the argument on the $T\backslash A$ section. Then, the fact that the adjoint state constraint multifunction P is independent of x simplifies the process even further in as much as Propositions 1 and 2 become easier and the set $\mathcal{E}_{\hat{x}}$ is the same for all $\hat{x}$. The details of this argument can be found in [5].

4. An existence theorem for calculus of variations. To obtain an existence theorem, we need to place conditions on the problem so that

$$\{x \in \mathcal{B}\colon I(x) \leqq \alpha\} \quad \text{and} \quad \{x \in \mathcal{B}\colon \Phi(x) \leqq \alpha\}$$

will be compact in the weak* topology.

Assumptions (G).

Assumption (G1). There exists a function $\bar{p}$ which is continuously differentiable on T constants γ_1 and γ_2 with γ_1 nonnegative, and a $\delta > 0$ with $\delta \geqq \|\bar{p}\|_\infty$ such that for all (t, x)

$$H(t, x, p) \leqq \gamma_1|x| - \dot{\bar{p}}(t)x - \gamma_2 \quad \text{when } |p - \bar{p}(t)| \leqq \delta.$$

Notice that assumptions (B1), (B2) and the assumption in the statement of Theorem 1 are implied by (G1).

Assumption (G2). For the function $\bar{p}$ above, the following holds for all a, b in $\mathbb{R}^n$

$$l(a, b) \geqq k(|a|) - \bar{p}(t_1)b - \rho|b|$$

where $\rho \geqq 0$ and $k\colon [0, \infty) \to \mathbb{R} \cup \{+\infty\}$ is a nondecreasing function such that

$$\lim_{s \to +\infty} \frac{k(s)}{s} = +\infty.$$

Assumption (G3). The parameters δ, ρ and γ_1 in assumptions (G1) and (G2) can be chosen such that $\delta - \rho - \gamma_1$ is greater than zero.

THEOREM 2. *Let assumptions* (G) *hold. Then for each* $\alpha \in \mathbb{R}$, *there exists an* $M > 0$ *such that*

$$\{\tilde{x} \in \mathscr{A}_\theta(x): x \in \mathscr{B} \text{ and } \Phi_\theta(\tilde{x}, x) \leqq \alpha\} \subset \{\tilde{x} \in \mathscr{A}_\theta(x): x \in \mathscr{B} \text{ and } \|\tilde{x}\|_V \leqq M\}.$$

Using this result we can replace Φ_M with Φ in (2.7) and thereby obtain an existence theorem for the extended calculus of variations problem.

Proof. Because $L(t, x, \cdot)$ is convex and lowersemicontinuous, the inequality in (G1) is equivalent to

$$\begin{aligned} L(t, x, v) &= \sup\{pv - H(t, x, p)\} \\ &= \sup\{(p + \bar{p}(t))v - H(t, x, p + \bar{p}(t))\} \\ &\geqq \bar{p}(t)v + \delta|v| - \gamma_1|x| + \bar{p}(t)x + \gamma_2. \end{aligned}$$

Fix $\alpha \in \mathbb{R}$ and choose an $x \in \mathscr{B}$ and an $\tilde{x} \in \mathscr{A}_\theta(x)$ such that $\Phi_\theta(\tilde{x}, x) \leqq \alpha$ where Φ_θ is defined in (2.10)

$$\begin{aligned} \alpha &\geqq l(\tilde{x}(\eta_0), \tilde{x}(\eta_1)) + \int_{\eta_0}^{\eta} L\left(\zeta(\eta), \tilde{x}(\eta), \frac{d\tilde{x}}{d\eta}(\eta)\right)(1 - h(\eta))\, d\eta \\ &\quad + \int_{\eta_0}^{\eta_1} r\left(\zeta(\eta), \tilde{x}(\eta), \frac{d\tilde{x}}{d\eta}(\eta)\right)h(\eta)\, d\eta \\ &\geqq l(\tilde{x}(\eta_0), \tilde{x}(\eta_1)) + \int_{\eta_0}^{\eta_1} \bar{p}(\zeta(\eta))\frac{d\tilde{x}}{d\eta}(\eta)\, d\eta + \delta\int_{\eta_0}^{\eta_1}\left|\frac{d\tilde{x}}{d\eta}(\eta)\right| d\eta \\ &\quad - \gamma_1\int_{\eta_0}^{\eta_1}|\tilde{x}(\eta)|\, d\eta + \int_{\eta_0}^{\eta_1}\frac{d\bar{p}}{d\eta}(\zeta(\eta))\tilde{x}(\eta)\, d\eta + \gamma_2 T \\ &\geqq k(|\tilde{x}(\eta_0)|) - \bar{p}(\zeta(\eta_1))\tilde{x}(\eta_1) - \rho|\tilde{x}(\eta_1)| + \int_{\eta_0}^{\eta_1}\bar{p}(\zeta(\eta))\frac{d\tilde{x}}{d\eta}(\eta)\, d\eta \\ &\quad + \delta\int_{\eta_0}^{\eta_1}\left|\frac{d\tilde{x}}{d\eta}(\eta)\right| d\eta - \gamma_1\int_{\eta_0}^{\eta_1}|\tilde{x}(\eta)|\, d\eta + \int_{\eta_0}^{\eta_1}\frac{d\bar{p}}{d\eta}(\zeta(\eta))\tilde{x}(\eta)\, d\eta + \gamma_2 T \\ &= k(|\tilde{x}(\eta_0)|) - \rho|\tilde{x}(\eta_1)| - \bar{p}(\zeta(\eta_0))\tilde{x}(\eta_0) \\ &\quad + \delta\int_{\eta_0}^{\eta_1}\left|\frac{d\tilde{x}(\eta)}{d\eta}\right| d\eta - \gamma_1\int_{\eta_0}^{\eta_1}|\tilde{x}(\eta)|\, d\eta + \gamma_2 T \\ &\geqq k(|\tilde{x}(\eta_0)|) - |\tilde{x}(\eta_0)|(\rho + |\bar{p}(\zeta(\eta_0))| + \gamma_1) \\ &\quad - (\rho + \gamma_1 - \delta)\int_{\eta_0}^{\eta_1}\left|\frac{d\tilde{x}(\eta)}{d\eta}\right| d\eta + \gamma_2 T. \end{aligned}$$

By the definition of k there exists a $\lambda \in \mathbb{R}$ such that for $s \geqq 0$

$$k(s) - (|\bar{p}(\eta_0)| + \delta)s \geqq \lambda.$$

Hence

$$\alpha - \lambda - \gamma_2 T \geqq (\delta - \rho - \gamma_1)\left(|\tilde{x}(\eta_0)| + \int_{\eta_0}^{\eta_1}\left|\frac{d\tilde{x}}{d\eta}(\eta)\right| d\eta\right).$$

By (G3), $\delta - \rho - \gamma_1$ is positive and we can choose our M in the theorem to be

$$\frac{\alpha - \lambda - \gamma_2 T}{\delta - \rho - \gamma_1}.$$ Q.E.D.

Combining Theorems 1 and 2, we obtain the following existence theorem.

THEOREM 3. *Let assumptions* (B) *and* (G) *hold. Then for each* $\alpha \in \mathbb{R}$, *the level sets*

$$S_\alpha = \{x \in \mathscr{B}: \Phi(x) \leqq \alpha\}$$

are compact in the weak topology. In particular, an optimal solution exists for problem* (2.5) *if there is a feasible solution.*

Remark. Theorem 2 and the remark following the proof of Theorem 1 imply that under the above conditions the infima in the definition of Φ are attained so that we can replace inf with min. Applying similar reasoning to Theorem 1A, we obtain

THEOREM 3A. *Let assumptions* (A) *and* (G) *hold. Then for each* $\alpha \in \mathbb{R}$, *the level sets*

$$\tilde{S}_\alpha = \{x \in \mathscr{B}: I(x) \leqq \alpha\}$$

are compact in the weak topology. In particular, an optimal solution exists for problem* (2.2) *if there is a feasible solution.*

Remark. As far as Theorem 3A is concerned, one can make do with a weaker version of assumption (G). For example we only need $\bar{p} \in \mathscr{A}$ and it is not necessary that $\|\bar{p}\|_\infty \leqq \delta$.

Although the optimal solution will, in general, not be absolutely continuous, there are cases when it is. The most obvious of these is when the recession function is trivial.

COROLLARY 2. *Let assumptions* (A) *and* (G) *hold, and*

$$\bar{r}(t, \xi) = \begin{cases} 0 & \text{if } \xi = 0, \\ +\infty & \text{otherwise.} \end{cases}$$

If problem (2.2) is feasible, then an optimal solution will exist and will belong to $\mathscr{A}$.

Note. The condition in the above corollary is equivalent to

$$H(t, x, p) < \infty \ \forall p, \quad \text{or} \quad P(t, x) \equiv \mathbb{R}^n.$$

An example of a class of problems that has an optimal solution belonging to $\mathscr{A}$ although other optimal solutions will exist that belong to $\mathscr{B} \setminus \mathscr{A}$, is contained in the following corollary to Theorem 3.

COROLLARY 3. *Let assumptions* (B) *and* (G) *hold. Also assume that L is parametric, in other words, that L is independent of t and* $L(x, \lambda \dot{x}) = \lambda L(x, \dot{x})$ *for* $\lambda \geqq 0$. *If problem* (2.4) *is feasible, then an optimal solution that belongs to* $\mathscr{A}$ *will exist.*

Proof. If $(x, \dot{x}) \in \operatorname{dom} L$ then $(x, 0) \in \operatorname{dom} L^\infty$ and

$$r(x, \xi) = \lim_{\lambda \to \infty} \frac{L(x, \lambda \xi)}{\lambda} = L(x, \xi).$$

Because of this, assumption (B3) will automatically be satisfied and

$$\Phi(x) = \min_{\tilde{x} \in \mathscr{A}_\theta(x)} \left\{ l(\tilde{x}(\eta_0), \tilde{x}(\eta_1)) + \int_{\eta_0}^{\eta_1} \left[L\left(\tilde{x}(\eta), \frac{d\tilde{x}}{d\eta}(\eta)\right)(1 - h(\eta)) + r\left(\tilde{x}(\eta), \frac{d\tilde{x}}{d\eta}(\eta)\right) h(\eta) \right] d\eta \right\}$$

$$= \min_{\tilde{x} \in \mathscr{A}_\theta(x)} \left\{ l(\tilde{x}(\eta_0), \tilde{x}(\eta_1)) + \int_{\eta_0}^{\eta_1} L\left(\tilde{x}(\eta), \frac{d\tilde{x}}{d\eta}(\eta)\right) d\eta \right\}.$$

Since an optimal solution $\bar{x}$ of (2.5) exists if the problem is feasible, we can find an $\tilde{\bar{x}} \in \mathscr{A}_{\tilde{\theta}}(\bar{x})$ that achieves the minimum in the above. Reparametrizing $\tilde{\bar{x}}$, we can find an $\hat{x} \in \mathscr{A}$ such that

$$\Phi(\bar{x}) = l(\hat{x}(t_0), \hat{x}(t_1)) + \int_T L(\hat{x}(t), \dot{\hat{x}}(t))\, dt$$

and so $\hat{x}$ is optimal. Q.E.D.

Example 3. Consider the model of a fishery with irreversible investment (see [3]). Its description is as follows

$$\text{minimize } \mathscr{J} = \int_0^\infty e^{-\delta t}[\pi I(t) + cE(t) - pqE(t)x(t)]\, dt$$

subject to

$$\dot{x}(t) = F(x(t)) - qE(t)x(t), \qquad x(0) = x_0,$$
$$\dot{K}(t) = I(t) - \gamma K(t), \qquad K(0) = K_0,$$

and

$$0 \leqq E(t) \leqq K(t), \qquad 0 \leqq I(t), \qquad 0 \leqq x(t)$$

where

$x(t)$ = fish population biomass at time t

$K(t)$ = capital at time t

$E(t)$ = fishing effort at time t

$I(t)$ = investment rate at time t

$F(\cdot)$ = natural growth function of the biomass

π = price of capital

c = operating cost per unit effort

p = price of landed fish

q = catchability coefficient

δ = instantaneous rate of discount

γ = rate of depreciation.

The above problem is not a simple one to solve for several reasons. It is over an infinite interval, it has mixed control and state constraints ($0 \leqq E \leqq K$) and it allows unbounded investment rate without the objective function penalizing it via a growth condition, as would be necessary to apply the usual existence theorems. In fact, the optimal solution, as calculated in [3] can have jumps in K, brought about by an infinite investment rate. So attempting to minimize J over $\mathscr{A}$ will be a fruitless search. If we were to replace the infinite time interval with a finite one, we could analyse the problem via one of our extended problems. In that case, what would be the extended functional? The next section deals with transforming control problems to calculus of variations problems but in this case we can do it more directly.

Replace I with $\dot{K}+\gamma K$ and E with $(F(x)-\dot{x})/qx$. Then the corresponding calculus of variations problem is

$$\text{minimize } J=\int_0^\infty e^{-\delta t}\left[\pi(K(t)+\gamma K(t))+\left(\frac{c}{qx(t)}-p\right)(F(x(t))-\dot{x}(t))\right]dt$$

subject to $x(0)=x_0$, $K(0)=K_0$

$$0\leqq\frac{F(x)-\dot{x}}{qx}\leqq K, \qquad 0\leqq x, \quad 0\leqq\dot{K}.$$

Let $y=(x, K)^T$, $v=(\dot{x}, \dot{K})^T$; then the Lagrangian is convex in v for each (t, y) and the recession function is independent of y and has the form

$$\bar{r}(t, \xi)=\begin{cases} e^{-\delta t}\pi\xi_2 & \text{if } \xi=(\xi_1, \xi_2)^T \text{ with } \xi_1=0,\ \xi_2\geqq 0,\\ +\infty & \text{otherwise.}\end{cases}$$

Extending the problem to $\mathcal{B}$ then leads to

$$\text{minimize } \hat{J}=\int_0^\infty e^{-\delta t}\left[\pi(\dot{K}(t)+\gamma K(t))+\left(\frac{c}{qx(t)}-p\right)(F(x(t))-\dot{x}(t))\right]dt + \int_0^\infty e^{-\delta t}\pi\xi_2(t)\,d\theta(t)$$

where

$$dK(t)=\dot{K}(t)\,dt+\xi_2(t)\,d\theta(t), \qquad x(0)=x_0, \quad \theta\in\mathcal{M},$$
$$K(0)=K_0,$$
$$0\leqq\frac{F(x)-\dot{x}}{qx}\leqq K, \qquad 0\leqq x, \quad 0\leqq\dot{K}, \quad 0\leqq\xi_2.$$

This is essentially the extension of J to arcs in $\mathcal{B}$ that is used in [3].

5. An existence theorem for optimal control. Many optimal control problems can be stated in the following compact form

$$\text{minimize } \tilde{\Psi}(x, u)=l(x(t_0), x(t_1))+\int_T K(t, x(t), \dot{x}(t), u(t))\,dt \tag{5.1}$$

over all $x\in\mathcal{A}$ and $u\in L^1_m$, where

$$K: T\times\mathbb{R}^n\times\mathbb{R}^n\times\mathbb{R}^m\to\mathbb{R}\cup\{+\infty\}.$$

We will assume that K is a Lebesgue normal integrand and $K(t, x, \cdot, \cdot)$ is convex for each (t, x); l will have the same properties as it did for the calculus of variations problem.

To see how the problem (5.1) and the corresponding assumptions relate to an optimal control problem in the usual form, we refer the reader to [10].

We can convert (5.1) to a calculus of variations problem, by defining

$$L(t, x, v)=\inf_u K(t, x, v, u). \tag{5.2}$$

The convexity of $K(t, x, \cdot, \cdot)$ ensures the convexity of $L(t, x, \cdot)$. If K satisfies assumption (U1) below, we will then have a problem with the same form as (2.1).

We wish to extend the functional in (5.1) so that it can evaluate any $x \in \mathcal{B}$. Let $\hat{r}$ be the recession function of K, that is

$$\hat{r}(t, x, \xi, \mu) = \lim_{\lambda \to \infty} \left\{ \frac{K(t, x, v+\lambda\xi, u+\lambda\mu) - K(t, x, v, u)}{\lambda} \right\}$$

where $(x, v, u) \in \operatorname{dom} K(t, \cdot, \cdot, \cdot)$.

Our extension of (5.1) is then

$$\begin{aligned} \text{minimize } \Psi(x, \nu) = l(x(t_0), x(t_1)) &+ \int_T K(t, x(t), \dot{x}(t), u(t))\, dt \\ &+ \int_T \hat{r}(t, x(t), \xi_s(t), \mu_s(t))\, d\theta_s(t) \\ &+ \sum_{t \in T} \hat{q}(t, x(t^-), x(t^+)) \end{aligned} \tag{5.3}$$

where

$$\hat{q}(t, a, b) = \inf \left\{ \int_0^1 \hat{r}(t, y(s), \dot{y}(s), z(s))\, ds : y(0) = a,\ y(1) = b,\ y \in \mathcal{A}[0, 1] \text{ and } z \in L^1_m[0, 1] \right\}$$

and $x \in \mathcal{B}$, $\nu \in \mathcal{B}_m$ and

$$\begin{aligned} dx(t) &= \dot{x}(t)\, dt + \xi_s(t)\, d\theta_s(t) + \xi_a(t)\, d\theta_a(t), \\ d\nu(t) &= u(t)\, dt + \mu_s(t)\, d\theta_s(t) + \mu_a(t)\, d\theta_a(t) \end{aligned}$$

where, as before, the subscripts s and a denote, respectively, the smooth (nonatomic) and atomic parts of the singular measure. Let us also define

$$r(t, x, \xi) = \inf_{\mu} \hat{r}(t, x, \xi, \mu). \tag{5.4}$$

Assumption (U1) (inf-boundedness). For each fixed $t \in T$, $\alpha \in \mathbb{R}$ and bounded set $D \subset \mathbb{R}^n \times \mathbb{R}^n$, the set $\{u \in \mathbb{R}^m : (x, v) \in D \text{ with } K(t, x, v, u) \leqq \alpha\}$ is bounded.

Under this assumption $L(t, \cdot, \cdot)$ is lower semicontinuous and the r in (5.4) is the recession function for L (see [5]). In the following theorem, the assumptions on L and r apply to those functions as obtained above.

THEOREM 4. *Let assumptions* (B), (G) *and* (U1) *hold. Then if problem* (5.3) *has a feasible solution, it has an optimal solution.*

Proof. This result is shown by first proving that the problems (5.3) and (2.5) are in some sense equivalent, by a modified version of the Equivalence Theorem of [11]. Then we may use Theorem 3. Q.E.D.

Thus we can obtain existence results for the extended optimal control problem of (5.3), as well as the simpler versions that correspond to the calculus of variation problems in Corollaries 2 and 3, and those modified problems satisfying the conditions of Corollary 1.

As is true for the calculus of variations problem, the analogue of the functional I for optimal control has a much simpler form than (5.3).

Before we introduce this extension, we need to make the following definitions and assumptions. Define the multifunction $X : T \rightrightarrows \mathbb{R}^n$ by

$$X(t) = \{x : \exists (v, u) \text{ with } K(t, x, v, u) < \infty\}.$$

Using (5.2), we can re-express this as

$$X(t) = \{x: \exists v \text{ with } L(t, x, v) < \infty\}$$

which is the same as our earlier X. Define

$$\tilde{H}(t, x, p, w) = \sup_{v,u} \{p \cdot v + w \cdot u - K(t, x, v, u)\}.$$

Let

$$\tilde{F}(t, x) = \operatorname{dom} \tilde{H}(t, x, \cdot, \cdot) = \{(p, w): \tilde{H}(t, x, p, w) < \infty\}.$$

The following will imply assumption (A1), so that the recession function for the control problem will be independent of x.

Assumption (U2). There exists a multifunction $F: T \rightrightarrows \mathbb{R}^n$ such that for all $x \in X(t)$

$$\operatorname{cl} \tilde{F}(t, x) = F(t) \quad \text{for all } t \in T.$$

We can define the recession function $\hat{r}$ by

$$\hat{r}(t, \xi, \mu) = \sup \{p \cdot \xi + w \cdot \mu: (p, w) \in F(t)\}.$$

Assumption (U2). The extension of the functional in (2.2) to optimal control problems is then given by,

$$\begin{aligned} \text{minimize } \bar{\Psi}(x, v) = l(x(t_0), x(t_1)) &+ \int_T K(t, x(t), \dot{x}(t), u(t))\, dt \\ &+ \int_T \hat{r}(t, \xi(t), \mu(t))\, d\theta(t) \end{aligned} \tag{5.5}$$

where $x \in \mathcal{B}$, $v \in \mathcal{B}_m$ and

$$dx(t) = \dot{x}(t)\, dt + \xi(t)\, d\theta(t),$$
$$dv(t) = u(t)\, dt + \mu(t)\, d\theta(t)$$

and $\theta \in \mathcal{M}$.

Finally, we have the analogue of Theorem 4 for the simpler case.

THEOREM 4A. *Let assumptions* (A), (G), (U1) *and* (U2) *hold. Then if problem* (5.5) *has a feasible solution, it has an optimal solution.*

6. Discussion. One of the purposes in developing these existence theorems was to give foundation to the functionals Φ, I and Ψ and $\bar{\Psi}$ being valid extensions of the functionals J and $\tilde{\Psi}$ which only cover cases where $x \in \mathcal{A}$. We are not implying that they are the only extensions to problems where the states can jump and the controls can exhibit impulsive behaviour, but from Corollary 1 and the resulting discussion we can say with some justification that they form "upper bounds" for such extensions. If we look at it from an economic viewpoint, this should not be surprising. Thinking of the L and g terms of Corollary 1 as two producers, the difference being that the second producer g, can supply any amount of product instantaneously whereas the first needs some amount of time to do so, if g charges more than the limit that L charges when the period over which it supplies the product goes to zero, we would buy the product from L. But this limit of L is none other than the recession function r. So g must charge less than this limit to sell its goods, that is, for there to be any instantaneous change in the product x, we must have g less than or equal to r.

In many problems that one encounters in the economic literature where the states are allowed to have jumps, the term involving the nonatomic singular part of dx is

absent. However it is not difficult to construct a sequence of functions that are absolutely continuous and where the limit of the sequence is a continuous function of bounded variation with a nontrivial singular term. Therefore, such problems may not be well defined.

Acknowledgment. The author is deeply grateful to Professor R. T. Rockafellar for his guidance and assistance.

REFERENCES

[1] L. D. BERKOVITZ, *Optimal Control Theory*, Springer-Verlag, New York, 1974.
[2] O. CALIGARIS, F. FERRO AND P. OLIVA, *Sull'esistenza del minimo per problemi di calcolo delle variazioni relativi ad archi di variazione limitata*, Boll. Un. Mat. Ital., B(5) 14 (1977), pp. 340-369.
[3] C. W. CLARK, F. H. CLARKE AND G. R. MUNRO, *The optimal exploitation of renewable resource stocks: problems of irreversible investment*, Econometrica, 47 (1979), pp. 25-47.
[4] L. M. GRAVES, *The Theory of Functions of a Real Variable*, McGraw-Hill, New York, 1956.
[5] J. M. MURRAY, *On the proper extension of optimal control problems to admit impulses*, Doctoral dissertation, Univ. Washington, Seattle, 1983.
[6] L. W. NEUSTADT, *A general theory of minimum-fuel space trajectories*, SIAM J. Control, 3 (1965), pp. 317-356.
[7] R. W. RISHEL, *An extended Pontryagin principle for control systems whose control laws contain measures*, SIAM J. Control, 3 (1965), pp. 191-205.
[8] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton Univ. Press, Princeton, NJ, 1970.
[9] ———, *Integrals which are convex functionals, II*, Pacific J. Math., 39 (1971), pp. 439-469.
[10] ———, *Optimal arcs and the minimum value function in problems of Lagrange*, Trans. Amer. Math. Soc., 180 (1973), pp. 53-84.
[11] ———, *Existence theorems for general control problems of Bolza and Lagrange*, Advances in Math., 15 (1975), pp. 312-333.
[12] ———, *Integral functionals, normal integrands and measurable selections*, in Nonlinear Operators and the Calculus of Variations, L. Waelbroeck, ed., Springer-Verlag, Berlin, 1976, pp. 157-207.
[13] ———, *Dual problems of Lagrange for arcs of bounded variation*, in Calculus of Variations and Control Theory, D. L. Russell, ed., Academic Press, New York, 1976, pp. 155-192.
[14] ———, *Optimality conditions for convex control problems with nonnegative states and the possibility of jumps*, in Game Theory and Mathematical Economics, O. Moeschlin ed., North-Holland, New York, 1981, pp. 339-349.
[15] W. W. SCHMAEDEKE, *Optimal control theory for nonlinear vector differential equations containing measures*, SIAM J. Control, 3 (1965), pp. 231-280.
[16] J. WARGA, *Variational problems with unbounded controls*, SIAM J. Control, 3 (1966), pp. 424-438.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 3, May 1986

007

OPTIMAL CONTROL FOR VARIATIONAL INEQUALITIES*

AVNER FRIEDMAN†

Abstract. Consider the problem of maximizing a functional which depends on a control function k and on the solution of an elliptic variational inequality with k appearing in the data. The variational problem for k is nondifferentiable and nonconvex. We obtain necessary conditions on a maximizer k_0 and then use them to determine the structure of k_0 in some cases.

Key words. optimal control, variational inequality

AMS(MOS) subject classification. 93C20

Introduction. Consider an elliptic variational inequality with control k appearing either as an inhomogeneous term on the right-hand side or in the boundary data. Denote the corresponding solution by u and consider the functional

$$J(k)=\int F(x, u)\,dx+\int \Phi(k). \tag{0.1}$$

We are interested in studying the solutions k_0 of the optimization problem

$$J(k_0)=\max_{k\in\mathcal{A}} J(k), \qquad k_0\in\mathcal{A}, \tag{0.2}$$

where $\mathcal{A}$ is the class of controls.

This problem differs significantly for similar problems for solutions of linear or quasilinear elliptic (or parabolic) equations (as dealt with by Lions [10], [11] and the references given there) in that the functional $J(k)$ is generally nondifferentiable in the case of a variational inequality. Although some necessary conditions for k_0 have been derived by Barbu [1] and Mignot and Puel [12] (see also the references in [1], [12]), these conditions seem too implicit for extracting from them significant properties of the maximizer. In some specific elliptic and parabolic variational inequalities [4], [5], [7] the structures of the optimal controls have been determined; however the corresponding functionals $J(k)$ in these instances are actually differentiable.

In this paper we consider a large class of optimal control problems for elliptic variational inequalities with nondifferentiable $J(k)$; the main assumption regarding F is that $F_u\geqq 0$. In § 1 we shall derive an effective necessary condition for the maximizer, and in § 2 we use it to determine the structure of k_0 for a certain class $\mathcal{A}$. In § 3 we give several extensions.

1. A necessary condition. Let Ω be a bounded domain in $\mathbb{R}^n$ with C^2 boundary $\partial\Omega$ and let p be a positive number satisfying

$$p>\frac{n}{2}, \qquad p\geqq 2. \tag{1.1}$$

Let U^0 be a given function satisfying

$$U^0\in W^{2,p}(\Omega), \qquad U^0>0 \quad \text{in } \bar{\Omega}, \tag{1.2}$$

* Received by the editors September 18, 1984, and in revised form March 25, 1985. This work is partially supported by National Science Foundation under grant MCS-8300293.

† Department of Mathematics, Northwestern University, Evanston, Illinois 60201.

and let f be a given function satisfying

$$f \in L^p(\Omega). \tag{1.3}$$

We introduce a class $\mathscr{A}$ of control functions $k(x)$ with the properties

(1.4) $\mathscr{A}$ is a closed, bounded and convex subset of $L^p(\Omega)$.

For any $k \in \mathscr{A}$ consider the variational inequality

$$\int_\Omega \nabla u \cdot \nabla(\psi - u) \geqq -\int_\Omega (f+k)(\psi - u) \quad \forall \psi \in K,\ u \in K, \tag{1.5}$$

where

$$K = \{\psi \in H^1(\Omega),\ \psi \geqq 0 \text{ a.e.},\ \psi - U^0 \in H_0^1(\Omega)\}, \tag{1.6}$$

and the functional

$$J(k) = \int_\Omega F(x, u)\, dx + \int_\Omega \Phi(k)\, dx, \tag{1.7}$$

we assume that

(1.8) $\int_\Omega \Phi(k)$ is upper semicontinuous under weak convergence in $L^p(\Omega)$, i.e., if $k_m \to k$ weakly in $L^p(\Omega)$ where $k_m, k \in \mathscr{A}$ then

$$\int_\Omega \Phi(k) \geqq \limsup_{m\to\infty} \int_\Omega \Phi(k_m),$$

and that

(1.9) $F(x, u)$ is continuous in $\bar{\Omega} \times \mathbb{R}^1$.

Consider the problem

$$J(k_0) = \max_{k \in \mathscr{A}} J(k), \qquad k_0 \in \mathscr{A}. \tag{1.10}$$

By general theory for variational inequalities [3], [8] the solutions u of (1.5) belong to a bounded set of $W^{2,p}(\Omega)$ and, therefore, by (1.1) and Sobolev's imbedding,

$$|u|_{C^\alpha(\bar{\Omega})} \leqq C, \quad \text{where } \alpha = 2 - \frac{n}{p}, \qquad C < \infty. \tag{1.11}$$

Taking a maximizing sequence k_m and their corresponding solutions $u = u_m$ and using (1.8), (1.9) and (1.11), we find that, for a subsequence,

$$k_m \to k_0 \quad \text{weakly in } L^p(\Omega), \qquad k_0 \in \mathscr{A},$$
$$u_m \to u_0 \quad \text{in } C^\beta(\bar{\Omega}) \quad \text{for any } 0 < \beta < \alpha,$$

and k_0 is the solution of (1.10); u_0 is the solution of (1.5) corresponding to k_0.

In order to derive necessary conditions on k_0 we further assume that

$$\begin{aligned} &F_u(x, u) \text{ is continuous in } \bar{\Omega} \times \mathbb{R}^1, \\ &F_u(x, u) \geqq 0, \end{aligned} \tag{1.12}$$

and that $\Phi(k)$ has a derivative $\Phi'(k)$, that is,

(1.13) for any $k \in \mathscr{A}$ there is a function in $L^q(\Omega)$ $(1/q+1/p=1)$ denoted by $\Phi'(k)$ such that for any $l \in L^p(\Omega)$ with $k+\varepsilon l \in \mathscr{A}$ if $0<\varepsilon<1$ there holds

$$\frac{1}{\varepsilon}\int_\Omega [\Phi(k+\varepsilon l)-\Phi(k)]\,dx \to \int_\Omega \Phi'(k) l\,dx$$

if $\varepsilon \to 0$.

Let k_0 be a solution of (1.10) and denote by u_0 the corresponding solution of (1.5). Introduce the noncoincidence set

$$\Omega_0 = \{x \in \Omega,\ u_0(x) > 0\},$$

(which is an open set since u_0 is continuous in $\bar{\Omega}$) and let Q be the solution of the Dirichlet problem

$$\int_\Omega \nabla Q \cdot \nabla \psi = \int_\Omega F_u(x, u_0)\psi\,dx \quad \forall \psi \in H_0^1(\Omega_0), \quad Q \in H_0^1(\Omega_0). \tag{1.14}$$

By elliptic regularity $Q \in W^{2,p}(G)$ for any open set G with $\bar{G} \subset \Omega_0 \cup \partial\Omega$. Since the free boundary

$$\Gamma = \partial\Omega_0 \cap \Omega$$

is not regular, in general, one cannot assert that Q is continuous up to the free boundary Γ with $Q=0$ on Γ.

Taking $\psi = Q^-$ in (1.14) and using (1.12) we easily find that

$$\int_{\Omega_0} |\nabla Q^-|^2 = 0,$$

which implies (since $Q^- \in H_0^1(\Omega_0)$) that $Q^- = 0$, i.e., $Q \geqq 0$. By the strong maximum principle we further infer that

$$Q > 0 \quad \text{in } \Omega_0. \tag{1.15}$$

By comparison with the solution Q^1 of $\Delta Q^1 = -A$ in Ω, $Q^1 = 0$ on $\partial\Omega$ where $A > F_u(x, u_0)$, we find that $Q < Q^1$ in Ω_0, so that

$$Q \leqq C < \infty \quad \text{in } \Omega_0. \tag{1.16}$$

We can now state the main result of this section.

THEOREM 1.1. *Let k_0 be a solution of* (1.10) *and suppose that $k_0+\varepsilon l \in \mathscr{A}$ for all $0<\varepsilon<\varepsilon_0$, for some $\varepsilon_0>0$, and* supp $l \subset \Omega_0$. *Then*

$$\int_{\Omega_0} (Q-\Phi'(k_0))l\,dx \geqq 0. \tag{1.17}$$

Proof. Let U^1 be a nonnegative function in $W^{2,p}(\Omega)$ such that $U^1 = U^0$ in a neighborhood of $\partial\Omega$ and $U^1 = 0$ in a neighborhood of $\Omega \setminus \Omega_0$.

Denote by U_ε the solution of the Dirichlet problem

$$\begin{aligned} &\int_{\Omega_0} \nabla U_\varepsilon \cdot \nabla(\phi - U_\varepsilon) = -\int_{\Omega_0} (f+k_0+\varepsilon l)(\phi - U_\varepsilon) \quad \text{for any } \phi, \\ &\qquad \phi - U^1 \in H_0^1(\Omega_0), \quad U_\varepsilon - U^1 \in H_0^1(\Omega_0). \end{aligned} \tag{1.18}$$

Denote by u_ε the solution of the variational inequality (1.5) corresponding to $k_0+\varepsilon l$, i.e.,

$$\int_\Omega \nabla u_\varepsilon \cdot \nabla(\psi - u_\varepsilon) \geqq -\int_\Omega (f+k_0+\varepsilon l)(\psi - u_\varepsilon) \quad \text{for any } \psi \in K, \qquad u_\varepsilon \in K. \tag{1.19}$$

Taking $\psi = u_\varepsilon + (U_\varepsilon - u_\varepsilon)^+$ in (1.19) (where $U_\varepsilon = 0$ outside Ω_0) and $\phi = U_\varepsilon - (U_\varepsilon - u_\varepsilon)^+$ in (1.18) and adding, we obtain

$$-\int_{\Omega_0} \nabla(U_\varepsilon - u_\varepsilon)\cdot\nabla(U_\varepsilon - u_\varepsilon)^+ \geqq 0,$$

so that $\nabla(U_\varepsilon - u_\varepsilon)^+ = 0$; consequently

$$U_\varepsilon \leqq u_\varepsilon \quad \text{in } \Omega_0. \tag{1.20}$$

LEMMA 1.2. *There holds*

$$u_0 - U^1 \in H_0^1(\Omega_0), \tag{1.21}$$

and, for a subsequence $\varepsilon \to 0$,

$$\frac{U_\varepsilon - u_0}{\varepsilon} \to z \quad \textit{weakly in } H_0^1(\Omega_0), \tag{1.22}$$

where $z \in H_0^1(\Omega_0)$, *and* $\Delta z = l$ *in* Ω_0.

Proof. For any small $\delta > 0$, the function $(u_0 - \delta)^+ - (U^1 - \delta)^+$ has compact support in Ω_0 and hence belongs to $H_0^1(\Omega_0)$. Taking $\delta \to 0$ it follows that $u_0 - U^1$ is also in $H_0^1(\Omega_0)$.

To prove (1.22) notice that since $\Delta u_0 = f + k_0$ in Ω_0,

$$\int_{\Omega_0} \nabla u_0 \cdot \nabla \phi = -\int_{\Omega_0} (f+k_0)\phi,$$

for any $\phi \in C_0^1(\Omega_0)$ and, by approximation, also for any $\phi \in H_0^1(\Omega_0)$. Recalling (1.21) we then have

$$\int_{\Omega_0} \nabla u_0 \cdot \nabla(\psi - u_0) = -\int_{\Omega_0} (f+k_0)(\psi - u_0) \quad \text{if } \psi - U^1 \in H_0^1(\Omega_0).$$

Taking $\psi = U_\varepsilon$ and taking $\phi = u_0$ in (1.18) we obtain, by adding,

$$\int \Omega_0 |\nabla(u_0 - U_\varepsilon)|^2 \leqq \varepsilon \int_{\Omega_0} |l(u_0 - U_\varepsilon)| \leqq C\varepsilon \left\{ \int_{\Omega_0} |u_0 - U_\varepsilon|^2 \right\}^{1/2},$$

since $l \in L^p(\Omega)$, $p \geqq 2$. It follows that

$$\left\| \frac{u_0 - U_\varepsilon}{\varepsilon} \right\|_{H_0^1(\Omega_0)} \leqq C, \tag{1.23}$$

and the assertion (1.22) follows. Clearly $z \in H_0^1(\Omega_0)$. Also, since $\Delta(U_\varepsilon - u_0) = \varepsilon l$ in Ω_0, $\Delta z = l$ in Ω_0.

Having proved Lemma 1.2, we now proceed to establish the assertion (1.17). We begin with the inequality

$$J(k_0 + \varepsilon l) \leqq J(k_0),$$

or

$$\frac{1}{\varepsilon}\int_\Omega [F(x, u_\varepsilon) - F(x, u_0)] + \frac{1}{\varepsilon}\int_\Omega [\Phi(k_0 + \varepsilon l) - \Phi(k_0)] \leqq 0. \tag{1.24}$$

By the mean value theorem,

$$\frac{1}{\varepsilon}[F(x, u_\varepsilon) - F(x, u_0)] = F_u(x, \tilde{u}_\varepsilon(x))(u_\varepsilon - u_0)/\varepsilon,$$

where $\tilde{u}_\varepsilon(x)$ lies in the interval $(u_0(x), u_\varepsilon(x))$. Recalling (1.20) and the assumption $F_u(x, u) \geqq 0$, we obtain

$$\frac{1}{\varepsilon}[F(x, u_\varepsilon) - F(x, u_0)] \geqq F_u(x, \tilde{u}_\varepsilon(x))(U_\varepsilon - u_0)/\varepsilon. \tag{1.25}$$

Since $u_\varepsilon \to u_0$ uniformly in $\bar{\Omega}$, we also have

$$F_u(x, \tilde{u}_\varepsilon(x)) \to F_u(x, u_0(x)) \quad \text{uniformly with respect to } x \in \Omega. \tag{1.26}$$

Let Q_ε be the solution of the Dirichlet problem

$$\int_{\Omega_0} \nabla Q_\varepsilon \cdot \nabla \psi = \int_{\Omega_0} F_u(x, \tilde{u}_\varepsilon(x))\psi, \qquad \forall \psi \in H_0^1(\Omega_0), \quad Q_\varepsilon \in H_0^1(\Omega_0). \tag{1.27}$$

Taking $\psi = Q_\varepsilon - Q$ in (1.14) and $\psi = Q - Q_\varepsilon$ in (1.27) and adding we easily deduce, using (1.26), that

$$\|Q_\varepsilon - Q\|_{H_0^1(\Omega_0)} \to 0. \tag{1.28}$$

Substituting (1.25) into (1.24) and using (1.27) with $\psi = (U_\varepsilon - u_0)/\varepsilon$, we get

$$\int_{\Omega_0} \nabla Q_\varepsilon \cdot \nabla \frac{U_\varepsilon - u_0}{\varepsilon} + \frac{1}{\varepsilon} \int_\Omega (\Phi(k_0 + \varepsilon l) - \Phi(k_0)) \leqq 0.$$

Taking $\varepsilon \to 0$ and using (1.28), (1.22) and (1.13), we obtain

$$\int_{\Omega_0} \nabla Q \cdot \nabla z + \int_{\Omega_0} \Phi'(k_0) l \leqq 0. \tag{1.29}$$

Let $Q_j \in C_0^1(\Omega_0)$, $Q_j \to Q$ in $H_0^1(\Omega_0)$. Then

$$\int_{\Omega_0} \nabla Q_j \cdot \nabla z = -\int_{\Omega_0} Q_j \Delta z = -\int_{\Omega_0} Q_j l.$$

Taking $j \to \infty$ we obtain

$$\int_{\Omega_0} \nabla Q \cdot \nabla z = -\int_{\Omega_0} Q l.$$

Substituting this into (1.29), the assertion (1.17) follows.

Remark 1.1. Consider the case of a general obstacle g; that is, u is a solution of (1.5) with K replaced by

$$K_g = \{\psi \in H_0^1(\Omega), \psi \geqq g \text{ a.e.}, \psi - U^0 \in H_0^1(\Omega)\}. \tag{1.30}$$

Taking $\tilde{u} = u - g$, $\tilde{U}^0 = U^0 - g$ the problem is reduced to the one studied before with f replaced by $\tilde{f} = f - \Delta g$. Thus Theorem 1.1 can immediately be applied.

Remark 1.2. Theorem 1.1 can obviously be extended to general elliptic variational inequalities with variable coefficients (with K defined by (1.6) or (1.30)).

Remark 1.3. One can prove that, for a subsequence $\varepsilon \to 0$,

$$\frac{u_\varepsilon - u_0}{\varepsilon} \to \tilde{z} \text{ weakly in } H^1(\Omega_0),$$

but we cannot assert that $(u_\varepsilon - u_0)/\varepsilon \in H_0^1(\Omega_0)$ or even that $\tilde{z} \in H_0^1(\Omega_0)$ (unless the free

boundary is known to be smooth). It is for this reason that we have introduced the auxiliary solution U_ε; fortunately, replacing u_ε by U_ε does not change the inequality (1.24).

2. An application. Consider the case

$$\mathscr{A} = \left\{ 0 \leqq k(x) \leqq N, \int_\Omega k(x)\, dx = M \right\}, \tag{2.1}$$

where M, N are positive constants. We assume that

$$N \cdot \text{meas}\,(\Omega) > M \tag{2.2}$$

so that $\mathscr{A}$ is nontrivial. We take the functional

$$J(k) = \int_\Omega F(x, u)\, dx \tag{2.3}$$

and assume in addition to (1.9), (1.12) that

$$F_u(x, u) > 0 \quad \text{if } u > 0. \tag{2.4}$$

We also assume that (1.2) and (1.3) hold, as before.

Let k_0 be a solution of (1.10) and u_0 the solution of (1.5) and (1.6) corresponding to k_0. Set $\Omega_0 = \{x \in \Omega, u_0(x) > 0\}$ and define Q as in (1.14).

Notice that on any level set $\{Q = \lambda\}$ there holds $\Delta Q = 0$ a.e. and, therefore, also $F_u(x, u_0) = 0$. In view of (2.4) we conclude that

$$\text{meas}\,\{Q = \lambda\} = 0 \quad \text{for any } \lambda. \tag{2.5}$$

THEOREM 2.1. *For any maximizer k_0 there is a $\lambda \geqq 0$ such that*

$$\begin{aligned} &k_0 = N \text{ a.e. in the set } \Omega_0 \cap \{Q < \lambda\}, \\ &k_0 = 0 \text{ a.e. in the set } \Omega_0 \cap \{Q > \lambda\}. \end{aligned} \tag{2.6}$$

Proof. We first claim that

$$\text{meas}\,\{x \in \Omega_0, 0 < k_0(x) < N\} = 0. \tag{2.7}$$

Indeed, otherwise there exists a Lebesgue point x_0 of k_0, $x_0 \in \Omega_0$, such that

$$2\delta < k(x_0) < N - 2\delta, \qquad 0 < 4\delta < N.$$

Hence for any $\varepsilon > 0$ there is a set G of positive measure in the ball $B_\varepsilon(x_0) = \{|x - x_0| < \varepsilon\}$ such that

$$\delta < k(x) < N - 2\delta \quad \text{in } G.$$

Let $l(x)$ be any bounded function with support in G such that $\int l(x) = 0$. Then $k_0 \pm \varepsilon l \in \mathscr{A}$ if ε is small enough, and Theorem 1.1 gives

$$\int_G Ql = 0.$$

It follows that $Q = \text{const}$ in G, which contradicts (2.5).

From (2.7) we see that

$$k_0 = N\chi_A \quad \text{in } \Omega_0$$

where χ_A denotes the characteristic function of a set A. Let x_0 be any Lebesgue point

of A and of $k(x)$, and let y_0 be any Lebesgue point of $\Omega_0 \backslash A$ and of $k(x)$. Then we have

$$Q(x_0) \leqq Q(y_0). \tag{2.8}$$

Indeed, otherwise there is a $\lambda_0 > 0$ such that

$$k_0(x) = N, \quad k_0(y) = 0, \quad Q(x) > \lambda_0, \quad Q(y) < \lambda_0 \tag{2.9}$$

for $x \in G$, $y \in G'$ where $G \subset B_\varepsilon(x_0)$, $G' \subset B_\varepsilon(y_0)$ for some small $\varepsilon > 0$ and meas $(G) =$ meas $(G') > 0$. Take $l = -1$ in G, $l = 1$ in G' and $l = 0$ elsewhere. Then $k_0 + \varepsilon l \in \mathscr{A}$ if $\varepsilon > 0$, ε small, and Theorem 1.1 gives

$$\int_G Ql + \int_{G'} Ql \geqq 0$$

or $\int_{G'} Q \geqq \int_G Q$, a contradiction to (2.9).

Having proved (2.8) we conclude that $Q(x) < Q(y)$ a.e. for $x \in A$, $y \in \Omega_0 \backslash A$. This yields the assertion of the lemma for some $\lambda \geqq 0$.

THEOREM 2.2. *If*

$$\text{meas } \{x \in \Omega_0, k_0(x) > 0\} > 0 \tag{2.10}$$

then

$$k_0 = N \quad \text{a.e. in } \Omega \backslash \Omega_0. \tag{2.11}$$

Theorems 2.1 and 2.2 exhibit the bang-bang nature of the optimal controls.

Proof. Suppose (2.11) is not true. Then for any small $\varepsilon > 0$ one can find a set $G \subset (\Omega \backslash \Omega_0)$ with small positive measure such that $k(x) < N - \delta$ in G.

By (2.10) there exists a set G' in Ω_0 such that $k(x) > \delta$ in G' and meas $G' =$ meas G. Define $l = 1$ in G, $l = -1$ in G' and $l = 0$ elsewhere. Then $k_0 + \varepsilon l \in \mathscr{A}$ if $\varepsilon < \delta$. Denoting the corresponding solution of (1.5) by u_ε, we claim that

$$u_\varepsilon \geqq u_0, \qquad u_\varepsilon \not\equiv u_0. \tag{2.12}$$

Indeed, we substitute $\psi = u_0 - (u_0 - u_\varepsilon)^+$ in the variational inequality for u_0 and $\psi = u_\varepsilon + (u_0 - u_\varepsilon)^+$ in the variational inequality for u_ε. Adding the two inequalities we get

$$\int_\Omega |\nabla (u_0 - u_\varepsilon)^+|^2 \leqq \varepsilon \int_\Omega l(u_0 - u_\varepsilon)^+ = -\varepsilon \int_{G'} (u_0 - u_\varepsilon)^+ \leqq 0.$$

Hence $(u_0 - u_\varepsilon)^+ = 0$ and, since $\Delta u \not\equiv \Delta u_\varepsilon$ on G', (2.12) follows.

From (2.12) we conclude that $J(k_0 + \varepsilon l) > J(k_0)$, a contradiction.

Remark 2.1. If $f \leqq 0$ then the condition (2.10) is satisfied. Indeed, otherwise we have $\Delta u_0 = f \leqq 0$ in Ω_0. Take a ball B in Ω_0 with $\partial B \cap \Gamma = \{x_0\}$ (Γ is the free boundary). Since $u_0 > 0$ in B and $u_0(x_0) = 0$, the maximum principle gives $\nabla u(x_0) \neq 0$, a contradiction.

THEOREM 2.3. *If $f \geqq -N$ then, for any maximizer k_0,*

$$\text{meas } \{x \in \Omega_0, k_0(x) = N\} < \text{meas } \Omega_0. \tag{2.13}$$

Proof. Suppose (2.13) is not true. Then

$$-\Delta u_0 = -f - N \quad \text{in } \Omega_0.$$

Since also $-\Delta u_0 = 0 \geqq -f - N$ a.e. in $\Omega \backslash \Omega_0$, it follows that

$$-\Delta u_0 \geqq -f - N \quad \text{a.e. in } \Omega.$$

Also

$$u_0 \geqq 0, \quad u_0(-\Delta u_0 + f + N) = 0 \quad \text{a.e. in } \Omega.$$

Since $-(f+k) \geqq -(f+N)$ for any $k \in \mathscr{A}$, we then have, by comparison of u_0 with the solution u of (1.5), that $u \geqq u_0$. Since $u > 0$ (and, therefore, $\Delta u = f + k$) in some Ω-neighborhood $\tilde{\Omega}$ of $\partial\Omega$, if we choose k with $k \not\equiv N$ in $\tilde{\Omega}$ we then have $u \not\equiv u_0$ so that, by (2.4), $J(k) > J(k_0)$, a contradiction.

Remark 2.2. For any component T in Ω_0 of $\{Q < \lambda\}$ there holds

$$\partial T \cap \partial\Omega_0 \neq \varnothing. \tag{2.14}$$

Indeed, otherwise $Q = \lambda$ on ∂T and, since $-\Delta Q = F_u(x, u_0) > 0$ in T, the maximum principle gives $Q > \lambda$ in T, a contradiction. One of the components of $\{Q < \lambda\}$ must contain an Ω-neighborhood of $\partial\Omega$. If the free boundary $\partial\Omega_0 \cap \Omega$ consists of a finite number m of smooth surface Γ_i then the set $\{Q < \lambda\}$ has at most $m+1$ components: for each component T we have that either $\partial T \supset \partial\Omega$ or $\partial T \supset \Gamma_i$ for some i.

Remark 2.3. In case Ω is a one-dimensional interval $\{0 < x < a\}$ and $f \geqq 0$, the coincidence set consists of a single interval (which may be empty) and $k_0 = N$ on precisely three intervals: $0 \leqq x \leqq \gamma_1$, $\gamma_2 \leqq x \leqq \gamma_3$, $\gamma_4 \leqq x \leqq a$ (provided the coincidence set consists of a nonzero interval). When $f = 0$ and $F(x, u) \equiv u$ the γ_i can be computed explicitly; this special case was analyzed by Yaniro [13] by a method which is strictly one-dimensional.

Remark 2.4. The results of §§ 1 and 2 extend to the case where $U^0 \geqq 0$ on $\partial\Omega$.

3. Various extensions. In this section we give additional applications of the methods of §§ 1 and 2.

We begin by considering the same problem (1.10) associated with (1.5) and (1.6) where $\mathscr{A}$ is given by

$$\mathscr{A} = \left\{ k \in L^p(\Omega),\ k \geqq 0 \text{ a.e.},\ \int_\Omega k^p \leqq M \right\}, \tag{3.1}$$

and

$$J(k) = \int_\Omega F(x, u)\, dx + \int_\Omega \Phi(k); \tag{3.2}$$

here p, F and Φ are as in § 1. For any maximizer k_0 we introduce the corresponding u_0, Ω_0 and Q as before.

THEOREM 3.1. *For any maximizer k_0 there exists a real number μ such that*

$$Q - \Phi'(k_0) = \mu k_0^{p-1} \quad \textit{a.e. in } \Omega_0 \cap \{k_0 > 0\}, \tag{3.3}$$

$$Q - \Phi'(k_0) \geqq 0 \quad \textit{a.e. in } \Omega_0 \cap \{k_0 = 0\}. \tag{3.4}$$

Proof. To prove (3.3) we may assume that $\Omega_0 \cap \{k_0 > 0\}$ has positive measure. Take any $\delta > 0$ such that the set $A_\delta = \{x \in \Omega_0,\ k_0(x) > \delta\}$ has positive measure and let l be any bounded measurable function with support in A_δ such that

$$\int k_0^{p-1} l = 0. \tag{3.5}$$

Let $\tilde{k}_0 = k_0$ if $x \in A_\delta$, $\tilde{k}_0 = 0$ if $x \in \Omega \setminus A_\delta$. The function

$$k_\varepsilon = k_0 + \varepsilon((\pm l) - \gamma \tilde{k}_0), \qquad \gamma > 0$$

is nonnegative if ε is small enough and

$$\int k_\varepsilon^p = \int k_0^p + \varepsilon p \int k_0^{p-1}((\pm l) - \gamma \tilde{k}_0) + O(\varepsilon^2)$$

$$\leqq M - \varepsilon p \gamma \int_{A_\delta} k_0^p + O(\varepsilon^2) < M,$$

if ε is small enough. Thus $k_\varepsilon \in \mathscr{A}$ and Theorem 1.1 gives

$$\int_{A_\delta} (Q - \Phi'(k_0))((\pm l) - \gamma \tilde{k}_0) \geqq 0.$$

Taking $\gamma \to 0$ we get

$$\int_{A_\delta} (Q - \Phi'(k_0)) l = 0. \tag{3.6}$$

We have thus proved that (3.5) implies (3.6) when l is any bounded measurable function with support in A_δ. Consequently,

$$Q - \Phi'(k_0) = \mu k_0^{p-1} \quad \text{in } A_\delta,$$

where μ is some real number. Since obviously μ is independent of δ, the assertion (3.3) follows.

To prove (3.4) let S be any compact subset of Ω_0 contained in $\{k_0(x) = 0\}$ and let l_0 be a bounded measurable function with support in $S \cup A_\delta$ (for some small $\delta > 0$) satisfying

$$l_0 \geqq 0 \quad \text{on } S, \qquad l_0 = -1 \quad \text{on } A_\delta.$$

Let $l = Dl_0$ in S, $l = l_0$ elsewhere. Then, for any large $D > 0$, the function $k_0 + \varepsilon l$ is in $\mathscr{A}$ if ε is small enough. Applying Theorem 1.1, we get

$$D \int_S (Q - \Phi'(k_0)) l_0 \geqq \int_{A_\delta} (Q - \Phi'(k_0)).$$

Dividing by D and letting $D \to \infty$, we find that

$$\int_S (Q - \Phi'(k_0)) l_0 \geqq 0.$$

Since l_0 and S are arbitrary, (3.4) follows.

Remark 3.1. If $\Phi(k) \equiv 0$ then the proof of Theorem 2.2 shows that supp k_0 is contained in the coincidence set of u_0, provided the latter has positive measure.

We next study problems with control on the boundary. Consider the variational inequality

$$\begin{aligned} &\int_\Omega \nabla u \cdot \nabla(\psi - u) \geqq -\int_\Omega f(\psi - u) \quad \text{if } \psi - u \in H_0^1(\Omega), \quad \psi \geqq 0 \quad \text{a.e.,} \\ &u \in H^1(\Omega), \quad u \geqq 0 \quad \text{a.e.,} \quad u = U^0 + k \quad \text{on } \partial\Omega \text{ in the sense of trace class,} \end{aligned} \tag{3.7}$$

where U^0 is as in § 1 and k is the control function varying in the class, say,

$$\mathscr{B} = \left\{ k \in L^\infty(\partial\Omega),\ 0 \leqq k \leqq N,\ \int_{\partial\Omega} k = M \right\}.$$

We associate with (3.7) a functional, say,

$$J(k)=\int_\Omega F(x,u)\,dx+\int_{\partial\Omega}\Phi(k)\,dS,$$

where F and Φ are as in §§ 1 and 2 (with Ω replaced by $\partial\Omega$ in (1.8) and (1.13)).

Let k_0 be a maximizer, i.e.,

$$J(k_0)=\max_{k\in\mathcal{B}} J(k),\qquad k_0\in\mathcal{B}, \tag{3.8}$$

and denote by u_0 the corresponding solution of (3.7). Set $\Omega_0=\{x\in\Omega,\ u_0(x)>0\}$ and define Q, as before, by (1.14).

THEOREM 3.2. *Let k_0 be a solution of* (3.8) *and suppose $k_0+\varepsilon l\in\mathcal{B}$ for all $0<\varepsilon<\varepsilon_0$, for some $\varepsilon_0>0$. Then*

$$\int_{\partial\Omega}\left(\frac{\partial Q}{\partial\nu}-\Phi'(k_0)\right)l\geqq 0, \tag{3.9}$$

where ν is the outward normal to $\partial\Omega$.

Proof. Denote by u_ε the solution of (3.7) corresponding to $k_0+\varepsilon l$ and let $\tilde{U}_\varepsilon$ be a function in $H^1(\Omega)$ such that $\tilde{U}_\varepsilon=u_0$ in a neighborhood of $\Omega\setminus\Omega_0$, $\tilde{U}_\varepsilon=u_0+\varepsilon l$ on $\partial\Omega$. Denote by U_ε the solution of

$$\Delta U_\varepsilon=f\quad\text{in }\Omega_0,$$
$$U_\varepsilon-\tilde{U}_\varepsilon\in H_0^1(\Omega_0).$$

Then we again have that $u_\varepsilon\geqq U_\varepsilon$. Proceeding as in the proof of Theorem 1.1 we can show that

$$\frac{U_\varepsilon-u_0}{\varepsilon}\to z\quad\text{weakly in } H^1(\Omega_0),$$
$$\Delta z=0\quad\text{in }\Omega_0,$$
$$z-U^*\in H_0^1(\Omega_0),$$

where U^* is a function in $H^1(\Omega)$ such that $U^*=l$ on $\partial\Omega$ (in the sense of trace class) and $U^*=0$ in a neighborhood of $\partial\Omega_0\cap\Omega$. Further

$$-\int_{\partial\Omega_0\cap\partial\Omega}\frac{\partial Q}{\partial\nu}z+\int_\Omega\nabla z\cdot\nabla Q+\int_{\partial\Omega}\Phi'(k_0)l\leqq 0.$$

Since the middle integral on the left-hand side vanishes, (3.9) follows.

Remark 3.2. From Theorem 3.2 we can easily deduce that there exists a λ such that

$$k_0=N\text{ on }\left\{x\in\partial\Omega;\ \frac{\partial Q}{\partial\nu}-\Phi'(k_0)>\lambda\right\},$$
$$k_0=0\text{ on }\left\{x\in\partial\Omega;\ \frac{\partial Q}{\partial\nu}-\Phi'(k_0)<\lambda\right\}.$$

Remark 3.3. Theorem 3.2 extends to the case where $U^0\geqq 0$ on $\partial\Omega$.

Remark 3.4. There are several difficulties in trying to extend the method of this paper to parabolic variational inequalities: (i) although the existence of a solution U_ε to the parabolic analogue of (1.18) is known, uniqueness does not seem to be known (see [9, p. 63]) since no a priori regularity properties of the free boundary are assumed;

thus the inequality $U_\varepsilon \leqq u_\varepsilon$ is not obvious; (ii) the same problem occurs for Q_ε, although in the case of increasing sets $S(t) \equiv \{x;\ u(x, t) > 0\}$ uniqueness is known; see [9, p. 63]; (iii) the regularity of U_ε and Q in t does not seem to be sufficient to justify the integration by parts steps needed to extend the proof of Theorem 1.1. In what follows we briefly give one example in which the free boundary is smooth so that our method can be applied.

Consider the one-phase exterior Stefan problem with inner core S for the water region. We impose the boundary condition

$$\frac{\partial \theta}{\partial \nu} = \gamma + k(t) \quad \text{on } S \qquad \theta = \text{temperature}, \tag{3.10}$$

where γ is a positive constant and k varies in the class

$$\mathcal{B}_0 = \left\{ o \leqq k(t) \leqq N, \int_0^T k(t) \leqq M \right\}.$$

It is well known that the problem for θ can be formulated as a variational inequality for u ($u_t = \theta$ and $u = 0$ on the free boundary); see [3], [8].

We assume that

$$S \text{ is a ball}, \qquad N < \gamma. \tag{3.11}$$

Introducing polar coordinates (r, φ) about the center of S (φ is $(n-1)$-dimensional), we have that $\partial u/\partial \varphi = 0$ on $S \times (0, T)$. From this we deduce, by a standard argument that $|\partial u/\partial \varphi| \leqq C$ in the entire domain $\Omega \times (0, T)$. Recalling (3.10) and (3.11), we deduce that

$$\frac{\partial u}{\partial l} > 0 \quad \text{on } S,$$

for any direction l close enough to the radial direction $-r$. This enables us to deduce, as in [6], that the free boundary is Lipschitz continuous and, by [2], it is C^1 (or, in fact, C^∞).

We are now in a position to be able to study maximization or minimization problems. Consider the problem of maximizing the volume of ice which has melted by time T. This translates into the problem of minimizing

$$J_0(k) = \int u_t(x, T)\, dx. \tag{3.12}$$

Since this functional is not very regular, we replace it by an average

$$\frac{1}{T - T_0} \int_{T_0}^{T} \int_\Omega u_t(x, t)\, dx\, dt, \qquad 0 < T_0 < T,$$

or by

$$J(k) = \int_\Omega u(x, T)\, dx - \int_\Omega u(x, T_0)\, dx. \tag{3.13}$$

We also replace $\mathcal{B}_0$ by

$$\mathcal{B} = \{k \in \mathcal{B}_0 \text{ and } k(t) = 0 \text{ if } t > T_0\}. \tag{3.14}$$

Denote by Q^s the solution of

$$\begin{aligned} &Q_t+\Delta Q=0 && \text{in } \{u_0>0\}\cap\{t<s\}\equiv\Omega_0^s,\\ &Q=1 && \text{on } \{t=s\},\\ &\frac{\partial Q}{\partial \nu}=0 && \text{if } x\in S,\quad 0<t<s,\\ &Q=0 && \text{on the free boundary in } \{0<t<s\}. \end{aligned}$$

Proceeding as in § 1, we find that

$$\int_0^{T_0}\int_S (Q^T-Q^{T_0})L(\tau)\,dS\,d\tau\geqq 0,$$

where

$$\frac{\partial z}{\partial \nu}=L(\tau)=\int_0^\tau l(\sigma)\,d\sigma \quad \text{on } S\times(0,T_0).$$

Hence the function

$$W(\tau)=\int_\tau^T\int_S Q^T-\int_\tau^{T_0}\int_S Q^{T_0}$$

satisfies

$$\int_0^{T_0} W(\tau)l(\tau)\geqq 0.$$

Since

$$W_\tau=-\int_S (Q^T-Q^{T_0})$$

and $Q^T<Q^{T_0}$ by the maximum principle, we find, as in the proof of Theorem 1.1, that

$$k_0(t)=\begin{cases} N, & \text{if } 0<t<t_0,\\ 0, & \text{if } t_0<t<T. \end{cases}$$

For $n=1$ the same result was obtained in [5] for the original functional (3.12); the restriction $N<\gamma$ is not needed in this case.

REFERENCES

[1] V. BARBU, *Optimal Control of Variational Inequalities*, Pitman, London, 1984.

[2] L. A. CAFFARELLI, *Regularity of free boundaries in higher dimensions*, Acta Math., 139 (1977), pp. 155-184.

[3] A. FRIEDMAN, *Variational Principles and Free Boundary Problems*, John Wiley, New York, 1982.

[4] ———, *Nonlinear optimal control for parabolic equations*, this Journal, 22 (1982), pp. 805-816.

[5] A. FRIEDMAN AND L. JIANG, *Nonlinear optimal control problems in heat conduction*, this Journal, 21 (1983), pp. 940-952.

[6] A. FRIEDMAN AND D. KINDERLEHRER, *A one-phase Stefan problem*, Indiana Univ. Math. J., 24 (1975), pp. 1005-1035.

[7] A. FRIEDMAN AND D. YANIRO, *Optimal control for the dam problem*, Optim. Appl. Math., 13 (1985), pp. 59-78.

[8] D. KINDERLEHRER AND G. STAMPACCHIA, *An Introduction to Variational Inequalities and Their Applications*, Academic Press, New York, 1980.

[9] J. L. LIONS, *Equations differentielles operationnelles et problèmes aux limites*, Springer-Verlag, Berlin, 1961.

[10] ———, *Sur le contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles*, Dunod, Paris, 1968.

[11] ———, *Some Aspects of the Optimal Control of Distribution Parameter Systems*, CBMS Regional Conference Series in Applied Mathematics 6, Society for Industrial and Applied Mathematics, Philadelphia, 1972.

[12] F. MIGNOT AND J. P. PUEL, *Optimal control in some variational inequalities*, this Journal, 22 (1984), pp. 466-476.

[13] D. YANIRO, *Optimal control problems for variational inequalities*, thesis, Northwestern Univ., Evanston, IL, August, 1984.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 3, May 1986

008

ON THE EFFICIENCY AND OPTIMALITY OF ALLOCATIONS II*

NIKOLAOS S. PAPAGEORGIOU†

Abstract. In this work we study the problem of characterizing efficient, price efficient and optimal allocations in a very general economic model with a continuum of agents and an infinite-dimensional commodity space. We do that using the theory of concave normal integrands and Clarke's theory of generalized gradients, as well as certain results from geometric functional analysis. We also consider approximations of those notions and study their properties using the theory of ε-subdifferentiation and Ekeland's variational principal. Then we examine those concepts in the context of a particular sector of the enonomy using single valued and multivalued conditional expectations and martingale theory. Finally we study stability questions using the Kuratowski-Mosco convergence of sets and the epi-convergence (τ-convergence) of closed functions.

Key words. normal integrand, generalized gradient, measurable multifunction, conditional expectation, subdifferential, ε-subdifferential, Kuratowski-Mosco convergence

AMS(MOS) subject classifications. primary 90A, 49E; secondary 46

1. Introduction. A central problem in economics is the characterization of efficient programs of resource allocation by a price system and the use of a price mechanism to attain such an allocation in economies where decision making is decentralized. Previous work in this area was done by Majumdar [23], [24], Peleg [31]–[34], Peleg-Yaari [35] and Radner [37]. However, their models were quite restrictive (one agent-discrete time models) and their assumptions were quite strict from a mathematical viewpoint because they did not address the problem in the framework of nonsmooth analysis. Also their work, when it was infinite-dimensional, was concentrated on the dual pair (commodity space-price space) (l^∞, l^1). Our work aims at overcoming the limitations of those papers, providing a more general and natural framework for addressing those issues and presenting several new results that, even when we put them in the context of the earlier models, are novel. Furthermore in this work, we are the first to introduce and study approximate efficiency and optimality, which provide a strong insight into the properties of complete efficiency and optimality. Also, since our model corresponds to a multisector economy, we study what happens when we restrict our attention to a specific sector and finally—and in this respect we believe that our work is pioneering—we examine the stability of all the concepts that we introduce, under perturbations of the data on which they depend.

2. Background material and basic notions. We start with a description of our model. We postpone the precise mathematical definitions until later in this section.

Our model is a very general and flexible one and so it admits several economic interpretations. We will briefly outline all of them. The mathematical objects that constitute our model are the following: A (nonatomic) finite measure space (Ω, Σ, μ), a separable Banach space X ordered by a nonempty, closed, convex, pointed cone X_+, a measurable multifunction $F: \Omega \to 2^X$ and a function $u: \Omega \times X \to \bar{\mathbb{R}}$. Now we will give their economic interpretation.

If we view our economy as a producer's economy, (Ω, Σ, μ) is the measure space of agents (producers); Ω is the set of agents, Σ is the family of all possible coalitions and $\mu(\cdot)$ gives the relative size of the coalitions. The idea of a continuum of agents

* Received by the editors April 1, 1984 and in revised form March 15, 1985. This research was supported by National Science Foundation grant DMS-8403135.

† Department of Mathematics, University of Illinois, Urbana, Illinois 61801.

was first introduced by Aumann [1] as a device to capture the spirit of pure competition; in other words the economic situation where each individual agent cannot alone influence the outcome of the collective activity, but can do so through coalitions. We put the word nonatomic in parenthesis because it is not always necessary. When the measure space is atomic we interpret the atoms as large traders who constitute the oligopolistic participants of the economic process. The space X is the commodity space. The assumption that commodities are not finite in number agrees with many classical situations for economic theory: intertemporal equilibrium with an infinite horizon, a world of uncertainty with an infinite number of states or differentiation of commodities. The multifunction $F(\cdot)$ describes the production feasibility set for each producer $\omega \in \Omega$. So we call $F(\cdot)$ the production multifunction. The function $u(\omega, \cdot)$ specifies the profit function for each producer $\omega \in \Omega$. Finally a function $p(\cdot) \in [L_X^1(\Omega)]_+^*$ will be called a "price system" and then $p(\omega)$ represents the price established by agent $\omega \in \Omega$ for his production. From the viewpoint of optimization theory prices can be identified with "Lagrange multipliers" or "dual variables".

We can also view the mathematical model as representing a consumer's economy with only public goods. Then (Ω, Σ, μ) is the measure space of consumers, X is again the commodity space, $F: \Omega \to 2^X$ is the consumption multifunction, i.e., for each agent $\omega \in \Omega$, $F(\omega)$ represents the set of his feasible consumption bundles and $u(\omega, \cdot)$ gives the utility function of agent $\omega \in \Omega$. Finally $p(\cdot) \in [L_X^1(\Omega)]_+^*$ assigns a different price (tax rate) for each consumer.

We can also have a probabilistic interpretation of our model. Namely, think of a world of uncertainty with only one agent. Then (Ω, Σ, μ) is the probability space of all possible outcomes, $F(\omega)$ is the agent's consumption set when the state of affairs is ω, $u(\omega, \cdot)$ is his utility function when the outcome is $\omega \in \Omega$ and finally $p(\omega)$ is the price system that he faces when the state of nature is ω.

Finally we can give a "dynamic" interpretation to our mathematical model. In that case we have that $\Omega = T$ is a closed interval in $\mathbb{R}_+$, Σ is the Lebesgue subsets of T and $\mu = \lambda$ is the Lebesgue measure. Think of $t \in T$ as a time variable. Then $F(t)$ is the set of all feasible consumption plans for the community at time t, $u(t, \cdot)$ is the utility function at that same instant of time and finally $p(t)$ is the price system prevailing in the market at time $t \in T$.

In general $F(\cdot)$ will be a closed valued, integrably bounded multifunction. In that case the set S_F^1 of feasible resource allocations is nonempty. For the function $u: \Omega \times X \to \bar{\mathbb{R}}$ we will assume initially that it is a concave, normal integrand (in the sense of Rockafellar [40]) which is monotone increasing. So we will rely heavily on Rockafellar's powerful theory of normal integrands. Then we generalize to functions which are not necessarily concave in the x variable, but they are locally Lipschitz. Here our main tool will be Clarke's theory of generalized gradients [7]–[9].

The three basic concepts that are central in this work are the following:

(1) We will say that $f(\cdot) \in L_X^1(\Omega)$ is an efficient allocation if and only if $f \in S_F^1$ and $(f(\omega) + \dot{X}_+) \cap F(\omega) = \varnothing$ μ-a.e. where $\dot{X}_+ = X_+ \backslash \{\varnothing\}$.

So efficient allocations are those under which no nonnull coalition can find another feasible redistribution of resources so as to improve the well being of all its members and at the same time keep the welfare status of the rest of the society at the same level. Viewed in another way an efficient allocation is one which to μ-almost all agents assigns a commodity bundle which is maximal within their feasibility set for the partial ordering on X induced by X_+.

(2) We will say that $f(\cdot) \in L_X^1(\Omega)$ is an optimal allocation if and only if $f \in S_F^1$ and $u(\omega, g(\omega)) \leqq u(\omega, f(\omega))\mu$-a.e. for all $g \in S_F^1$.

So an optimal allocation is one which maximizes the profit (or utility) of μ-almost all agents among all other feasible allocations.

We know that prices suggest how economic decisions can be decentralized. But in addition to that the study of price systems associated with efficient allocations (programs) can be used to provide qualitative information about such programs (like existence results, topological properties of the set of efficient allocations etc.) So it is natural to make the following definition.

(3) If $p(\cdot) \in [L_X^1(\Omega)]_+^* \setminus \{0\}$ we will say that $f(\cdot) \in L_X^1(\Omega)$ is a "$p(\cdot)$-efficient" (or "price efficient for $p(\cdot)$") allocation if and only if $\langle p, f \rangle = \sigma_{S_F^1}(p)$ where $\sigma_{S_F^1}(\cdot)$ is the support function of the set S_F^1 and $\langle \cdot, \cdot \rangle$ denotes the duality brackets between $L_X^1(\Omega)$ and $[L_X^1(\Omega)]^*$.

In [28] the author obtained some first relations between the above three concepts. Here we will extend those results and prove new ones.

Sometimes mathematical economists prefer to have resource allocation functions that belong in $L_X^\infty(\Omega)$. We therefore conduct a parallel investigation for this case. The main technical difficulty that we face in this case is that $[L_X^\infty(\Omega)]^*$ is not $L_{X^*}^1(\Omega)$ but a much larger Banach space. If we attempt to define price systems as elements of $[L_X^\infty(\Omega)]_+^*$, we cannot have a satisfactory economic interpretation for them. So we are forced to confine ourselves to price systems in $L_{X^*}^1(\Omega) \subsetneq [L_X^\infty(\Omega)]^*$. However this calls for trouble, because as our work in [28] suggests, the arguments that guarantee the existence of a value maximizing price system are classical duality arguments (separation theorems) which produce elements in $[L_X^\infty(\Omega)]^*$ which may not be in $L_{X^*}^1(\Omega)$. To overcome this difficulty we will use Levin's generalization of the Yosida–Hewitt theorem [22] and the work of Rockafellar [41], [42], which examined normal integral functionals within the dual system $(L_X^\infty(\Omega), L_{X^*}^1(\Omega))$.

When int $X_+ \neq \varnothing$ we can have a weaker notion of efficiency.

(4) We say that $f(\cdot) \in L_X^1(\Omega)$ is a "weakly efficient allocation" if and only if $f \in S_F^1$ and $(f(\omega) + \text{int } X_+) \cap F(\omega) = \varnothing$ μ-a.e.

Intuitively we can think of weakly efficient allocations as those allocations for which no nonnegligible coalition can find an alternative feasible redistribution of goods that will make all of its members strictly better without affecting the well being of the rest of the society. Clearly every efficient allocation is weakly efficient, but the converse is not true in general.

In certain cases it is quite difficult to determine efficient and optimal allocations. We want then to be able to get arbitrarily close to such allocations. This leads to the following two notions.

(1a) Given $\varepsilon(\cdot) \in [L^1(\Omega)]_+$ we say that $f(\cdot) \in L_X^1(\Omega)$ is "$\varepsilon(\cdot)$-optimal" if and only if $f \in S_F^1$ and $u(\omega, g(\omega)) - \varepsilon(\omega) \leqq u(\omega, f(\omega))$ μ-a.e. for all $g \in S_F^1$.

(2a) If $p(\cdot) \in [L_X^1(\Omega)]_+^* \setminus \{0\}$ and $\varepsilon > 0$, then $f(\cdot) \in L_X^1(\Omega)$ is "ε-$p(\cdot)$-efficient" if and only if $f \in S_F^1$ and $\langle p, f \rangle \geqq \sigma_{S_F^1}(p) - \varepsilon$.

If we are within the producer's or consumer's model and we want to examine only a sector of the economy, we model that by passing to a sub-σ-field Σ_0 of Σ. Then the profit (resp. utility) function is given by $E^{\Sigma_0} u(\cdot, \cdot)$ and the production (resp. consumption) multifunction by $E^{\Sigma_0} F(\cdot)$. In the context of the probabilistic model this means that we have only partial information about the underlying uncertainty.

Now we will pass to the mathematical definitions and we will recall very briefly some basic facts from the theory of measurable multifunctions. For details the reader can refer to Castaing–Valadier [6], Himmelberg [14] and Rockafellar [39], [40].

So let $F: \Omega \to 2^X \setminus \{\varnothing\}$ be a multivalued function (multifunction). We introduce the set $\text{Gr } F = \{(\omega, x) \in \Omega \times X: x \in F(\omega)\}$ which we will call the graph of $F(\cdot)$. If $V \subseteq X$, we define $F^-(V) = \{\omega \in \Omega: F(\omega) \cap V \neq \varnothing\}$.

When X is a topological vector space, by $P_f(X)$ (resp. $P_k(X)$) we will denote the nonempty closed (resp. compact) subsets of X. A w in front of f (resp. k) will mean that the closedness (resp. compactness) is with respect to the weak topology $w(X, X^*)$. Finally a c after f or k will mean that the set is in addition convex.

The next theorem summarizes the major results existing on the measurability of closed valued multifunctions.

THEOREM 2.1. *Let (Ω, Σ) be a measurable space and X a separable metric space. Let $F:\Omega \to P_f(X)$ be a multifunction.*

Consider the following statements:

(1) $F^-(B) \in \Sigma$ *for every* $B \in B(X) =$ *the Borel* σ*-field of* X.

(2) $F^-(C) \in \Sigma$ *for every* C *a closed subset of* X.

(3) $F^-(U) \in \Sigma$ *for every* U *an open subset of* X.

(4) $\omega \to d_{F(\omega)}(x) = \inf_{z \in F(\omega)} \|x - z\|$ *is measurable for all* $x \in X$.

(5) *There exists a sequence of measurable selectors* $f_n(\cdot)$ *of* $F(\cdot)$ *s.t.* $F(\omega) = \operatorname{cl}\{f_n(\omega)\}_{n \geq 1}$ (*Castaing representation*).

(6) $\operatorname{Gr} F \in \Sigma \times B(X)$.

Then we have the following results:

(i) (1)$\Rightarrow$(2)$\Rightarrow$(3)$\Leftrightarrow$(4)$\Rightarrow$(6).

(ii) *If X is a Polish space (i.e., in addition is complete) then* (3)$\Leftrightarrow$(5).

(iii) *If X is Polish and there is a complete σ-finite measure on Σ then* (1)-(6) *are all equivalent.*

Following Himmelberg [14], we say that a multifunction $F:\Omega \to P_f(X)$ satisfying (1) (resp. (2), (3)) is Borel (resp. strongly, weakly) measurable.

Let X be a separable Banach space. For any multifunction $F:\Omega \to 2^X$ we can define the set

$$S_F^1 = \{f(\cdot) \in L_X^1(\Omega): f(\omega) \in F(\omega) \mu\text{-a.e.}\},$$

i.e., S_F^1 contains all the integrable selectors of $F(\cdot)$. Clearly S_F^1 may be empty. If it is nonempty and $F(\cdot)$ is closed valued then it is easy to check that it is a closed subset of $L_X^1(\Omega)$.

Using this set we can define an integral for $F(\cdot)$:

$$\int_\Omega F(\omega)\, d\mu(\omega) = \left\{ \int_\Omega f(\omega)\, d\mu(\omega): f \in S_F^1 \right\}.$$

This was first introduced by Aumann [2] for $X = \mathbb{R}^n$. It is a natural generalization of both the integral of single valued functions and of the Minkowski sum of sets. It turned out to be a very powerful tool in several areas of applied mathematics especially in optimal control and mathematical economics.

In our case the integrals $\int_\Omega f(\omega)\, d\mu(\omega) f \in S_F^1$ are taken in the sense of Bochner [10]. Clearly if $S_F^1 = \varnothing$ then $\int_\Omega F(\omega)\, d\mu(\omega) = \varnothing$.

We will say that a measurable multifunction $F:\Omega \to P_f(X)$ is integrably bounded if there is a $\varphi(\cdot) \in L^1(\Omega)$ s.t.

$$|F(\omega)| = \sup_{x \in F(\omega)} \|x\| \leqq \varphi(\omega) \mu\text{-a.e.}$$

Using the Kuratowski-Ryll-Nardzewski selection theorem [21] we can see that for an integrably bounded multifunction $S_F^1 \neq \varnothing$ and so $\int_\Omega F(\omega)\, d\mu(\omega) \neq \varnothing$.

According to Hiai-Umegaki [13] for Σ_0 a sub-σ-field of Σ and $F:\Omega \to P_f(X)$ an integrably bounded multifunction, there exists a unique Σ_0-measurable multifunction $E^{\Sigma_0}F:\Omega \to P_f(X)$ which is integrable bounded also and for which we have that

$$S_{E^{\Sigma_0}F}^1 = \operatorname{cl}\{E^{\Sigma_0} f: f \in S_F^1\}$$

where the closure is taken in the $L^1_X(\Omega)$-norm topology. From here on for simplicity we will write $S^1_{E^{\Sigma_0}F} = S^1(\Sigma_0)$. The multifunction $E^{\Sigma_0}F(\cdot)$ is called the set valued conditional expectation of $F(\cdot)$ with respect to the sub-σ-field Σ_0.

Finally we will introduce a notion of convergence of closed sets, which is different from the convergence in the Hausdorff metric and which is more appropriate in the study of the stability of optimization and variational problems. This mode of set convergence was first introduced by Mosco [26], [27] and was extensively studied by Salinetti-Wets [45]-[47]. So let X be a Banach space and let $\{A_n\}_{n\geq 1}$ be a sequence of nonempty, closed subsets of X. Let t be a topology on X. We will say that A_n t-converges in the Kuratowski-Mosco sense to A if

$$t\text{-}\overline{\lim_{n\to\infty}} A_n \subseteq A \subseteq t\text{-}\underline{\lim_{n\to\infty}} A_n,$$

where

$$t\text{-}\overline{\lim_{n\to\infty}} A_n = \{x = t\text{-}\lim_{m\to\infty} x_m : x_m \in A_m,\ m \in M \subseteq N\}$$

and

$$t\text{-}\underline{\lim_{n\to\infty}} A_n = \{x = t\text{-}\lim_{n\to\infty} x_n : x_n \in A_n,\ n \in N\}.$$

Since we always have that

$$t\text{-}\underline{\lim_{n\to\infty}} A_n \subseteq t\text{-}\overline{\lim_{n\to\infty}} A_n,$$

we deduce that $\{A_n\}_{n\geq 1}$ t-converges to A in the Kuratowski-Mosco sense if and only if $t\text{-}\underline{\lim}_{n\to\infty} A_n = A = t\text{-}\overline{\lim}_{n\to\infty} A_n$. In that case we write that

$$A_n \xrightarrow{t\text{K-M}} A \quad \text{as } n \to \infty.$$

When $w\text{-}\overline{\lim}_{n\to\infty} A_n = A = s\text{-}\underline{\lim}_{n\to\infty}$ (where w denotes the weak topology and s the strong topology on X), we say that A_n converges to A in the Kuratowski-Mosco sense and we write

$$A_n \xrightarrow{\text{K-M}} A \quad \text{as } n \to \infty.$$

Finally if $\{f_n, f\}_{n\geq 1}$ is a sequence of $\bar{\mathbb{R}}$-valued closed convex functions on X, we say that $f_n \xrightarrow{\tau} f$ as $n \to \infty$ if and only if

$$\operatorname{epi} f_n \xrightarrow{\text{K-M}} \operatorname{epi} f \quad \text{as } n \to \infty.$$

We have defined price systems to be functions in $[L^1_X(\Omega)]^*_+$. We know that in general $[L^1_X(\Omega)]^* = L^\infty_{X^*_{w^*}}(\Omega)$ (see [20]) where $X^*_{w^*}$ is the space X^* with the w^*-topology. If X^* has the Radon-Nikodym property then $[L^1_X(\Omega)]^* = L^\infty_{X^*}(\Omega)$.

Throughout this work (Ω, Σ, μ) will be a complete probability space and X a separable Banach space. Additional hypotheses will be introduced as needed. Since we will be dealing with concave functions all the notions and results of convex analysis will be used in a "concave-context". Thus if $f: X \to \bar{\mathbb{R}}$ ($\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty\}$) is concave, we take $\partial_\varepsilon f(x) = \{x^* \in X^* : f(z) - f(x) \leqq (x^*, z - x) + \varepsilon \text{ for all } z \in X\}$, $\varepsilon \geqq 0$. The conjugate $f^*: X^* \to \bar{\mathbb{R}}$ of $f(\cdot)$ is defined by $f^*(x^*) = \inf_{x\in X} \{(x^*, x) - f(x)\}$.

Briefly the organization of the paper is as follows: In the next section we study the properties of efficient, price efficient and weakly efficient allocations. In § 4 we conduct an analogous study for optimal allocations. In § 5 we deal with the approximate

versions of those concepts. Finally in § 6 we examine the stability (robustness) of those notions under perturbations of the data and we also examine what happens to them when we restrict our attention to a particular economic sector.

We believe that our work illustrates in a rather convincing way, how some elegant mathematical theories, like Rockafellar's theory of normal integrands, Clarke's theory of generalized gradients and the Kuratowski-Mosco convergence of sets and functions can have nontrivial applications in mathematical economics.

3. Efficient and price efficient allocations. We will start with an existence result concerning efficient allocations. For that purpose assume that $\operatorname{int} X_+^* \neq \varnothing$, $F: \Omega \to P_{fc}(X)$ is integrably bounded and S_F^1 is locally w-compact in $L_X^1(\Omega)$. Then we have:

THEOREM 3.1. *Under the above hypotheses the set of efficient allocations is nonempty.*

Proof. Since by hypothesis $F(\cdot)$ is integrably bounded we have that S_F^1 is nonempty, bounded, closed and convex. So $\operatorname{As}(S_F^1) = \{0\}$ where $\operatorname{As}(\cdot)$ denotes the asymptotic cone of the set under consideration.

For any $g(\cdot) \in S_F^1$ we have that

$$(*) \qquad \operatorname{As}[S_F^1 \cap (g + (L_X^1)_+] = \operatorname{As}(S_F^1) \cap \operatorname{As}(g + (L_X^1)_+) = \operatorname{As} S_F^1 \cap (L_X^1)_+ = \{0\}.$$

By hypothesis S_F^1 is locally w-compact. Then so is $(S_F^1 \cap (g + (L_X^1)_+))$. Hence using (*) and Theorem I-9 of Castaing-Valadier [6] we conclude that $K_g = S_F^1 \cap (g + (L_X^1)_+)$ is w-compact in $L_X^1(\Omega)$.

Since $\operatorname{int} X_+^* \neq \varnothing$, we can see that $\operatorname{int}(L_{X^*}^\infty)_+ \neq \varnothing$. So let $p(\cdot) \in \operatorname{int}(L_{X^*}^\infty)_+$. Because K_g is w-compact, there exists $f(\cdot) \in K_g$ where $p(\cdot)$ achieves its supremum on K_g. We claim that $f(\cdot)$ is an efficient allocation. Suppose not. Then we can find $\hat{f} \in S_F^1$ s.t. $\hat{f} > f$ (i.e. $\mu\{\omega \in \Omega: \hat{f}(\omega) - f(\omega) \in \dot{X}_+\} > 0$). Clearly then $\hat{f} \in K_g$ and furthermore since $p(\cdot) \in \operatorname{int}(L_{X^*}^\infty)_+$, we have that

$$\langle p, \hat{f} - f \rangle > 0 \Rightarrow \langle p, f \rangle > \sigma_{K_g}(p),$$

a blatant contradiction.

So indeed $f(\cdot)$ is an efficient allocation. Q.E.D.

Remark. In [30] Theorem 3.4 gives us a necessary and sufficient condition for S_F^1 to be locally w-compact. This is the following:

> S_F^1 is a locally w-compact subset of $L_X^1(\Omega)$ if and only if the polar $(S_F^1)^0$ has a nonempty relative interior for the Mackey topology $m(L_{X^*_{w^*}}, L_X^1)$ and span $(S_F^1)^0$ is closed and of finite codimension.

Now we turn to the study of the geometric properties of the efficient and price efficient allocations.

We will start with a topological property of efficient allocations. So assume that $F: \Omega \to P_f(X)$ is integrably bounded.

PROPOSITION 3.1. *If $f(\cdot) \in L_X^1(\Omega)$ is an efficient allocation then $f(\omega) \in \operatorname{bd} F(\omega)$, $\omega \in \Omega$.*

Proof. Without any loss of generality we may assume that $\operatorname{int} F(\omega) \neq \varnothing$ for all $\omega \in \Omega$. Suppose that for $A \in \Sigma$ with $\mu(A) > 0$ and for $\omega \in A$ we have that $f(\omega) \in \operatorname{int} F(\omega)$. This means that there exists an $\varepsilon > 0$ depending on ω s.t.

$$[f(\omega) + B_\varepsilon^X(0)] \subseteq F(\omega),$$

where $B_\varepsilon^X(0) = \{x \in X: \|x\| \leqq \varepsilon\}$.

Consider the following multifunction defined on $A \subseteq \Omega$:

$$R(\omega) = \{\varepsilon > 0: f(\omega) + B_\varepsilon^X(0) \subseteq F(\omega)\} = \{\varepsilon > 0: \|f(\omega) - \operatorname{bd} F(\omega)\| \geqq \varepsilon\}.$$

From Theorem 4.6 of Himmelberg [14] we know that $\omega \to \operatorname{bd} F(\omega)$ is a $P_f(X)$-valued, measurable multifunction. So by Castaing's representation theorem there exist $\{u_n(\cdot)\}_{n\geq 1} \subseteq S^1_{\operatorname{bd} F}$ s.t.

$$\operatorname{cl}\{u_n(\omega)\}_{n\geq 1} = \operatorname{bd} F(\omega).$$

Then $\|f(\omega) - \operatorname{bd} F(\omega)\| = \inf_{n\geq 1} \|f(\omega) - u_n(\omega)\|$. But note that for all $n \geq 1$, $\omega \to \|f(\omega) - u_n(\omega)\|$ is measurable. Hence $\omega \to \|f(\omega) - \operatorname{bd} F(\omega)\|$ is measurable.

Let $\varphi(\omega, \varepsilon) = \|f(\omega) - \operatorname{bd} F(\omega)\| - \varepsilon$. Clearly this is a Caratheodory function and so is jointly measurable. Observe that

$$\operatorname{Gr} R = \{(\omega, \varepsilon) \in \Omega \times \mathbb{R}_+ \colon \varphi(\omega, \varepsilon) \geq 0\}.$$

Hence $\operatorname{Gr} R \in \Sigma_A \times B(\mathbb{R})$ where $\Sigma_A = \Sigma \cap A$. So we can apply Aumann's measurable selection theorem and find $\varepsilon : A \to \mathbb{R}_+$ measurable s.t. $\varepsilon(\omega) \in R(\omega)\omega \in A$. Therefore,

$$f(\omega) + B^X_{\varepsilon(\omega)}(0) \subseteq F(\omega) \quad \text{for } \omega \in A.$$

Now consider the multifunction $\Gamma(\omega) = B_{\varepsilon(\omega)}(0) \subseteq X_+$. Theorem 2.4 of Himmelberg [14] tells us that $\Gamma(\cdot)$ is measurable and closed valued. Applying the Kuratowski-Ryll-Nardzewski measurable section theorem we can find $e : A \to X_+$ measurable s.t. $e(\omega) \in \Gamma(\omega)$ for all $\omega \in A$. Then $g(\omega) = f(\omega) + e(\omega) \in F(\omega)$ for $\omega \in A$.

Define

$$\hat{g}(\omega) = \begin{cases} g(\omega) & \text{if } \omega \in A, \\ f(\omega) & \text{if } \omega \in \Omega \backslash A. \end{cases}$$

Clearly $\hat{g} \in S^1_F$ and $\hat{g} > f$, contradicting the hypothesis that $f(\cdot)$ is an efficient allocation. Q.E.D.

The next result provides an interesting global geometric characterization of efficient allocations. Recall that in a topological vector space a closed, convex subset is said to be rotund if every boundary point of the set is an extreme point.

PROPOSITION 3.2. *If $F : \Omega \to P_f(X)$ is an integrably bounded multifunction with rotund values and $f(\cdot) \in L^1_X(\Omega)$ is an efficient allocation then $f(\cdot) \in \operatorname{ext} S^1_F$.*

Proof. From Proposition 3.1 we know that $f(\omega) \in \operatorname{bd} F(\omega)$, $\omega \in \Omega$. Because of the rotundity of the set $F(\omega)$ we have that $f(\omega) \in \operatorname{ext} F(\omega)$, $\omega \in \Omega$. Using the work of Benamara [3], we then conclude that

$$f \in S^1_{\operatorname{ext} F} = \operatorname{ext} S^1_F. \qquad \text{Q.E.D.}$$

We can have something similar for price efficient allocations. First we need to introduce a new piece of terminology. If $f(\cdot) \in L^1_X(\Omega)$ and we have that $f(\omega) \in \operatorname{ext} F(\omega)$ μ-a.e. we will say that $f(\cdot)$ is an extremal allocation.

The next result tells us that any price efficient allocation can be substituted by an extremal allocation which is still price efficient for the same price system.

We will need an auxiliary lemma, which actually is interesting in its own right. This result first appeared in [29] and we repeat it here for the convenience of the reader.

LEMMA I. *If $F : \Omega \to P_{wkc}(X)$ is integrably bounded then S^1_F is a w-compact convex subset of $L^1_X(\Omega)$.*

Proof. We know that S^1_F is a nonempty, bounded, closed and convex subset of $L^1_X(\Omega)$. Let $q(\cdot) \in [L^1_X(\Omega)]^*$. From the Dinculeanu-Foias theorem (see [20]) we know that $[L^1_X(\Omega)]^* = L^\infty_{X^*_{w^*}}(\Omega)$. So $q(\cdot) \in L^\infty_{X^*_{w^*}}(\Omega)$. Then

$$\sup_{f \in S^1_F} \langle q, f \rangle = \sup_{f \in S^1_F} \int_\Omega (q(\omega), f(\omega))\, d\mu(\omega).$$

Using Theorem 2.2 of [13] we get that

$$\sup_{f\in S_F^1}\langle q,f\rangle=\int_\Omega \sup_{y\in F(\omega)}(q(\omega),y)\,d\mu(\omega).$$

Consider the multifunction

$$\Gamma(\omega)=\{h\in G(\omega):(q(\omega),h)=\sigma_{F(\omega)}(p(\omega))\}.$$

Because $F(\cdot)$ is $P_{wkc}(X)$-valued we can see that $\Gamma(\cdot)$ takes values in $P_f(X)$. Furthermore for the same reason $\sigma_{F(\cdot)}(\cdot)$ is a Caratheodory function. Hence $\omega\to\sigma_{F(\omega)}(q(\omega))$ is measurable. Then $\varphi(\omega,h)=(q(\omega),h)-\sigma_{F(\omega)}(q(\omega))$ being a Caratheodory function is jointly measurable. Observe that

$$\Gamma(\omega)=\{h\in F(\omega):\varphi(\omega,h)=0\}.$$

So $\operatorname{Gr}\Gamma\in\Sigma\times B(X)$. Applying Aumann's selection theorem we can find $h:\Omega\to X$ measurable s.t. $\sigma_{F(\omega)}(q(\omega))=(q(\omega),h(\omega))$, $\omega\in\Omega$. Therefore we have that

$$\sup_{f\in S_F^1}\langle q,f\rangle=\int_\Omega (q(\omega),h(\omega))\,d\mu(\omega)=\langle q,h\rangle.$$

Since $h(\cdot)\in S_F^1$ we conclude using James' theorem (see [19]) that S_F^1 is w-compact as claimed. Q.E.D.

Now we can proceed and state our result on price efficient allocations.

PROPOSITION 3.3. *If $F:\Omega\to P_{wkc}(X)$ is integrably bounded then for any $p(\cdot)\in[L^\infty_{X^*_{w^*}}]\backslash\{0\}$ we can find an extremal allocation which is price efficient for that price system $p(\cdot)$.*

Proof. From the previous lemma we know that S_F^1 is a w-compact subset of $L_X^1(\Omega)$. So $\sup_{f\in S_F^1}\langle p,f\rangle$ is attained at some element of S_F^1. From Bauer's maximum principle we know that this element can be taken to belong in ext S_F^1. So for some $\hat f(\cdot)\in\operatorname{ext} S_F^1$, $\langle p,\hat f\rangle=\sigma_{S_F^1}(p)$. On the other hand, from Benamara [3] we know that ext $S_F^1=S^1_{\operatorname{ext} F}$. Hence $\hat f(\cdot)\in S^1_{\operatorname{ext} F}$ which means that $\hat f(\cdot)$ is an extremal allocation. Q.E.D.

Now let $F:\Omega\to P_{fc}(X)$ be an integrably bounded multifunction and consider the following set

$$G(F)=\{(f,p)\in L_X^1(\Omega)\times[L^\infty_{X^*_{w^*}}(\Omega)]_+:f(\cdot)\text{ is }p\text{-efficient}\},$$

i.e., $G(F)$ contains all pairs of allocations and price systems s.t. the allocation is price efficient for that price system.

PROPOSITION 3.4. (1) *$G(F)$ is closed in $(L_X^1(\Omega),s)\times(L^\infty_{X^*_{w^*}}(\Omega),w^*)$.*

(2) *If X has the Schur property then $G(F)$ is closed in $(L_X^1(\Omega),w)\times(L^\infty_{X^*_{w^*}},w)$.*

Proof. (1) Let (f_a,p_a) be a net in $G(F)$ s.t. $(f_a,p_a)\xrightarrow{s\times w^*}(f,p)$. Then by definition we have that

$$\langle p_a,f_a\rangle=\sigma_{S_F^1}(p_a).$$

Because

$$f_a\xrightarrow{s\text{-}L_X^1}f\quad\text{and}\quad p_a\xrightarrow{w^*\text{-}L^\infty_{X^*_{w^*}}}p$$

we have that $\langle p_a,f_a\rangle\to\langle p,f\rangle$. On the other hand, $\underline{\lim}_a\,\sigma_{S_F^1}(p_a)\geqq\sigma_{S_F^1}(p)$. So $\sigma_{S_F^1}(p)\leqq\langle p,f\rangle$. But note that $f\in S_F^1$. Hence $\sigma_{S_F^1}(p)=\langle p,f\rangle$ which means that $f(\cdot)$ is p-efficient, i.e., $(f,p)\in G(F)$.

(2) If X has the Schur property, then we know from Khurana [49] that $L^1_X(\Omega)$ has the Dunford-Pettis property. So if

$$f_a \xrightarrow{w\text{-}L^1_X} f \quad \text{and} \quad p_a \xrightarrow{w\text{-}L^\infty_{X^*_{w^*}}} p$$

then $\langle p_a, f_a\rangle \to \langle p, f\rangle$. Again we can get that $\sigma_{S^1_F}(p) \leqq \langle p, f\rangle$ and since $f \in S^1_F$ equality holds. So $(f, p) \in G(F)$. Q.E.D.

Now we will pass to a series of results that compare efficiency and price efficiency.

Although a price efficient allocation is not necessarily efficient the next result tells us that price efficient allocations are w-dense in the set of efficient allocations. For that purpose assume that Σ is countably generated and int $X^*_+ \neq \varnothing$.

THEOREM 3.2. *If $F: \Omega \to P_{wkc}(X)$ is integrably bounded then the $w(L^1_X(\Omega), L^\infty_{X^*_{w^*}}(\Omega))$-closure of the set of all price efficient allocations contains the set of efficient allocations.*

Proof. Observe that if $f(\cdot) \in S^1_F$ is efficient for $F(\cdot)$ it is also efficient for $(F(\cdot) - X_+)$ and vice versa. So from Proposition 3.1 we know that $f(\omega) \in bd(F(\omega) - X_+)\omega \in \Omega$, which implies that $f(\cdot) \in bd(S^1_{F-X_+})$.

From the Bishop-Phelps theorem (see [12]) we know that there exist $\{f_n\}_{n\geqq 1} \subseteq bd(S^1_{F-X_+})$ s.t. $f_n(\cdot)$ is $p_n(\cdot)$-efficient with $\|p_n\|_{L^\infty_{X^*_{w^*}}} \leqq 1$ and

$$f_n \xrightarrow{s\text{-}L^1_X} f \quad \text{as } n \to \infty.$$

Hence

$$\langle p_n, f_n\rangle = \sigma_{S^1_{F-X^+}}(p_n) \Rightarrow \langle p_n, f_n\rangle \geqq \langle p_n, f_n - e\rangle$$

for all $e(\cdot) \in [L^1_X(\Omega)]_+$. So $\langle p_n, e\rangle \geqq 0$ for all $e(\cdot) \in [L^1_X(\Omega)]_+$ which means that $p_n(\cdot) \in [L^\infty_{X^*_{w^*}}(\Omega)]_+$ for all $n \geqq 1$.

Because Σ is countably generated $L^1_X(\Omega)$ is separable and so we can assume that

$$p_n \xrightarrow{w^*\text{-}L^\infty_{X^*_{w^*}}} p \quad \text{as } n \to \infty$$

and for all $n \geqq 1 \langle p_n, f_n\rangle = \sigma_{S^1_{F-X^+}}(p_n) = \sigma_{S^1_F}(p_n)$.

For $n \geqq 1$ define $\Delta(f_n) = \{g(\cdot) \in S^1_F : g \geqq f_n\}$. Since $\Delta(f_n)$ is a closed, convex subset of S^1_F and S^1_F is w-compact we deduce that $\Delta(f_n)$ is w-compact for all $n \geqq 1$. Our claim is that $\Delta(f_n)$ has a maximal element with respect to the cone $[L^1_X(\Omega)]_+$. Suppose not. Then for every $g(\cdot) \in \Delta(f_n)$ we can find $g'(\cdot) \in \Delta(f_n)$ s.t. $g' > g$. Since by hypothesis int $X^*_+ \neq \varnothing$ we can see that int $[L^\infty_{X^*_{w^*}}(\Omega)]_+ \neq \varnothing$. Let $\hat{p}(\cdot) \in$ int $(L^\infty_{X^*_{w^*}}(\Omega))_+$. Because $\Delta(f_n)$ is w-compact, there exists $\hat{g}(\cdot) \in \Delta(f_n)$ s.t. $\sigma_{\Delta(f_n)}(\hat{p}) = \langle \hat{p}, \hat{g}\rangle$. From what we said above we can find $\hat{\hat{g}} \in \Delta(f_n)$ s.t. $\hat{\hat{g}} > \hat{g}$. Then $\langle \hat{p}, \hat{\hat{g}} - \hat{g}\rangle > 0$ which means that $\langle \hat{p}, \hat{\hat{g}}\rangle > \sigma_{\Delta(f_n)}(\hat{p})$, a contradiction.

So let $g_n(\cdot) \in \Delta(f_n)$ be maximal. Then clearly $g_n(\cdot)$ is maximal in S^1_F. But then this means that $g_n(\cdot)$ is efficient. Because of the w-compactness of S^1_F and the Eberlein-Smulian theorem (see [11, p. 430]) we may assume, without loss of generality, that

$$g_n \xrightarrow{w\text{-}L^1_X} g \in S^1_F \quad \text{as } n \to \infty.$$

Then $g \geqq f$ and since $f(\cdot)$ is by hypothesis efficient we may conclude (after changing $g(\cdot)$ on a μ-null set) that $g(\omega) = f(\omega)$ for all $\omega \in \Omega$. Q.E.D.

We can have an analogue of Theorem 3.2 for $L^\infty_X(\Omega)$ allocations. Assume that X is reflexive and int $X^*_+ \neq \varnothing$.

THEOREM 3.3. *If $F:\Omega \to P_{wkc}(X)$ is integrably bounded by $\varphi(\cdot) \in L^{\infty}(\Omega)$ then the norm-closure of the set of price efficient allocations with price systems in $[L^1_{X^*}(\Omega)]_+$ contains the set of efficient allocations.*

The proof of this result follows the pattern of the proof of the previous theorem, but instead of the Bishop-Phelps theorem we use a generalization of it due to Phelps [36, Thm. 1].

We can now go further and determine a situation where price efficiency implies efficiency.

So again assume that $\operatorname{int} X_+^+ \neq \varnothing$. Then we have the following result.

PROPOSITION 3.5. *If $F:\Omega \to P_f(X)$ is integrably bounded and $f(\cdot) \in L^1_X(\Omega)$ is price efficient for the price system $p(\cdot) \in \operatorname{int}[L^{\infty}_{X^*_{w^*}}(\Omega)]_+$ then $f(\cdot)$ is efficient.*

Proof. Suppose not. Then as before we can find $\hat{f}(\cdot) \in S^1_F$ s.t. $(\hat{f}-f)(\cdot) \in [L^1_X(\Omega)]_+ \backslash \{0\}$. Then from Borwein [5] we know that

$$\langle p, \hat{f}-f\rangle > 0 \Rightarrow \langle p, \hat{f}\rangle > \sigma_{S^1_F}(p),$$

a contradiction. Q.E.D.

Once again we turn our attention to $L^{\infty}_X(\Omega)$ allocations. As we already mentioned in § 2, the difficulty that we encounter in this case is that $L^1_{X^*}(\Omega) \subsetneq [L^{\infty}_X(\Omega)]^*$ and the elements of $[L^{\infty}_X(\Omega)]^*$ outside $L^1_{X^*}(\Omega)$ do not have in general a satisfactory economic interpretation. So the natural question to ask, is whether we can always choose our price system to belong in $L^1_{X^*}(\Omega)$. The next theorem gives us conditions that produce an affirmative answer to this question.

Before going into the theorem we would like to recall a few facts about $L^{\infty}_X(\Omega)$. Details can be found in Levin [22] and Rockafellar [41]. A functional $x^* \in [L^{\infty}_X(\Omega)]^*$ is said to be "absolutely continuous" relative to μ if $\langle x^*, x(\cdot)\rangle = \int_{\Omega} (x^*(\omega), x(\omega))\, d\mu(\omega)$ for all $x(\cdot) \in L^{\infty}_X(\Omega)$, where $x^*(\cdot) \in L^1_{X^*}(\Omega)$. In this case we call $x^*(\cdot)$ the density of the functional x^*. The absolutely continuous functionals form a closed subspace of $[L^{\infty}_X(\Omega)]^*$ which is isometrically isomorphic to $L^1_{X^*}(\Omega)$. This subspace has a natural complement in $[L^{\infty}_X(\Omega)]^*$, the subspace of singular functionals. A functional $x^* \in [L^{\infty}_X(\Omega)]^*$ is said to be "singular" relative to μ if there exists a sequence of sets $A_n \in \Sigma$ s.t. $A_{n+1} \subseteq A_n$ for all $n \geqq 1$, $\mu(A_n) \downarrow 0$ as $n \to \infty$ and $\langle x^*, x(\cdot)\rangle = 0$ for every $x(\cdot) \in L^{\infty}_X(\Omega)$ which vanishes on some A_n. Then each element $x^* \in [L^{\infty}_X(\Omega)]^*$ has a unique decomposition of the form $x^* = x^*_a + x^*_s$ where x^*_a is absolutely continuous relative to μ and x^*_s is singular relative to μ. Moreover $\|x^*\| = \|x^*_a\| + \|x^*_s\|$.

Assume that $\operatorname{int} X_+ \neq \varnothing$.

THEOREM 3.4. *If $F:\Omega \to P_{fc}(X)$ is integrably bounded by $\varphi(\cdot) \in L^{\infty}(\Omega)$ and $f(\cdot) \in S^{\infty}_F$ is an efficient allocation then there exists $p(\cdot) \in [L^1_{X^*}(\Omega)]_+$ s.t. $f(\cdot)$ is price efficient for that price system $p(\cdot)$.*

Proof. Since by hypothesis $f(\cdot)$ is efficient we know that

$$(f + [\dot{L}^{\infty}_X(\Omega)]_+) \cap S^{\infty}_F = \varnothing \qquad ([\dot{L}^{\infty}_X(\Omega)]_+ = [L^{\infty}_X(\Omega)]\backslash\{0\}).$$

Recall that $\operatorname{int}[L^{\infty}_X]_+ \neq \varnothing$ and that S^{∞}_F is convex. So we can apply the first separation theorem for convex sets and find $p(\cdot) \in [L^{\infty}_X(\Omega)]^* \backslash \{0\}$ s.t.

$$\langle p, f+e\rangle \geqq \langle p, g\rangle$$

for all $e(\cdot) \in [L^{\infty}_X(\Omega)]_+$ and all $g(\cdot) \in S^{\infty}_F$. (Here $\langle \cdot, \cdot \rangle$ denotes the duality brackets between $L^{\infty}_X(\Omega)$ and $[L^{\infty}_X(\Omega)]^*$.) Taking $g = f$ we get that $\langle p, e\rangle \geqq 0$ and since $e(\cdot) \in [L^{\infty}_X(\Omega)]_+$ is arbitrary we conclude that $p(\cdot) \in [L^{\infty}_X(\Omega)]^*_+$.

Next note that $\langle p, f\rangle = \sigma_{S^{\infty}_F}(p)$. We claim that $\langle p_a, f\rangle = \sigma_{S^{\infty}_F}(p_a)$ where p_a is the absolutely continuous part of the functional $p \in [L^{\infty}_X(\Omega)]^*$. Let $\{A_n\}_{n \geqq 1} \subseteq \Sigma$ be the sets

that have, relative to p_s, the property described in the definition of the singular part of p. So $A_{n+1} \subseteq A_n n \geqq 1$ and $\mu(A_n) \downarrow 0$ as $n \to \infty$ while $\langle p_s, g \rangle = 0$ if $g|_{A_n} = 0$ for some $n \geqq 1$. Define

$$g_n(\omega) = \begin{cases} f(\omega) & \text{for } \omega \in A_n, \\ g(\omega) & \text{for } \omega \in \Omega \backslash A_n, \end{cases}$$

where $g(\cdot) \in S_F^\infty$ arbitrary. Clearly $g_n(\cdot) \in S_F^\infty$.

Note that

$$g_n(\cdot) \xrightarrow{\mu} g(\omega) \quad \text{as } n \to \infty.$$

By passing to a subsequence, if necessary, we may assume that $g_n(\omega) \to g(\omega)$ μ-a.e. as $n \to \infty$. Also note that $\langle p_s, g_n - f \rangle = 0$. So we have that

$$\langle p, g_n - f \rangle = \langle p_a + p_s, g_n - f \rangle = \langle p_a, g_n - f \rangle \leqq 0.$$

Applying Lebesgue's dominated convergence theorem we get that $\langle p_a, g_n - f \rangle \to \langle p_a, g - f \rangle \leqq 0$ as $n \to \infty$ and since $g(\cdot) \in S_F^\infty$ was arbitrary we conclude that $\langle p_a, f \rangle = \sigma_{S_F^\infty}(p_a)$ which proves that $f(\cdot)$ is price efficient for the price system $p_a(\cdot) \in L_{X^*}^1(\Omega)$. Q.E.D.

In the next result we examine what happens to price efficient allocations when we have two consumption (or production) multifunctions and we pass to their intersection.

PROPOSITION 3.6. *If $F_1 : \Omega \to P_{kc}(X)$ and $F_2 : \Omega \to P_{fc}(X)$ are integrably bounded, $\operatorname{int} S_{F_2}^1 \neq \varnothing$, $S_{F_1}^1 \cap \operatorname{int} S_{F_2}^1 \neq \varnothing$ and $f(\cdot)$ is price efficient for the multifunction $(F_1 \cap F_2)(\cdot)$ and for the price system $p(\cdot)$ then $f(\cdot)$ is price efficient for $(p(\cdot), F_1(\cdot))$ and $(p(\cdot), F_2(\cdot))$.*

Proof. Because by hypothesis $f(\cdot)$ is price efficient for $p(\cdot)$ and $(F_1 \cap F_2)(\cdot)$ we have that $\langle p, f \rangle = \sigma_{S_{F_1 \cap F_2}^1}(p) = \sigma_{S_{F_1}^1 \cap S_{F_2}^1}(p)$. Since $S_{F_1}^1 \cap \operatorname{int} S_{F_2}^1 \neq \varnothing$ from Moreau [25] we know that

$$\begin{aligned} \sigma_{S_{F_1}^1 \cap S_{F_2}^1}(p) &= [\sigma_{S_{F_1}^1} \square \sigma_{S_{F_2}^1}](p) \\ &= \inf_{p' \in L_{X^{**}(\Omega)}^\infty} [\sigma_{S_{F_1}^1}(p - p') + \sigma_{S_{F_2}^1}(p')] \\ &\geqq \langle p - p', f \rangle + \langle p', f \rangle = \langle p, f \rangle. \end{aligned}$$

Since $\langle p, f \rangle = \sigma_{S_{F_1}^1 \cap S_{F_2}^1}(p)$, we get that the above infimal convolution is exact at $p = p + 0$ and $p = 0 + p$. So $f(\cdot)$ is $(p(\cdot), F_1(\cdot))$ and $(p(\cdot), F_2(\cdot))$ efficient. Q.E.D.

Now we will study the concepts of efficiency and price efficiency for a particular sector of the economy. Recall that we model sectors by sub-σ-fields of Σ.

For that we will need the following lemma, the proof of which can be found in [28, § 5].

LEMMA II. *If $p(\cdot) \in L_{X^*}^\infty(\Omega, \Sigma_0)$ and $f(\cdot) \in L_X^1(\Omega)$ then $\langle p, f \rangle = \langle p, E^{\Sigma_0} f \rangle$.*

For the next theorem assume that X^* is separable. Then from the Dunford-Pettis theorem (see [10]) we know that X^* has the Radon-Nikodym property and so $[L_X^1(\Omega)]^* = L_{X^*}^\infty(\Omega)$. Also assume that X_+^* is generating.

THEOREM 3.5. (i) *If $F : \Omega \to P_{wkc}(X)$ is integrably bounded and $v(\cdot) \in S_{E^{\Sigma_0} F}^1$ is efficient for $E^{\Sigma_0} F(\cdot)$ then there exists $f(\cdot) \in S_F^1$ which is efficient for $F(\cdot)$ and such that $v(\cdot) = E^{\Sigma_0} f(\cdot)$.*

(2) *Suppose $F(\cdot)$ is as above and $p(\cdot) \in [L_{X^*}^\infty(\Omega, \Sigma_0)]_+$. Then: $f(\cdot) \in S_F^1$ is price efficient for $p(\cdot)$, $F(\cdot)$ if and only if $E^{\Sigma_0} f(\cdot) \in S^1(\Sigma_0)$ is price efficient for $p(\cdot)$, $E^{\Sigma_0} F(\cdot)$.*

Proof. (1) From Hiai-Umegaki [13] we know that

$$S^1(\Sigma_0) = \operatorname{cl} E(S_F^1 | \Sigma_0)$$

where the closure is taken in the $L_X^1(\Omega, \Sigma_0)$-norm.

But from Lemma I we know that S_F^1 is w-compact and convex in $L_X^1(\Omega)$. So $E(S_F^1|\Sigma_0) = S^1(\Sigma_0)$. Therefore $v(\cdot) = E^{\Sigma_0} f(\cdot)$ for some $f(\cdot) \in S_F^1$. Our claim is that $f(\cdot)$ is efficient for $F(\cdot)$. Suppose not. Then there exists $A \in \Sigma$ with $\mu(A) > 0$ s.t.

$$G(\omega) = (f(\omega) + \dot{X}_+) \cap F(\omega) \neq \varnothing$$

for all $\omega \in A$. Note that the multifunction $\omega \to f(\omega) + \dot{X}_+$ has a measurable graph. To see that consider $h: (\omega, x) \to (\omega, x - f(\omega))$. Then $h^{-1}(\Omega \times \dot{X}_+) = \operatorname{Gr} G_1$ where $G_1(\omega) = f(\omega) + \dot{X}_+$, and $h(\cdot, \cdot)$ is $(\Sigma \cap A) \times B(X)$-measurable. So $\operatorname{Gr} G = \operatorname{Gr} G_1 \cap \operatorname{Gr} F \in (\Sigma \cap A) \times B(X)$. Hence we can apply Aumann's measurable selection theorem and find $g: A \to X$ measurable s.t. $g(\omega) \in (f(\omega) + X_+) \cap F(\omega) \omega \in A$. Next define

$$\tilde{g}(\omega) = \begin{cases} g(\omega) & \text{for } \omega \in A, \\ f(\omega) & \text{for } \omega \in \Omega \backslash A. \end{cases}$$

Clearly $\tilde{g} > f$. Then for all $p(\cdot) \in [L_{X^*}^{\infty}(\Omega, \Sigma_0)]^*$ we have that

$$\langle p, \tilde{g} - f \rangle \geqq 0 \Rightarrow \langle p, E^{\Sigma_0}(g - f) \rangle \geqq 0 \quad \text{(by Lemma II)}$$

$$\Rightarrow E^{\Sigma_0} \tilde{g} \geqq E^{\Sigma_0} f.$$

We claim that $E^{\Sigma_0} \tilde{g} > E^{\Sigma_0} f$. Because if $E^{\Sigma_0} \tilde{g}(\omega) = E^{\Sigma_0} f(\omega)$ μ-a.e. then $\int_\Omega E^{\Sigma_0} \tilde{g}(\omega)\, d\mu(\omega) = \int_\Omega E^{\Sigma_0} f(\omega)\, d\mu(\omega)$. Hence

$$\int_\Omega \tilde{g}(\omega)\, d\mu(\omega) = \int_\Omega f(\omega)\, d\mu(\omega)$$

$$\Rightarrow \left(x^*, \int_\Omega \tilde{g}(\omega)\, d\mu(\omega) \right) = \left(x^*, \int_\Omega f(\omega)\, d\mu(\omega) \right) \quad \text{for all } x \in X_+^*$$

$$\Rightarrow \int_\Omega (x^*, \tilde{g}(\omega))\, d\mu(\omega) = \int_\Omega (x^*, f(\omega))\, d\mu(\omega)$$

$$\Rightarrow \int_\Omega (x^*, \tilde{g}(\omega) - f(\omega))\, d\mu(\omega) = 0.$$

Recall that $(x^*, \tilde{g}(\omega) - f(\omega)) \geqq 0$ μ-a.e. Hence $(x^*, \tilde{g}(\omega) - f(\omega)) = 0$ μ-a.e. for $x^* \in X_+^*$. Since by hypothesis X_+^* is generating, we get that for all $x^* \in X^*(x^*, \tilde{g}(\omega) - f(\omega)) = 0$ μ-a.e. and so $\tilde{g}(\omega) = f(\omega)$ μ-a.e., a contradiction. So $E^{\Sigma_0} f < E^{\Sigma_0} \tilde{g}$ which in turn contradicts the efficiency of $E^{\Sigma_0} f(\cdot) = v(\cdot)$ because $E^{\Sigma_0} \tilde{g} \in E^{\Sigma_0} S_F^1 = S_{E^{\Sigma_0} F}^1$. Therefore $f(\cdot)$ is indeed efficient for $F(\cdot)$.

(2) Suppose $f(\cdot) \in S_F^1$ is price efficient for $p(\cdot)$ and $F(\cdot)$, with $p(\cdot) \in [L_{X^*}^{\infty}(\Omega, \Sigma_0)]_+$. This means that

$$\langle p, f \rangle = \sigma_{S_F^1}(p)$$

$$\Rightarrow \int_\Omega (p(\omega), f(\omega))\, d\mu(\omega) = \sup_{g(\cdot) \in S_F^1} \int_\Omega (p(\omega), g(\omega))\, d\mu(\omega)$$

$$= \int_\Omega \sup_{y \in F(\omega)} (p(\omega), y)\, d\mu(\omega)$$

$$= \int_\Omega \sigma_{F(\omega)}(p(\omega))\, d\mu(\omega).$$

Using Lemma II and Bismut [4] we get that

$$\int_\Omega (p(\omega), E^{\Sigma_0} f(\omega))\, d\mu(\omega) = \int_\Omega E^{\Sigma_0} \sigma_{F(\omega)}(p(\omega))\, d\mu(\omega).$$

From [13, Thm. 5.5] (see also Valadier [48]) we know that

$$\int_\Omega E^{\Sigma_0}\sigma_{F(\omega)}(p(\omega))\,d\mu(\omega)=\int_\Omega \sigma_{\Sigma_0(\omega)}(p(\omega))\,d\mu(\omega)=\sigma_{S^1(\Sigma_0)}(p)$$

where $\sigma_{\Sigma_0(\omega)}(\cdot)=\sigma_{E^{\Sigma_0}F(\omega)}(\cdot)$.

Hence $\langle p, E^{\Sigma_0}f\rangle=\sigma_{S^1(\Sigma_0)}(p)$ which proves that $E^{\Sigma_0}f(\cdot)$ is price efficient for $p(\cdot)$ and $E^{\Sigma_0}F(\cdot)$.

Now suppose that $E^{\Sigma_0}f\in S^1(\Sigma_0)$ is price efficient for $p(\cdot)$ and $E^{\Sigma_0}F(\cdot)$. Then we have

$$\langle p, E^{\Sigma_0}f\rangle=\sigma_{S^1(\Sigma_0)}(p)=\int_\Omega \sigma_{\Sigma_0(\omega)}(p(\omega))\,d\mu(\omega)=\int_\Omega E^{\Sigma_0}\sigma_{F(\omega)}(p(\omega))\,d\mu(\omega).$$

Again by Lemma II and [4] we get that

$$\langle p, f\rangle=\int_\Omega \sigma_{F(\omega)}(p(\omega))\,d\mu(\omega)=\sigma_{S^1_F}(p).$$

which proves that $f(\cdot)$ is price efficient for $p(\cdot)$ and $F(\cdot)$. Q.E.D.

We will conclude this section by studying the topological properties of the set of weakly efficient allocations. We will denote that set by $wE(F)$.

Our first result is the following.

THEOREM 3.6. *If $F:\Omega\to P_f(X)$ is integrably bounded then $wE(F)$ is a w-closed subset of $L^1_X(\Omega)$.*

Proof. Let $\{f_\delta(\cdot)\}_{\delta\in\Delta}$ be a net in $wE(F)$ and suppose that $f_\delta \xrightarrow{w\text{-}L^1_X} f$. We will show that $f(\cdot)\in wE(F)$. Suppose not. Then there exists $B\in\Sigma$ with $\mu(B)>0$ s.t. for all $\omega\in B(f(\omega)+\operatorname{int} X_+)\cap F(\omega)\neq\varnothing$. For $\omega\in B$ let $R(\omega)=(f(\omega)+\operatorname{int} X_+)\cap F(\omega)$. Note that $\operatorname{Gr} R=\operatorname{Gr}(f(\cdot)+\operatorname{int} X_+)\cap \operatorname{Gr} F$ and as in the proof of Theorem 3.5 we can show that $\operatorname{Gr} R\in\Sigma_B\times B(X)$. Applying Aumann's measurable selection theorem we can find $r:B\to X$ measurable s.t. $r(\omega)\in R(\omega)\omega\in B$. So $r(\omega)\gg f(\omega)\in B$ (i.e., $r(\omega)-f(\omega)\in \operatorname{int} X_+$ for all $\omega\in B$). Therefore for every $A\in\Sigma$, $A\subseteq B$ and every $x^*\in X^*_+$ we have that

$$\int_A (x^*, r(\omega))\,d\mu(\omega)>\int_A (x^*, f(\omega))\,d\mu(\omega).$$

Since by hypothesis $f_\delta \xrightarrow{w\text{-}L^1_X} f$ we can find $\delta_0\in\Delta$ s.t. for $\delta\in\Delta$ $\delta\geqq\delta_0$ we have that

$$\int_A (x^*, r(\omega))\,d\mu(\omega)>\int_A (x^*, f_\delta(\omega))\,d\mu(\omega)\Rightarrow\int_A (x^*, r(\omega)-f_\delta(\omega))\,d\mu(\omega)>0$$

and this is true for all $A\in\Sigma$, $A\subseteq B$, $x^*\in X^*_+$ and $\delta\geqq\delta_0$. So we have that

$$(x^*, r(\omega)-f_\delta(\omega))>0\quad \mu_{\hat{A}}\text{-a.e.}$$

for some $\hat{A}\in\Sigma$, $\hat{A}\subseteq B\mu(\hat{A})>0$ and for all $x^*\in \dot{X}^*_+$. Hence $r(\omega))\gg f_\delta(\omega)$ $\mu_{\hat{A}}$-a.e. From that we get that

$$(f_\delta(\omega)+\operatorname{int} X_+)\cap F(\omega)\neq\varnothing$$

for all $\omega\in\hat{A}$ and all $\delta\geqq\delta_0$, a contradiction. So $f(\cdot)\in wE(F)$ and so $wE(F)$ is w-closed in $L^1_X(\Omega)$. Q.E.D.

If we strengthen our hypotheses on $F(\cdot)$ and X we can deduce even more about $wE(F)$. So assume that X is finite-dimensional and that $wE(F)\neq\varnothing$.

THEOREM 3.7. *If $F:\Omega\to P_k(X)$ is integrably bounded and for all $\omega\in\Omega$ $\operatorname{int} F(\omega)\neq\varnothing$ then $wE(F)$ is a w-compact subset of $L^1_X(\Omega)$.*

Proof. We already know that $wE(F)$ is w-closed. Clearly it is also bounded. So if we can show that every element in $[L_X^1(\Omega)]^* = L_{X^*}^\infty(\Omega)$ achieves its supremum on $wE(F)$, by James' theorem we are done.

So let $x^*(\cdot) \in L_{X^*}^\infty(\Omega)$. Then we have that

$$\sup_{f(\cdot)\in wE(F)} \langle x^*, f\rangle = \sup_{f(\cdot)\in wE(F)} \int_\Omega (x^*(\omega), f(\omega))\, d\mu(\omega).$$

Consider the multifunction $\omega \to wE(\omega) = \{x \in F(\omega): (x + \operatorname{int} X_+) \cap F(\omega) = \varnothing\}$. Clearly this is nonempty. We will show that it is also closed valued. Let $\{x_n\} \subseteq wE(\omega)$ and $x_n \to x$ and $n \to \infty$. Then $x \in F(\omega)$. Also for all $n \geqq 1$ $(x_n + \operatorname{int} X_+) \cap F(\omega) = \varnothing$. So $(x + \operatorname{int} X_+) \cap \operatorname{int} F(\omega) = \varnothing$. Suppose $x \notin wE(F)$. Then $(x + \operatorname{int} X_+) \cap F(\omega) \neq \varnothing$ and so $(x + \operatorname{int} X_+) \cap \operatorname{int} F(\omega) \neq \varnothing$. Since $(x_n + \operatorname{int} X_+) \to (x + \operatorname{int} X_+)$ in the Kuratowski sense, from Theorem 2.2 of Salinetti-Wets [47] we have that $(x_n + \operatorname{int} X_+) \cap \operatorname{int} F(\omega) \neq \varnothing$ for all $n \geqq n_0$, a contradiction. Hence for all $\omega \in \Omega$, $wE(\omega)$ is closed. Now observe that

$$\operatorname{Gr} wE(\cdot) = [\operatorname{dom} \Phi]^c \cap \operatorname{Gr} F,$$

where $\Phi(\omega, x) = (x + \operatorname{int} X_+) \cap F(\omega)$. It is easy to see that the graph of $\Phi(\cdot, \cdot)$ is $\Sigma \times B(X) \times B(X)$-measurable. So by Theorem 3.4 and Propositions 2.1 and 2.2 of Himmelberg [14] we have that $\operatorname{dom} \Phi \in \Sigma \times B(X)$. Therefore $\operatorname{Gr} wE(\cdot) \in \Sigma \times B(X)$ and then this tells us that $\omega \to wE(\omega)$ is measurable. Applying [13, Thm. 2.2] we get that

$$\sup_{f(\cdot)\in wE(F)} \langle x^*, f\rangle = \int_\Omega \sup_{x\in wE(\omega)} (x^*(\omega), x)\, d\mu(\omega).$$

Since $wE(\omega)$ is a closed subset of $F(\omega)$ it is compact. Hence if we define

$$M(\omega) = \{\hat{x} \in wA(\omega): (x^*(\omega), \hat{x}) = \sup_{x\in wE(\omega)} (x^*(\omega), x)\}$$

this set is nonempty and closed for all $\omega \in \Omega$. Also from [14, Thm. 6.4] we have that $\omega \to M(\omega)$ is measurable. Then by the Kuratowski-Ryll Nardzewski selection theorem we can find $\hat{x}: \Omega \to X$ measurable s.t. $\hat{x}(\omega) \in wE(\omega)\omega \in \Omega$. So $\hat{x}(\cdot) \in wE(F)$ and

$$\sup_{f\in wE(F)} \langle x^*, f\rangle = \langle x^*, \hat{x}\rangle$$

which means that $wE(F)$ is w-compact in $L_X^1(\Omega)$. Q.E.D.

Remark. If $F(\cdot)$ is also convex valued then we know that automatically we have $\operatorname{rint} F(\omega) \neq \varnothing$ for all $\omega \in B$. So the interiority assumption on $F(\cdot)$ is redundant.

4. Optimal allocations. In this section we focus our attention on optimal allocations. We will start with a powerful existence result. For that purpose we need to introduce the following notion:

a function $\varphi: X \to \bar{\mathbb{R}}$ is said to be quasi-sup-w-compact if and only if there exists $\hat{\beta} \in \mathbb{R}$ s.t. for $\beta \geqq \hat{\beta}$ we have that $U_\beta^\varphi = \{x \in X: \varphi(x) \geqq \beta\}$ is w-compact.

Using this concept we can have the following existence result.

THEOREM 4.1. *If $u: \Omega \times X \to \bar{\mathbb{R}}$ is an integrand s.t.*

(i) *$u(\cdot, \cdot)$ is $\Sigma \times B(X)$-measurable and as a function of x is weakly upper semicontinuous,*

(ii) *$u(\omega, \cdot)$ is μ-a.e. quasi-sup-w-compact, and if $F: \Omega \to P_{fc}(X)$ is integrably bounded then there exists an optimal allocation.*

Proof. Let $\beta \geqq \hat{\beta}$. Then for all $\omega \in \Omega$ we have that

$$\begin{aligned} U^{u(\omega,\cdot)} &= \{x \in X: u(\omega, x) \geqq \beta\} \\ &\supseteq \{x \in X: u(\omega, x) - \delta_{F(\omega)}(x) \geqq \beta\} \\ &= U_{\beta}^{[u(\omega,\cdot)-\delta_{F(\omega)}(\cdot)]} = \Lambda_{\beta}(\omega). \end{aligned}$$

So for μ-almost all $\omega \in \Omega$ and for $\beta \geqq \hat{\beta}$ we have that $\Lambda_{\beta}(\omega)$ is w-compact. Now observe that

$$\sup_{x \in X} [u(\omega, x) - \delta_{F(\omega)}(x)] = \sup_{x \in \Lambda_{\beta}(\omega)} [u(\omega, x) - \delta_{F(\omega)}(x)] = \sup_{x \in \Lambda_{\beta}(\omega)} u(\omega, x).$$

By the Weierstrass theorem we know that the above supremum is attained for μ-almost all $\omega \in \Omega$. Define

$$G(\omega) = \{\hat{x} \in F(\omega): u(\omega, \hat{x}) = m(\omega)\}$$

where $m(\omega) = \sup_{x \in F(\omega)} u(\omega, x)$. Using Castaing's representation theorem and the weak upper-semicontinuity of $u(\omega, \cdot)$, it is easy to see that $m(\cdot)$ is measurable. By redefining $G(\cdot)$ on a μ-null set (if necessary) we may assume that $G(\omega) \neq \varnothing$ for all $\omega \in \Omega$. Also note that $\operatorname{Gr} G = \{(\omega, \hat{x}) \in \Omega \times X: u(\omega, \hat{x}) = m(\omega)\} \cap \operatorname{Gr} F \in \Sigma \times B(X)$. Hence we can apply Aumann's measurable selection theorem to find $f: \Omega \to X$ measurable s.t. $f(\omega) \in G(\omega) \omega \in \Omega$. So $f(\cdot) \in S_F^1$ and

$$u(\omega, g(\omega)) \leqq u(\omega, f(\omega)) \quad \mu\text{-a.e.}$$

for all $g(\cdot) \in S_F^1$. Thus $f(\cdot)$ is an optimal allocation. Q.E.D.

The next result is another existence result. It provides a set of conditions on $u(\cdot, \cdot)$ and $F(\cdot)$ under which there exists an optimal allocation which is also price efficient. Such an allocation is sometimes called "competitive optimal allocation".

THEOREM 4.2. *If $F: \Omega \to P_{wkc}(X)$ is integrably bounded and $u: \Omega \times X \to \bar{\mathbb{R}}$ is such that*

(i) *$u(\cdot, \cdot)$ is a $\Sigma \times B(X)$-measurable,*

(ii) *$u(\omega, \cdot)$ is monotone increasing and concave for all $\omega \in \Omega$,*

(iii) *for every $g(\cdot) \in S_F^1 \left| \int_{\Omega} u(\omega, g(\omega))\, d\mu(\omega) \right| < +\infty$,*

(iv) *there exists $x^*(\cdot) \in L_{X^*}^{\infty}(\Omega)$ s.t. $\int_{\Omega} u^*(\omega, x^*(\omega))\, d\mu(\omega) > -\infty$,*

then there exists a competitive optimal allocation.

Remark. Here $u^*(\cdot, \cdot)$ denotes the concave conjugate of $u(\omega, \cdot)$.

Proof. From Lemma I of § 3 we know that S_F^1 is a w-compact subset of $L_X^1(\Omega)$. Also from hypothesis (iii) and using the results of Rockafellar [42, Thm. 21] applied in this case to concave functions, we get that

$$(I_u)^*(\cdot) = I_{u^*}(\cdot).$$

Furthermore hypothesis (iv) allows us to apply once more that result of Rockafellar and get that

$$(I_{u^*})^*(\cdot) = I_{u^{**}}(\cdot).$$

But recall that $u^{**}(\cdot) = u(\cdot)$. Hence

$$(I_u)^{**}(\cdot) = I_u(\cdot)$$

which means $I_u(\cdot)$ is continuous with respect to any topology on $L_X^1(\Omega)$ compatible with the duality $(L_X^1(\Omega), L_{X_{w^*}^*}^{\infty}(\Omega))$. So the Weierstrass theorem tells us that $I_u(\cdot)$ achieves its supremum on the w-compact set S_F^1. Let $\hat{f}(\cdot) \in S_F^1$ s.t.

$$I_u(\hat{f}) = \sup_{g(\cdot) \in S_F^1} I_u(g).$$

Also because of the monotonicity of $u(\omega, \cdot)$ for all $\omega \in \Omega$ we can easily see that

$$I_u(\hat{f}) = \sup_{g(\cdot) \in [S_F^1 - (L_X^1)_+]} I_u(g).$$

If we set

$$\hat{\delta}(g) = \begin{cases} 0 & \text{if } g \in S_F^1 - (L_X^1)_+, \\ -\infty & \text{otherwise,} \end{cases}$$

(it is easy to see that this is concave and u.s.c.) then from the well-known extremality condition of convex analysis (applied to concave functions and their "superdifferential" $\partial(\cdot)$) we have that

$$0 \in \partial[I_u + \tilde{\delta}](\hat{f}).$$

Now apply the Moreau-Rockafellar theorem [38] and get that

$$0 \in \partial I_u(\hat{f}) + \partial \tilde{\delta}(\hat{f}).$$

So there exists $p(\cdot) \in \partial I_u(\hat{f})$ s.t. $-p(\cdot) \in \partial \tilde{\delta}(\hat{f})$. This second fact implies that

$$\langle p, \hat{f} \rangle \geqq \langle p, g \rangle$$

for all $g(\cdot) \in [S_F^1 - (L_X^1)_+]$. Hence

$$\langle p, \hat{f} \rangle \geqq \langle p, g \rangle - \langle p, e \rangle$$

for all $g(\cdot) \in S_F^1$ and all $e(\cdot) \in (L_X^1)_+$. Take $g = \hat{f}$ to conclude that $p(\cdot) \in [L_{X_w^*}^\infty(\Omega)]_+$. Also from Rockafellar [42] we know that $p(\cdot) \in \partial I_u(\hat{f})$ implies that $p(\omega) \in \partial u(\omega, \hat{f}(\omega))$ μ-a.e. and so

$$(*) \qquad u(\omega, g(\omega)) - u(\omega, \hat{f}(\omega)) \leqq (p(\omega), g(\omega) - \hat{f}(\omega)) \quad \mu\text{-a.e.}$$

for all $g(\cdot) \in S_F^1$. Also it is easy to see that

$$\tilde{\delta}_{S_F^1}(g) = \int_\Omega \tilde{\delta}_{F(\omega)}(g(\omega))\, d\mu(\omega)\, g(\cdot) \in S_F^1.$$

Since $-p(\cdot) \in \partial \tilde{\delta}_{S_F^1}(\hat{f})$ we get that $-p(\omega) \in \partial \tilde{\delta}_{F(\omega)}(\hat{f}(\omega))$, u-a.e. which means that $(p(\omega), \hat{f}(\omega)) \geqq (p(\omega), g(\omega))$ μ-a.e. for all $g(\cdot) \in S_F^1$. Using this fact and (*) we conclude that

$$u(\omega, g(\omega)) \leqq u(\omega, \hat{f}(\omega)) \quad \mu\text{-a.e.}$$

for all $g(\cdot) \in S_F^1$. So $\hat{f}(\cdot)$ is a competitive optimal allocation. Q.E.D.

Now we will pass to utility (profit) functions $u(\cdot, \cdot)$ that are not concave.

We start with a result that gives us a necessary condition for the optimality of an allocation.

THEOREM 4.3. *If $F: \Omega \to P_f(X)$ is integrably bounded $u: \Omega \times X \to \mathbb{R}$ is an integrand s.t.*

(i) *$u(\omega, \cdot)$ is Lipschitz and increasing for all $\omega \in \Omega$,*

(ii) *$u(\cdot, x)$ is measurable for all $x \in X$,*

and if $f(\cdot) \in S_F^1$ is an optimal allocation then there exists $p(\cdot) \in [L_{X_{w^*}^*}^\infty(\Omega)]_+$ s.t. $p(\omega) \in \partial u(\omega, f(\omega))$ μ-a.e.

Remark. Here ∂ denotes Clarke's generalized subdifferential. For details we refer to Clarke [7], [9] and Rockafellar [43].

Proof. Because $f(\cdot) \in S_F^1$ is by hypothesis an optimal allocation we have that

$$u(\omega, g(\omega)) \leqq u(\omega, f(\omega)) \quad \mu\text{-a.e.}$$

for all $g(\cdot) \in S_F^1$. Since $u(\omega, \cdot)$ is by hypothesis increasing we have that

$$I_u(g) \leqq I_u(f)$$

for all $g(\cdot) \in S_F^1 - (L_X^1(\Omega))_+$. Hence

$$\sup_{g(\cdot) \in S_F^1 - (L_X^1)_+} I_u(g) = I_u(f).$$

Note that if $A = S_F^1 - (L_X^1)_+$ for some $k > 0$ large enough

$$\sup_{g(\cdot) \in A} I_u(g) = \sup_{g(\cdot) \in L_X^1} [I_u(g) - kd_A(g)] = I_u(f).$$

We know that $d_A(\cdot)$ is Lipschitz. Then from Clarke [9] we know that $0 \in \partial[I_u - kd_A](g)$. Recall that Clarke's generalized subdifferential is subadditive and $\partial(-kd_A)(\cdot) = -\partial kd_A(\cdot)$ (see [9]). So we get that

$$0 \in \partial I_u(f) - \partial kd_A(F) \Rightarrow \partial I_u(f) \cap \partial kd_A(f) \neq \varnothing.$$

Let $p(\cdot) \in \partial I_u(f) \cap \partial kd_A(f)$. Then we have that

$$\langle p, \tilde{g} - f \rangle \leqq kd_A^0(f; \tilde{g} - f)$$

for all $\tilde{g}(\cdot) \in L_X^1(\Omega)$. Fix $e(\cdot) \in (L_X^1)_-$, otherwise arbitrary. Then

$$\langle p, f + e - f \rangle = \langle p, e \rangle \leqq kd_A^0(f; e).$$

From Hiriart-Urruty [15] we know that

$$d_A^0(f; e) = \overline{\lim_{\substack{g' \to f \\ g' \in A \\ \lambda \downarrow 0}}} \frac{d_A(g' + \lambda e)}{\lambda} = 0 \Rightarrow \langle p, e \rangle \leqq 0$$

$$\Rightarrow p(\cdot) \in [L_{X_{w^*}^*}^{\infty}(\Omega)]_+.$$

Also from Clarke [9] we know that if $p(\cdot) \in \partial I_u(f)$ then $p(\omega) \in \partial u(\omega, f(\omega))$ μ-a.e. Q.E.D.

Remark. If $F(\cdot)$ is also convex valued, then S_F^1 is convex and so is $d_A(\cdot)$. In this case generalized and convex subdifferentials coincide and so $\partial kd_A(f) = N_A(f)$. This means that $\langle p, f \rangle = \sigma_A(p) = \sigma_{S_F^1}(p)$. So we obtain the following extension of Theorem 4.3.

COROLLARY. *If all hypotheses of Theorem* 4.3 *hold and* $F(\cdot)$ *is, in addition, convex valued then* $f(\cdot)$ *is a competitive optimal allocation.*

We will conclude this section with another necessary condition for optimality in the absence of concavity. Now take $u(\omega, \cdot)$ to be locally Lipschitz $\omega \in \Omega$.

THEOREM 4.4. *If* $u(\cdot, \cdot)$ *and* $F(\cdot)$ *are as in Theorem* 4.3 *and* $f(\cdot) \in S_F^1$ *is an optimal allocation then for all* $e \in X_+$, $u^0(f(\omega); e) \geqq 0$ μ-*a.e.*

Proof. Using the Castaing representation of $F(\cdot)$ and the continuity of $u(\omega, \cdot)$ we can easily see that for all $x \in F(\omega)$

$$u(\omega, x) \leqq u(\omega, f(\omega)) \quad \mu\text{-a.e.}$$

Let $v(\omega, x) = -u(\omega, x)$. Then for all $x \in F(\omega)$ we have that

$$v(\omega, x) \geqq v(\omega, f(\omega)) \quad \mu\text{-a.e.}$$

So $f(\omega)$ is μ-a.e. the minimum of the locally Lipschitz function $v(\omega, \cdot)$ on the set $F(\omega)$. Furthermore because of the monotonicity of $u(\omega, \cdot)$ we can say that $f(\omega)$ is

μ-a.e. the minimum of $v(\omega, \cdot)$ on $F(\omega) - X_+$. From Hiriart-Urruty [16] we deduce that

$$v^0(\omega, f(\omega); h) \geqq 0 \quad \mu\text{-a.e.}$$

for all $h \in \tau_{F(\omega)-X_+}(f(\omega))$ (where τ denotes Clarke's tangent cone (see [7], [16], [43])). So

$$(-u)^0(\omega, f(\omega); h) = u^0(\omega, f(\omega); -h) \geqq 0 \quad \mu\text{-a.e.}$$

Note that $X_- \subseteq \tau_{F(\omega)-X_+}(f(\omega))$. Hence for all $e \in X_+$ we have that

$$u^0(\omega, f(\omega); e) \geqq 0 \quad \mu\text{-a.e.} \qquad \text{Q.E.D.}$$

5. Approximate efficiency and optimality. Sometimes price efficient and optimal allocations may not exist or it may be difficult to obtain them. In cases like these our objective is to approximate as closely as possible those ideal situations. This motivated the introduction of the $\varepsilon(\cdot)$-optimal and the ε-price efficient allocations. In this section we examine such allocations.

Naturally enough we will start with an existence result. Clearly from their definition, ε-price efficient allocations always exist. Let us see what is the situation with $\varepsilon(\cdot)$-optimal allocations.

PROPOSITION 5.1. *If $u: \Omega \times X \to \bar{\mathbb{R}}$ is a $\Sigma \times B(X)$-measurable integrand and $F: \Omega \to P_f(X)$ is an integrably bounded multifunction then for any $\varepsilon(\cdot) \in [L^1(\Omega)]_+$ s.t. $\varepsilon(\omega) > 0$ μ-a.e. $\varepsilon(\cdot)$-optimal allocation always exist.*

Proof. Consider the following multifunction

$$G(\omega) = \{\bar{x} \in F(\omega): \sup_{x \in F(\omega)} u(\omega, x) - \varepsilon(\omega) \leqq u(\omega, \bar{x})\}.$$

Clearly $G(\omega) \neq \varnothing$. By appropriately redefining $G(\cdot)$ on the exceptional μ-null set we may assume that $G(\cdot)$ is nonempty and closed valued. Also if $m(\omega) = \sup_{s \in F(\omega)} u(\omega, x)$ then we have that $m(\omega) > \lambda$ if and only if there exists $x \in F(\omega)$ s.t. $u(\omega, x) > \lambda$. So we see that

$$\{\omega \in \Omega;\ m(\omega) > \lambda\} = \mathrm{pr}_\Omega [\{(\omega, x) \in \Omega \times X: u(\omega, x) > \lambda\} \cap \mathrm{Gr}\, F].$$

From Saint-Beuve's projection theorem [44] we know that $\mathrm{pr}_\Omega [\{(\omega, x) \in \Omega \times X: u(\omega, x) > \lambda\} \cap \mathrm{Gr}\, F] \in \Sigma$. So $m(\cdot)$ is Σ-measurable. Let $\psi(\omega, x) = m(\omega) - \varepsilon(\omega) - u(\omega, x)$. Clearly this integrand is $\Sigma \times B(X)$-measurable and also $G(\omega) = \{x \in F(\omega): \psi(\omega, x) \leqq 0\}$. So

$$\mathrm{Gr}\, G = \{(\omega, x) \in \Omega \times X: \psi(\omega, x) \leqq 0\} \cap \mathrm{Gr}\, F \in \Sigma \times B(X).$$

This allows us to apply Aumann's measurable selection theorem and get $f: \Omega \to X$ measurable s.t. $f(\omega) \in G(\omega) \omega \in \Omega$. Hence

$$m(\omega) - \varepsilon(\omega) \leqq u(\omega, f(\omega)) \quad \mu\text{-a.e.}$$

$$\Rightarrow u(\omega, g(\omega)) - \varepsilon(\omega) \leqq u(\omega, f(\omega)) \quad \mu\text{-a.e.}$$

for all $g(\cdot) \in S_F^1$. This means that $f(\cdot) \in S_F^1$ is $\varepsilon(\cdot)$-optimal. Q.E.D.

Now we will compare $\varepsilon(\cdot)$-optimal allocations with ε-price efficient allocations.

PROPOSITION 5.2. *If $u: \Omega \times X \to \mathbb{R}$ is a Caratheodory, concave, increasing integrand s.t. for all $x \in X |u(\omega, x)| \leqq \varphi(\omega)$ μ-a.e. where $\varphi(\cdot) \in L^1(\Omega)$ and if $F: \Omega \to P_{fc}(X)$ is integrably bounded then every $\varepsilon(\cdot)$-optimal allocation is ε'-price efficient where $\varepsilon' \leqq \int_\Omega \varepsilon(\omega)\, d\mu(\omega) = \varepsilon$.*

Proof. Since by hypothesis $f(\cdot) \in S_F^1$ is $\varepsilon(\cdot)$-optimal we have for all $g(\cdot) \in S_F^1$ that

$$u(\omega, g(\omega)) - u(\omega, f(\omega)) \leqq \varepsilon(\omega) \quad \mu\text{-a.e.}$$

$$\Rightarrow \int_\Omega u(\omega, g(\omega))\, d\mu(\omega) - \int_\Omega u(\omega, f(\omega))\, d\mu(\omega) \leqq \int_\Omega \varepsilon(\omega)\, d\mu(\omega)$$

$$\Rightarrow I_u(g) - I_u(f) \leqq \varepsilon.$$

Because $u(\omega, \cdot)$ is monotone increasing for all $\omega \in \Omega$, we can also say that

$$I_u(g') - I_u(f) \leqq \varepsilon$$

for all $g'(\cdot) \in S_F^1 - (L_X^1)_+$. This means that

$$0 \in \partial_\varepsilon (I_u + \tilde{\delta}_{S_F^1 - (L_X^1)_+})(f).$$

From Hiriart-Urruty [18, Thm. 2.1] we know that there exist $\varepsilon_1, \varepsilon_2 \geqq 0$ s.t. $\varepsilon_1 + \varepsilon_2 = \varepsilon$ and

$$0 \in \partial_{\varepsilon_1} I_u(f) + \partial_{\varepsilon_2} \tilde{\delta}(f).$$

Hence there exists $p(\cdot) \in \partial_{\varepsilon_1} I_u(f)$ s.t. $-p(\cdot) \in \partial_{\varepsilon_2} \tilde{\delta}(f)$. This last fact tells us that

$$\langle -p, g' \rangle - \varepsilon_2 \leqq \langle -p, f \rangle$$

for all $g'(\cdot) \in S_F^1 - (L_X^1)_+$. Let $e(\cdot) \in (L_X^1)_+$ be arbitrary and set $g' = f - e$. Then we have that

$$\langle -p, -e \rangle \leqq \varepsilon_2 \Rightarrow \langle p, e \rangle \leqq \varepsilon_2$$

$$\Rightarrow p(\cdot) \in -(L_X^1)_+.$$

Next observe that for all $g(\cdot) \in S_F^1$

$$\langle -p, g \rangle - \varepsilon_2 \leqq \langle -p, f \rangle, \quad (-p(\cdot) \in (L_X^1)_+)$$

which means that $f(\cdot)$ is ε_2-price efficient and $\varepsilon_2 \leqq \varepsilon$. Q.E.D.

The next result tells us that given an approximately optimal allocation and any number $\delta > 0$, we can find another, at least as good allocation, which is within a distance δ from the original one for the sup-norm.

THEOREM 5.1. *If $u: \Omega \times X \to \mathbb{R}$ is a Caratheodory integrand $F: \Omega \to P_f(X)$ is integrably bounded, $f(\cdot) \in S_F^1$ is $\varepsilon(\cdot)$-optimal and $\lambda > 0$ then we can find $\hat{f}(\cdot) \in S_F^1$ which is $\hat{\varepsilon}(\cdot)$-optimal with $\hat{\varepsilon}(\omega) \leqq \varepsilon(\omega)$ μ-a.e. and $\|\hat{f}(\omega) - f(\omega)\| \leqq \lambda$ μ-a.e.*

Proof. Consider the following multifunction

$$\Gamma(\omega) = \Big\{ x \in F(\omega) : u(\omega, f(\omega)) \leqq u(\omega, x), \|f(\omega) - x\| \leqq \lambda$$

$$u(\omega, z) < u(\omega, x) + \frac{\varepsilon(\omega)}{\lambda} \|x - z\| \text{ for all } z \neq x z \in F(\omega) \Big\}.$$

From the Ekeland variational principle [12] we know that for all $\omega \in \Omega$, $\Gamma(\omega) \neq \varnothing$. Also recalling that $\operatorname{Gr} F \in \Sigma \times B(X)$ and that $u(\cdot, \cdot)$ being a Caratheodory integrand is $\Sigma \times B(X)$-measurable, we can easily conclude that $\operatorname{Gr} \Gamma \in \Sigma \times B(X)$.

Apply Aumann's measurable selection theorem to find $\hat{f}: \Omega \to X$ measurable s.t. $\hat{f}(\omega) \in \Gamma(\omega)$, $\omega \in \Omega$. So we have that

$$u(\omega, z) - \varepsilon(\omega) \leqq u(\omega, f(\omega)) \leqq u(\omega, \hat{f}(\omega)) \quad \mu\text{-a.e.} \qquad z \in F(\omega), \tag{5.1}$$

$$\|\hat{f}(\omega)-f(\omega)\| \leqq \lambda \quad \mu\text{-a.e.}, \tag{5.2}$$

$$u(\omega, z)-\frac{\varepsilon(\omega)}{\lambda}\|z-\hat{f}(\omega)\| \leqq u(\omega, f(\omega)) \quad \text{for all } z \in F(\omega)\backslash\{\hat{f}(\omega)\}. \tag{5.3}$$

From (5.3) we get that

$$u(\omega, z)-\frac{\varepsilon(\omega)}{\lambda}\operatorname{diam} F(\omega) \leqq u(\omega, \hat{f}(\omega)). \tag{5.3'}$$

Note that diam $F(\omega)<\infty$ μ-a.e. Set

$$\hat{\varepsilon}(\omega)=\min\left[\varepsilon(\omega), \frac{\varepsilon(\omega)}{\lambda}\operatorname{diam} F(\omega)\right].$$

Using Castaing's representation it is easy to see that $\omega \to \operatorname{diam} F(\omega)$ is measurable. So $\omega \to \hat{\varepsilon}(\omega)$ is measurable. From (5.1), (5.2) and (5.3') we get that

$$\|\hat{f}(\omega)-f(\omega)\| \leqq \lambda \quad \mu\text{-a.e.}$$

and

$$u(\omega, z)-\hat{\varepsilon}(\omega) \leqq u(\omega, \hat{f}(\omega)) \quad \mu\text{-a.e.}$$

which tells us that $\hat{f}(\cdot) \in S_F^1$ is the desired allocation. Q.E.D.

Sometimes ε-price efficient allocations are called ε-profit maximizing allocations. Using this concept of ε-profit maximization we can have a necessary and sufficient condition for efficiency. Recall that a set $V \subseteq X$ is said to be X_+-convex if and only if $V-X_+$ is convex. Assume that int $X_+^* \neq \varnothing$ and denote by $E(F)$ the set of efficient allocations of $F(\cdot)$.

THEOREM 5.2. *Assume that $F:\Omega \to 2^X$ is an integrably bounded multifunction with nonempty, w-compact and X_+-convex values. Then we have: $f(\cdot) \in S_F^1$ is an efficient allocation if and only if for every $e(\cdot) \in [L_X^1(\Omega)]_+$ there exists a price system $p(\cdot) \in [L_{X^*}^\infty(\Omega)]_+$ s.t. $f(\cdot)$ is ε-price efficient for $p(\cdot)$ with $\varepsilon=\langle p, e\rangle$.*

Proof (*Necessity*). Suppose $f(\cdot) \in S_F^1$ is an efficient allocation.

Consider the multifunction $G(\omega)=F(\omega)-X_+$. Clearly this is closed, convex valued and measurable. Also S_G^1 is nonempty, closed and convex. Note that $E(F)=E(G)$. So $f(\cdot)$ is efficient for $G(\cdot)$. For $e(\cdot) \in [L_X^1(\Omega)]_+$ consider the function

$$g(\omega)=f(\omega)+e(\omega).$$

Because $f(\cdot) \in E(G)$, $g(\omega) \notin G(\omega)$ μ-a.e. So far $\omega \in \Omega\backslash N$ where $\mu(N)=0$, we can apply the second separation theorem. According to that there exists $p \in B_1^{X^*}(0)=\{x^* \in X^*: \|x^*\| \leqq 1\}$, $p \neq 0$ depending on $\omega \in \Omega$ s.t. $\sigma_{G(\omega)}(p)<(p, g(\omega))$. From the nature of $G(\cdot)$ we get that $p \in \dot{X}_+ = X_+\backslash\{0\}$. Consider the multifunction

$$\Gamma(\omega)=\{p \in B_1^{X^*}(0) \cap X_+^*: \sigma_{G(\omega)}(p)<(p, g(\omega))\}.$$

For $\omega \in \Omega\backslash N$ we know that $\Gamma(\omega) \neq \varnothing$. For $\omega \in N$ let $\Gamma(\omega)=\{0\}$. Then for all $\omega \in \Omega$ $\Gamma(\omega) \neq \varnothing$. Define $\varphi(\omega, p)=\sigma_{G(\omega)}(p)-(p, g(\omega))$. Observe that $\varphi(\omega, \cdot)$ is l.s.c. for all $\omega \in \Omega$ and $\varphi(\cdot, p)$ is measurable for all $p \in B_1^{X^*}(0)$. Furthermore dom $\varphi(\omega, \cdot) \supseteq B_1^{X^*}(0) \cap X_+^*$ and so int (dom $\varphi(\omega, \cdot)) \supseteq$ int $(B_1^X(0) \cap X_+^*)=$ int $B_1^X(0) \cap$ int $X_+^* \neq \varnothing$. Hence from Corollary 2E of Rockafellar [40] we deduce that $\varphi(\cdot, \cdot)$ is a $\Sigma \times B(X^*)$-measurable function. Next observe that

$$\operatorname{Gr} \Gamma=\{(\omega, p) \in \Omega \times B_1^{X^*}: \varphi(\omega, p)<0\} \cup (N \times \{0\}).$$

So $\operatorname{Gr}\Gamma \in \Sigma \times B(X^*)$. This means that we can find $P: \Omega \to B_1^{X^*} \cap X_+^*$ s.t. $p(\omega) \in \Gamma(\omega)$ $\omega \in \Omega$. Hence

$$(p(\omega), g(\omega)) > \sigma_{G(\omega)}(p(\omega)) \quad \mu\text{-a.e.}$$

$$\Rightarrow \int_\Omega (p(\omega), g(\omega))\, d\mu(\omega) > \int_\Omega \sigma_{G(\omega)}(p(\omega))\, d\mu(\omega)$$

$$\Rightarrow \int_\Omega (p(\omega), f(\omega))\, d\mu(\omega) + \int_\Omega (p(\omega), e(\omega))\, d\mu(\omega) > \int_\Omega \sigma_{G(\omega)}(p(\omega))\, d\mu(\omega).$$

Since $p(\cdot) \in L_{X^*}^\infty(\Omega))_+$ from the definition of $G(\cdot)$ we see that

$$\sigma_{G(\omega)}(p(\omega)) = \sigma_{F(\omega)}(p(\omega)) \omega \in \Omega.$$

Furthermore we know that

$$\int_\Omega \sigma_{F(\omega)}(p(\omega))\, d\mu(\omega) = \sigma_{S_F^1}(p).$$

So finally we have that

$$\langle p, f \rangle > \sigma_{S_F^1}(p) - \varepsilon$$

where $\varepsilon = \langle p, e \rangle = \int_\Omega (p(\omega), e(\omega))\, d\mu(\omega)$ which shows that indeed $f(\cdot)$ is ε-price efficient with $\varepsilon = \langle p, f \rangle$.

Sufficiency. Suppose that $f(\cdot)$ was not an efficient allocation. This means that there exists $A \in \Sigma$ with $\mu(A) > 0$ s.t. for $\omega \in A$

$$G(\omega) = [f(\omega) + X_+] \cap F(\omega) \neq \varnothing.$$

Clearly $\operatorname{Gr} G \in (\Sigma \cap A) \times B(X)$. So we can find $g: A \to X$ measurable s.t. $g(\omega) \in G(\omega)$ $\omega \in A$.

Define

$$\hat{g}(\omega) = \begin{cases} g(\omega) & \text{for } \omega \in A, \\ f(\omega) & \text{for } \omega \in \Omega \backslash A. \end{cases}$$

Clearly $\hat{g}(\cdot) \in S_F^1$ and $\hat{g} > f$. Let $e(\omega) = \hat{g}(\omega) - f(\omega)$. Then $e(\cdot) \in (L_X^1)_+ \backslash \{0\}$. So we can find $p(\cdot) \in (L_{X^*}^\infty)_+$ s.t.

$$\langle p, f \rangle > \sigma_{S_F^1}(p) - \langle p, e \rangle \Rightarrow \langle p, f \rangle > \langle p, \hat{g} \rangle - \langle p, e \rangle$$

$$\Rightarrow \langle p, f \rangle > \langle p, \hat{g} - e \rangle = \langle p, f \rangle$$

a contradiction. So $f(\cdot)$ has to be efficient. Q.E.D.

6. Stability results. In this final section of our work we examine the variation of the sets of efficient and optimal allocations as we perturb the data on which those notions depend.

We will start with a stability result for the set of efficient allocations $E(F)$, when the consumption (production) multifunction varies in the Kuratowski-Mosco sense described in § 2. For that purpose we need the following lemma. Assume that $\operatorname{int} X_+^* \neq \varnothing$.

LEMMA III. *If $F: \Omega \to P_{fc}(X)$ is integrably bounded and S_F^1 is locally w-compact then $E(F) - (L_X^1)_+ = S_F^1 - (L_X^1)_+$.*

Proof. Clearly $E(F) - (L_X^1)_+ \subseteq S_F^1 - (L_X^1)_+$. Let $g(\cdot) \in S_F^1 - (L_X^1)_+$. Then we have that $g(\cdot) \in g'(\cdot) - (L_X^1)_+$ where $g'(\cdot) \in S_F^1$. Consider the set $(g' + (L_X^1)_+) \cap S_F^1$ and

let $z(\cdot)\in E((g'+(L^1_X)_+)\cap S^1_F)$. Such an element exists by Theorem 3.1. We claim that $z(\cdot)\in E(F)$. Suppose not. Then $(z(\cdot)+(L^1_X)_+)\in S^1_F\neq\varnothing$. So there are $\hat{g}(\cdot)\in S^1_F$ and $e(\cdot)\in(L^1_X)_+$ s.t. $\hat{g}(\cdot)=z(\cdot)+e(\cdot)$. But note that $z(\cdot)=g'(\cdot)+e'(\cdot)$ for some $e'(\cdot)\in(L^1_X)_+$. So $\hat{g}(\cdot)=g'(\cdot)+e'(\cdot)+e(\cdot)\in g'(\cdot)+(L^1_X)_+$. Therefore $z(\cdot)\notin E((g'+(L^1_X)_+)\in S^1_F)$ a contradiction. Hence $z(\cdot)\in E(F)$ and $z(\cdot)=g'(\cdot)+e''(\cdot)(e''(\cdot)=e'(\cdot)+e(\cdot))$. Thus $z(\cdot)=g(\cdot)+e''(\cdot)+e'''(\cdot)(e'''(\cdot)\in(L^1_X)_+)$ which means that $z(\cdot)-e''(\cdot)-e'''(\cdot)-g\in E(F)-(L^1_X)_+$. Q.E.D.

Now we are ready for the stability result we promised.

THEOREM 6.1. *If F_n, $F:\Omega\to P_{fc}(X)$ are measurable multifunctions for all $n\geqq 1$ and $F_n(\omega)\subseteq W(\omega)$ μ-a.e. with $W:\Omega\to P_{wkc}(X)$ integrably bounded and if $F_n(\omega)\xrightarrow{K\text{-}M}F(\omega)$ μ-a.e. then*

$$E(F)\subseteq w\text{-}\limsup_{n\to\infty}E(F_n).$$

Proof. Let $f(\cdot)\in E(f)$. Then $f(\cdot)\in S^1_F$ and so $f(\omega)\in F(\omega)$ μ-a.e. Since by hypothesis

$$F_n(\omega)\xrightarrow{K\text{-}M}F(\omega)\quad\mu\text{-a.e.}$$

we know that

$$\|f(\omega)-F_n(\omega)\|\to 0\quad\mu\text{-a.e.}$$

as $n\to\infty$. Let $G_n(\omega)=\{x_n\in F_n(\omega)\colon\|f(\omega)-x_n\|=\|f(\omega)-F_n(\omega)\|\}$. Clearly for all $n\geqq 1$, $\omega\to G_n(\omega)$ is nonempty, closed valued and measurable. Apply the Kuratowski-Ryll Nardzewski measurable selection theorem to find $g_n:\Omega\to X$ measurable s.t. $g_n(\omega)\in G_n(\omega)$ for all $\omega\in\Omega$ and all $n\geqq 1$. Then $g_n(\cdot)\in S^1_{F_n}$ for all $n\geqq 1$ and

$$\|f(\omega)-g_n(\omega)\|\to 0\quad\mu\text{-a.e.}$$

as $n\to\infty$. From Lemma III we know that

$$g_n(\cdot)=f_n(\cdot)-e_n(\cdot)$$

where $f_n(\cdot)\in E(F_n)$ and $e_n(\cdot)\in(L^1_X)_+(n\geqq 1)$. Then

$$f_n(\omega)-e_n(\omega)\to f(\omega)\quad\mu\text{-a.e.}$$

as $n\to\infty$. Applying Lebesgue's dominated convergence theorem for Bochner integrals we get that

$$f_n-e_n\xrightarrow{s\text{-}L^1_X(\Omega)}f$$

as $n\to\infty$. Note that $S^1_{F_n}\subseteq S^1_W$ for all $n\geqq 1$ and from Lemma I we know that S^1_W is w-compact in $L^1_X(\Omega)$. From the Eberlein-Smulian theorem we can find a subsequence $\{n_k\}\subseteq\{n\}$ s.t.

$$f_{n_k}\xrightarrow{w\text{-}L^1_X}f'\in S^1_F\quad\text{as } k\to\infty.$$

Also

$$f_{n_k}-e_{n_k}\xrightarrow{s\text{-}L^1_X}f\quad\text{as } k\to\infty.$$

Then for all $p(\cdot)\in L^\infty_{X^*}(\Omega)$ we have that

$$\langle p, e_{n_k}+f-f'\rangle-\langle p, e_{n_k}-f_{n_k}+f\rangle+\langle p, f_{n_k}-f'\rangle\to 0$$

as $k \to \infty$. So we deduce that

$$e_{n_k} \xrightarrow{w\text{-}L^1_X} f' - f \quad \text{as } k \to \infty.$$

Since $(L^1_X)_+$ is closed, convex we get that $f' \geqq f$. Since $f(\cdot) \in E(F)$ and $f'(\cdot) \in S^1_F$ we must have $f' = f$. Hence

$$f(\cdot) \in w\text{-}\limsup_n E(F_n)$$

and since $f(\cdot) \in E(F)$ was arbitrary we conclude that

$$E(F) \subseteq w\text{-}\limsup_{n \to \infty} E(F_n). \qquad \text{Q.E.D.}$$

Now we pass to approximately price efficient allocations. In the next result we prove that the set of approximately price efficient allocations is continuous in the Hausdorff metric with respect to both the level of approximation $\varepsilon > 0$ and the efficient price system $p(\cdot)$.

Let X be reflexive (always separable). If $A \subseteq L^1_X(\cdot)$ then by $\partial_\varepsilon \sigma_A(\cdot)$ we will mean the $L^1_X(\Omega)$-subgradients of σ_A: $[L^1_X(\Omega)]^* = L^\infty_{X^*}(\Omega) \to \bar{\mathbb{R}}$. So we will be interested only in the absolutely continuous parts of the ε-subgradients in $\partial^*_\varepsilon \sigma_A(\cdot) \subseteq [L^\infty_{X^*}(\Omega)]^*$. We will denote the set of ε-price efficient allocations, with approximately efficient price system $p(\cdot)$ by $PE_F(\varepsilon, p)$. Then we have

PROPOSITION 6.1. *If $F: \Omega \to P_{fc}(X)$ is integrably bounded $\{\varepsilon_n, \varepsilon\}_{n \geqq 1} \subseteq \mathbb{R}_+ \setminus \{0\}$, $\{p_n, p\}_{n \geqq 1} \subseteq L^\infty_{X^*}(\Omega) \setminus \{0\}$ $\varepsilon_n \to \varepsilon$ and $p_n \xrightarrow{s\text{-}L^\infty_X} p$ as $n \to \infty$ then $PE_F(\varepsilon_n, p_n) \xrightarrow{h} PE_F(\varepsilon, p)$ as $n \to \infty$ (here $h(\cdot, \cdot)$ denote the Hausdorff metric).*

Proof. From the definition of approximate price efficiency and [8, Lemma 2] we can easily deduce that

$$PE_F(\varepsilon_n, p_n) = \partial_{\varepsilon_n}[\langle p_n, \cdot \rangle - d_{S^1_F}(\cdot)], \qquad PE_F(\varepsilon, p) = \partial_\varepsilon[\langle p, \cdot \rangle - d_{S^1_F}(\cdot)].$$

Both sets are nonempty, convex, closed in $L^1_X(\cdot)$. We know that $h(\cdot, \cdot)$ defines a metric on those sets. Also recall that since S^1_F is convex $d_{S^1_F}(\cdot)$ is convex. So $-d_{S^1_F}(\cdot)$ is a concave, Lipschitz function. Hence from Hiriart-Urruty [17] we know that for all $n \geqq 1$

$$h(\partial_{\varepsilon_n} d\langle p_n, \cdot \rangle - d_{S^1_F}(\cdot)], \partial_\varepsilon[\langle p, \cdot \rangle - d_{S^1_F}(\cdot)]) \leqq \frac{1}{\min(\varepsilon_n, \varepsilon)}[\|p_n - p\|_\infty + |\varepsilon_n - \varepsilon|]$$

$$\Rightarrow h(PE_F(\varepsilon_n, p_n), PE_F(\varepsilon, p)) \leqq \frac{1}{\min(\varepsilon_n, \varepsilon)}[\|p_n - p\|_\infty + |\varepsilon_n - \varepsilon|].$$

Let $n \to \infty$. We conclude that $PE_F(\varepsilon_n, p_n) \xrightarrow{h} PE_F(\varepsilon, p)$. Q.E.D.

So the above result tells us that: the nonempty, closed, convex valued multifunction $PE_F(\cdot, \cdot)$ is Hausdorff continuous on $L^1_X(\Omega) \times \dot{\mathbb{R}}_+ (\dot{\mathbb{R}}_+ = \mathbb{R}_+ \setminus \{0\})$ with the strong product topology.

When $\varepsilon = 0$, the result of Hiriart-Urruty [17] on which the proof of the previous proposition was based is not any more applicable. So in that case we cannot have any more convergence in the Hausdorff metric. However we can prove convergence in the Kuratowski-Mosco sense for a fixed price system $p(\cdot)$.

PROPOSITION 6.2. *If $F: \Omega \to P_{fc}(X)$ is integrably bounded, $p(\cdot) \in [L^\infty_{X^*}(\Omega)]_+ \setminus \{0\}$ and $\varepsilon_n \downarrow 0$ as $n \to \infty$ then $PE_F(\varepsilon_n, p) \xrightarrow{K\text{-}M} PE_F(\varepsilon, p)$ as $n \to \infty$.*

Proof. Let $f_m(\cdot) \in PE_F(\varepsilon_m, p)$ $m \in M \subseteq N$ and suppose that $f_m \xrightarrow{w\text{-}L^1_X} f$. Since $f_m(\cdot) \in S^1_F$ and S^1_F is closed and convex in $L^1_X(\Omega)$ we get that $f(\cdot) \in S^1_F$.

By definition for all $m \in M$ we have that

$$\sigma_{S_F^1}(p) - \varepsilon_m \leqq \langle p, f_m \rangle.$$

Passing to the limit as $m \to \infty$ we get that

$$\sigma_{S_F^1}(p) \leqq \langle p, f \rangle$$

and since $f(\cdot) \in S_F^1$ we get that $\sigma_{S_F^1}(p) = \langle p, f \rangle$. So $f(\cdot) \in PE_F(p)$. Hence we see that

$$w\text{-}\limsup_{n \to \infty} PE_F(\varepsilon_n, p) \subseteq PE_F(p). \tag{6.1}$$

Next let $f(\cdot) \in PE_F(p)$. Then for all $n \geqq 1$ $f(\cdot) \in PE_F(\varepsilon_n, p)$. Applying Ekeland's variational principle we can find $f_n(\cdot) \in S_F^1$ s.t.

$$\sigma_{S_F^1}(p) - \varepsilon_n \leqq \langle p, f_n \rangle, \tag{6.2}$$

$$\|f_n - f\|_1 \leqq \sqrt{\varepsilon_n}. \tag{6.3}$$

From (2) we have that $f_n(\cdot) \in PE_F(\varepsilon_n, p)$ and from (6.3) we get that

$$f_n \xrightarrow{s\text{-}L_X^1} f \quad \text{as } n \to \infty.$$

Hence we deduce that

$$PE_F(p) \subseteq s\text{-}\liminf_{n \to \infty} PE_F(\varepsilon_n, p). \tag{6.4}$$

From (6.1) and (6.4) above we finally conclude that

$$PE_F(\varepsilon_n, p) \xrightarrow{\text{K-M}} PE_F(\varepsilon, p) \quad \text{as } n \to \infty. \qquad \text{Q.E.D.}$$

In the remaining part of this section we examine the behavior of optimality, price efficiency and efficiency for a sector as the sector gets larger.

So let $\{\Sigma_n\}_{n \geqq 1}$ be a sequence of sub-σ-fields of Σ and suppose that $\bigvee_{n=1} \Sigma_n = \Sigma$. Also assume that X is finite-dimensional. Denote by $A(u, F)$ the set of all optimal allocations determined by the consumption (production) multifunction $F(\cdot)$ and the utility (profit) functions $u(\cdot, \cdot)$. Then we have

THEOREM 6.2. *If $u: \Omega \times X \to \mathbb{R}$ is a Caratheodory integrand s.t. $|u(\omega), x)| \leqq \varphi(\omega)$ μ-a.e. for all $x \in X$ where $\varphi(\cdot) \in L^1(\omega)$ and if $F: \Omega \to P_{fc}(X)$ is integrably bounded then*

$$s\text{-}\limsup_{n \to \infty} A(E^{\Sigma_n} u, E^{\Sigma_n} F) \subseteq A(u, F).$$

Proof. Let

$$f(\cdot) \in s\text{-}\limsup_{n \to \infty} A(E^{\Sigma_n} u, E^{\Sigma_n} F).$$

Then this means that there exist $r_m(\cdot) \in A(E^{\Sigma_m} u, E^{\Sigma_m} F)$, $m \in M \subseteq N$ s.t. $r_m \xrightarrow{s\text{-}L_X^1} f$ as $m \to \infty$.

Recall that $r_m(\cdot) \in S^1(\Sigma_m)$. From Hiai-Umegaki [13] (see also Valadier [48]) we know that $S^1(\Sigma_m) = \text{cl}\{E^{\Sigma_m} S_F^1\}$ with the closure taken in the $L_n^1(\Omega)$-norm. Since X is finite-dimensional, from Lemma I of § 3 we know that S_F^1 is a w-compact, convex subset of $L_X^1(\Omega)$. Hence $S^1(\Sigma_m) = E^{\Sigma_m} S_F^1$. So $r_m = E^{\Sigma_m} f_m$ where $f_m(\cdot) \in S_F^1$ $m \in M$. By passing from the beginning to a subsequence, if necessary, we may assume that

$$E^{\Sigma_m} f_m(\omega) \to f(\omega) \quad \mu\text{-a.e.}$$

as $m \to \infty$. Since $r_m(\cdot) \in A(E^{\Sigma_m} u, E^{\Sigma_m} g)$ and $S^1(\Sigma_m) = E^{\Sigma_m} S_F^1$ we have that for every

$g(\cdot) \in S_F^1$

$$E^{\Sigma_m} u(w, E^{\Sigma_m} g(\omega)) \leqq E^{\Sigma_m} u(\omega, E^{\Sigma_m} f(\omega)) \quad \mu\text{-a.e.}$$

Since $F(\cdot)$ is integrably bounded from martingale theory we know that $E^{\Sigma_m} g(\omega) \to g(\omega)$ μ-a.e. as $m \to \infty$. Furthermore from Bismut [4, Thm. 1, p. 668] we know that

$$E^{\Sigma_m} u(\omega, \cdot) \to u(\omega, \cdot) \quad \mu\text{-a.e.}$$

and the convergence is uniform on compact sets.

Hence we deduce that

$$E^{\Sigma_m} u(\omega, E^{\Sigma_m} g(\omega)) \to u(\omega, g(\omega)) \quad \mu\text{-a.e.}$$

and

$$E^{\Sigma_m} u(\omega, E^{\Sigma_m} f(\omega)) \to u(\omega, f(\omega)) \quad \mu\text{-a.e.}$$

Since for all $m \in M \subseteq N$, $E^{\Sigma_m} u(\omega, E^{\Sigma_m} g(\omega)) \leqq E^{\Sigma_m} u(\omega, E^{\Sigma_m} f(\omega))$ μ-a.e. we conclude that

$$u(\omega, g(\omega)) \leqq u(\omega, f(\omega)) \quad \mu\text{-a.e.}$$

and since this is true for all $g(\cdot) \in S_F^1$, we have that $f(\cdot) \in A(u, F)$. Q.E.D.

We have an analogous stability result for price efficient allocations. For that purpose assume that X is finite-dimensional and Σ is a countably generated σ-field. Then we have

THEOREM 6.3. *If $F: \Omega \to P_{fc}(X)$ is integrably bounded then*

$$w\text{-}\limsup_{n \to \infty} PE(E^{\Sigma_n} F) \subseteq PE_1(F).$$

Proof. Let

$$f(\cdot) \in w\text{-}\limsup_{n \to \infty} PE(E^{\Sigma_n} F).$$

This means that there exist

$$g_m(\cdot) \in PE(E^{\Sigma_m} F) m \in M \subseteq N \text{ s.t. } g_m \xrightarrow{w\text{-}L_n^1(\Omega)} f \quad \text{as } m \to \infty.$$

So $g_m(\cdot) \in S^1(\Sigma_m)$ and there exist $p_m(\cdot) \in B^{L_n^\infty(\Omega, \Sigma_m)}(0)$

$$\langle p_m, g_m \rangle = \sigma_{S^1(\Sigma_m)}(p_m) m \in M.$$

Recall that

$$\sigma_{S^1(\Sigma_m)}(p_m) = \int_\Omega \sigma_{\Sigma_m(\omega)}(p_m(\omega))\, d\mu(\omega).$$

From Valadier [48] we know that $\sigma_{\Sigma_m(\omega)}(p_m(\omega)) = E^{\Sigma_m} \sigma_{F(\omega)}(p_m(\omega))$ μ-a.e. But then from Bismut [4] we get that

$$\int_\Omega E^{\Sigma_n} \sigma_{F(\omega)}(p_n(\omega))\, d\mu(\omega) = \int_\Omega \sigma_{F(\omega)}(p_n(\omega))\, d\mu(\omega) = \sigma_{S_F^1}(p_n).$$

Because Σ is countably generated, $L_n^1(\Omega)$ is separable and so $B_1^{L_n^\infty(\Omega)}(0)$ is metrizable with the w^*-topology and so w^*-sequentially compact. Hence we can find

$$\{m_k\} \subseteq \{m\} \text{ s.t. } p_{m_k} \xrightarrow{w^*\text{-}L_n^\infty} p \quad \text{as } k \to \infty$$

and $p(\cdot) \in B^{L_n^\infty(\Omega)}(0)$. At this point use Mazur's theorem to find

$$z_{m_k}(\cdot) \in \text{conv} \bigcup_{n \geq k} g_{m_n}(\cdot) \text{ s.t. } z_{m_k}(\cdot) \xrightarrow{s\text{-}L_n^1} f(\cdot) \quad \text{as } k \to \infty.$$

So we have that

$$\langle p_{m_k}, z_{m_k} \rangle \to \langle p, f(\cdot) \rangle \quad \text{as } k \to \infty. \tag{6.5}$$

Also note that since S_F^1 is bounded, $\sigma_{S_F^1}(\cdot)$ is finite and so because of convexity we conclude that it is continuous on $(B^{L_n^\infty(\Omega)}(0), w^*)$. Hence we have that

$$\sigma_{S^1(\Sigma_m)}(p_{m_k}) = \sigma_{S_k^1}(p_{m_k}) \to \sigma_{S_F^1}(p) \quad \text{as } k \to \infty. \tag{6.6}$$

Since $\sigma_{S^1(\Sigma_m)}(p_{m_k}) = \langle p_{m_k}, z_{m_k} \rangle$ from (6.5) and (6.6) above we conclude that $\langle p, f \rangle = \sigma_{S_F^1}(p)$. Because S_F^1 is a closed, convex subset of $L_n^1(\Omega)$, Hormander's theorem tells us that $f(\cdot) \in S_F^1$. Therefore, $f(\cdot) \in PE(F)$. Q.E.D.

Remark. A careful examination of the above proof can convince the reader that if X is a general separable Banach space then

$$s\text{-}\limsup_{n \to \infty} PE(E^{\Sigma_n} F) \subseteq PE(F).$$

We will conclude the paper with a similar stability analysis for the set $E(F)$ of efficient allocations. Assume that X is finite-dimensional Banach space.

THEOREM 6.4. *If $F: \Omega \to P_{fc}(X)$ is integrably bounded by $\varphi(\cdot) \in L^1(\Omega, \Sigma_1)$ then $E(F) \subseteq w\text{-}\limsup_{n \to \infty} E(E^{\Sigma_n} F)$.*

Proof. From Valadier [48] we know that for all $x^* \in X^*$ we have

$$\sigma_{E^{\Sigma_n} F(\omega)}(x^*) = E^{\Sigma_n} \sigma_{F(\omega)}(x^*) \quad \mu\text{-a.e.}$$

the exceptional set independent of x^* because of continuity. From Bismut [4] we know that for all $x^* \in X^*$

$$E^{\Sigma_n} \sigma_{F(\omega)}(x^*) \to \sigma_{F(\omega)}(x^*) \quad \mu\text{-a.e.}$$

As in the proof of Theorem 6.2 we can get that $S^1_{E^{\Sigma_n} F} = E^{\Sigma_n} S_F^1$. Hence

$$|E^{\Sigma_n} F(\omega)| = \sup_{f \in S_F^1} \|E^{\Sigma_n} f(\omega)\| \leq \sup_{f \in S_F^1} E^{\Sigma_n} \|f(\omega)\| \leq E^{\Sigma_n} \varphi(\omega) = \varphi(\omega) < +\infty \ \mu\text{-a.e.}$$

So we can write that

$$\text{dom } \sigma_{E^{\Sigma_n} F(\omega)}(\cdot) = \text{dom } \sigma_{F(\omega)}(\cdot) = X^* \quad \mu\text{-a.e.}$$

Then Corollary 2E of Salinetti-Wets [45] tells us that

$$\sigma_{E^{\Sigma_n} F(\omega)}(\cdot) = E^{\Sigma_n} \sigma_{F(\omega)}(\cdot) \xrightarrow{\tau} \sigma_{F(\omega)}(\cdot) \quad \mu\text{-a.e.}$$

as $n \to \infty$. But then from Theorem 3.1 of Mosco [27] we deduce that $E^{\Sigma_n} F \xrightarrow{sK} F$ as $n \to \infty$. Now recall that for all $n \geq 1$ we have $|E^{\Sigma_n} F(\omega)| \leq \varphi(\omega)$ μ-a.e. So

$$|E^{\Sigma_n} F(\omega)| \leq B^X_{\varphi(\omega)}(0) = \{x \in X : \|x\| \leq \varphi(\omega)\}.$$

It is easy to see that $\omega \to B^X_{\varphi(\omega)}(0)$ is integrably bounded. Therefore we can apply Theorem 6.1 and get that

$$E(F) \subseteq w\text{-}\limsup_{n \to \infty} E(E^{\Sigma_n} F). \qquad \text{Q.E.D.}$$

If we impose more restrictions on $F(\cdot)$ we can have a stronger version of Theorem 6.4. Assume that X is a reflexive Banach space (always separable).

THEOREM 6.5. *If $F: \Omega \to P_{fc}(X)$ is measurable and $F(\omega) \subseteq W$ μ-a.e. with W a weakly compact, convex subset of X then*

$$E(E^{\Sigma_n}F) \overset{sK}{\to} E(F) \quad \text{as } n \to \infty.$$

Proof. From Theorem 6.1 of Hiai-Umegaki [13] we have that

$$h(E^{\Sigma_n}F(\omega), F(\omega)) \to 0 \quad \mu\text{-a.e.} \quad \text{as } n \to \infty.$$

Also since $F(\omega) \subseteq W$ μ-a.e., we have that $E^{\Sigma_n}F(\omega) \subseteq E^{\Sigma_n}W = W$ μ-a.e. Hence we can apply Theorem 5.4 of [28] to conclude that

$$E(E^{\Sigma_n}F) \overset{sK}{\to} E(F) \quad \text{as } n \to \infty. \qquad \text{Q.E.D.}$$

Concluding our work, we would like to point out that it will be interesting to consider also risk-aversely efficient allocations (see [33]), study their properties and compare them with the notions introduced and studied in this paper. Risk-aversely efficient allocations are important in economic theory.

Acknowledgment. The author is very grateful to the referee for reading the paper carefully and making several comments and suggestions that improved the presentation considerably.

REFERENCES

[1] R. AUMANN, *Markets with a continuum of traders*, Econometrica, 32 (1964), pp. 39-50.
[2] ———, *Integrals of set valued functions*, J. Math. Anal. Appl., 12 (1965), pp. 1-12.
[3] M. BENAMARA, *Sections extremales d'une multiapplication*, CRAS Paris, 278 (1974), pp. 1249-1252.
[4] J. M. BISMUT, *Intégrales convèxes et probabilités*, J. Math. Anal. Appl., 42 (1973), pp. 639-73.
[5] J. BORWEIN, *Proper efficient points for maximizations with respect to cones*, SIAM J. Control Optim., 15 (1977), pp. 57-63.
[6] C. CASTAING AND M. VALADIER, *Convex analysis and measurable multifunctions*, Lecture Notes in Mathematics 560, Springer, Berlin, 1977.
[7] F. CLARKE, *Generalized gradients and applications*, Trans. Amer. Math. Soc., 205 (1975), pp. 247-263.
[8] ———, *A new approach to Lagrange multipliers*, Math. Oper. Res., 1 (1976), pp. 165-174.
[9] ———, *Generalized gradients of Lipschitz functionals*, Adv. Math., 40 (1981), pp. 52-67.
[10] J. DIESTEL AND J. J. UHL, *Vector measures*, Math. Surveys, Vol. 15, American Mathematical Society, Providence (1977).
[11] N. DUNFORD AND J. SCHWARTZ, *Linear Operators, Part* I, Wiley-Interscience, New York, 1957.
[12] J. R. GILES, *Convex Analysis with Applications in Differentiation of Convex Functions*, Research Notes in Mathematics 58, Pitman, Boston, 1982.
[13] F. HIAI AND H. UMEGAKI, *Integrals, conditional expectations and martingales of multivalued functions*, J. Mult. Anal., 7 (1977), pp. 149-182.
[14] C. HIMMELBERG, *Measurable relations*, Fund. Math., 87 (1975), pp. 53-72.
[15] J. B. HIRIART-URRUTY, *On optimality conditions in nondifferentiable programming*, Math. Programming, 14 (1978), pp. 73-86.
[16] ———, *Tangent cones, generalized gradients and math programming in Banach spaces*, Math. Oper. Res., 4 (1979), pp. 79-97.
[17] ———, *Lipschitz r-continuity of the approximate subdifferential of a convex function*, Math. Scand., 47 (1980), pp. 123-134.
[18] ———, *ε-subdifferential calculus*, in Convex Analysis and Optimization, P. Aubin, R. Vinter, eds., Pitman, Boston, 1982.
[19] R. HOLMES, *Geometric Functional Analysis and Applications*, Graduate Texts in Mathematics, Vol. 27, Springer, Berlin, 1975.
[20] A. IONESCU TULCEA AND C. IONESCU TULCEA, *Topics in the Theory of Lifting*, Ergebnisse Math. Grenzgebiete, Band 48, Springer, Berlin, 1969.
[21] K. KURATOWSKI AND C. RYLL-NARDZEWSKI, *A general theorem on selectors*, Bull. Acad. Polon. Sci., 13 (1965), pp. 397-403.

[22] V. LEVIN, *Lebesgue decomposition for functionals on the vector space L_X^∞*, Functional Anal. Appl., 8 (1974), pp. 48-53.
[23] M. MAJUMDAR, *Some general theorems on efficiency prices with an infinite dimensional commodity space*, J. Econom. Theory, 5 (1972), pp. 1-13.
[24] ———, *Efficient programs in infinite dimensional spaces, a complete characterization*, J. Econom. Theory, 7 (1974), pp. 355-369.
[25] J. J. MOREAU, *Intersection of moving convex sets in normed spaces*, Math. Scand., 36 (1975), pp. 159-173.
[26] U. MOSCO, *Convergence of convex sets and of solutions of variational inequalities*, Adv. Math, 3 (1969), pp. 510-585.
[27] ———, *On the continuity of the Young-Fenchel transform*, J. Math. Anal. Appl., 35 (1971), pp. 518-535.
[28] N. S. PAPAGEORGIOU, *On the efficiency and optimality of random allocations*, J. Math. Anal. Appl., 105 (1985), pp. 113-136.
[29] ———, *On the theory of Banach valued multifunctions, Part* 1: *Integration and conditional expectation*, J. Mult. Anal., 16 (1985), to appear.
[30] ———, *Stochastic nonsmooth analysis and optimization I*, Trans. Amer. Math. Soc. (to appear).
[31] B. PELEG, *Efficiency prices for optimal consumption plans*, J. Math. Anal. Appl., 29 (1970), pp. 83-90.
[32] ———, *Efficiency prices for optimal consumption plans II*, Israel J. Math., 9 (1971), pp. 222-234.
[33] ———, *Efficient random variables*, J. Math. Econom., 5 (1978), pp. 242-252.
[34] ———, *Efficiency prices for optimal consumption plans*, J. Math. Anal. Appl., 32 (1970), pp. 630-638.
[35] B. PELEG AND M. YAARI, *Efficiency prices in infinite dimensional spaces*, J. Econom. Theory, 2 (1970), pp. 41-85.
[36] R. PHELPS, *Weak* support points of convex sets in E^**, Israel J. Math., 2 (1964), pp. 177-182.
[37] R. RADNER, *Efficiency prices for infinite horizon production programs*, Rev. Econom. St., 34 (1967), pp. 51-66.
[38] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton Univ. Press, Princeton, NJ, 1970.
[39] ———, *Measurable dependence of convex sets and functions on parameters*, J. Math. Anal. Appl., 28 (1969), pp. 4-25.
[40] ———, *Integral functionals, normal integrands and measurable selectors*, in Nonlinear Operators and Calculus of Variations, Gossez et al., eds., Lecture Notes in Mathematics 543, Springer, Berlin, 1976.
[41] ———, *Convex integral functionals and duality*, in Contributions to Nonlinear Functional Analysis, Ed Zarantonello, ed., Academic Press, New York, 1971, pp. 215-236.
[42] ———, *Conjugate Duality and Optimization*, CBS Regional Conference Series in Applied Mathematics 16, Society for Industrial and Applied Mathematics, Philadelphia, 1973.
[43] ———, *The Theory of Subgradients and its Applications to Problems of Optimization*, Heldermann Verlag, Berlin, 1981.
[44] M. F. SAINT-BEUVE, *On the extension of von Neumann-Aumann's theorem*, J. Funct. Anal., 17 (1974), pp. 112-129.
[45] G. SALINETTI AND R. WETS, *On the relations between two types of convergence for convex functionals*, J. Math. Anal. Appl., 60 (1977), pp. 211-226.
[46] ———, *Convergence of sequences of convex sets in finite dimensions*, SIAM Rev., 21 (1979), pp. 18-33.
[47] ———, *On the convergence of closed valued measurable multifunctions*, Trans. Amer. Math. Soc., 266 (1981), pp. 275-289.
[48] M. VALADIER, *On conditional expectation of random sets*, Annali di Mat. Pure Appl. (1981), pp. 81-91.
[49] S. S. KHURANA, *Weak sequential convergence in L_E and Dunford-Pettis property of L_E^1*, Proc. Amer. Math. Soc., 78 (1980), pp. 85-88.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 3, May 1986

009

AN APPROACH TO SIMULTANEOUS SYSTEM DESIGN, I. SEMIALGEBRAIC GEOMETRIC METHODS*

BIJOY K. GHOSH†

Abstract. This paper introduces semialgebraic parameterization as an approach to analyze simultaneous stabilization and pole placement problems. Rational families of plants of a given McMillan degree, that are simultaneously stabilizable by a fixed family of compensators, are parameterized. For a discrete family of plants, the parameterization problem reduces to the simultaneous stabilization or the pole placement problem of a r-tuple of multi input multi output plants by a nonswitching compensator. It is shown that by removing a semialgebraic subset of a proper algebraic set, the "space of plants" can be decomposed into components that are either simultaneously stabilizable or simultaneously unstabilizable. Under special cases, explicit parameterization of the semialgebraic set is obtained. Finally a necessary condition for the simultaneous stabilization of single input or single output plants is obtained.

Key words. plant, compensator, semialgebraic-set, decision-theory

AMS(MOS) subject classifications. 93, 14

1. Introduction. Classically, in control theory one considers a lumped, linear, time-invariant, proper or strictly proper plant and one of the design objectives is to construct an output feedback scheme that would stabilize the plant. Not all plants have a compensator that satisfies the specified design constraints and it is of interest to parameterize those plants that do. In this paper, the semialgebraic properties of the parameterization problem is studied. In a related part II we study this problem via algebraic geometric methods.

In order to introduce the class of problems to be investigated, we consider a $p \times m$ rational, transfer function matrix $G(s)$ modelling a m input p output plant and address the following problem:

Problem 1.1. Let S be a topological space. Consider a family of plants $G_\lambda(s)$ parameterized by $\lambda \in S$. Does there exist a compensator $K_{f(\lambda)}(s)$ where

$$f: S \to S_1 \subset S \tag{1.1}$$

such that the closed loop systems $G_\lambda[I + K_{f(\lambda)}(s)G_\lambda(s)]^{-1}$ are stable for all $\lambda \in S$.

If the answer to Problem 1.1 is "yes", one asks the following parameterization problem.

Problem 1.2. Let f be fixed. Describe the family of plants $G_\lambda(s)$ for which there exists a family of compensators $K_{f(\lambda)}(s)$ such that the closed loop system $G_\lambda(s)[I + K_{f(\lambda)}(s)G_\lambda(s)]^{-1}$ is stable for all $\lambda \in S$.

If $S = S_1$ and f is the identity map, an adaptive control problem, called the "switching compensator problem" is obtained. If f is a constant map, the so-called "nonswitching compensator problem" also known as the "blending problem" is obtained. An important class of the noswitching compensator problem, called the "simultaneous stabilization problem" arises when the set S is discrete. The problem is to describe the set of r-tuples of plants $G_1(s), \cdots, G_r(s)$ that admit a stabilizing compensator. Let us now consider the following two examples.

* Received by the editors August 21, 1984, and in revised form on February 11, 1985.

† Department of Systems Science and Mathematics, Washington University, St. Louis, Missouri 63130. This research was partially supported by National Aeronautics and Space Administration grant NSG 2265 while the author was at Harvard University, Cambridge, Massachusetts. This paper is part of the author's Ph.D. thesis at Harvard University, Cambridge, Massachusetts.

Example 1.1 (*a switching compensator paroblem*). Let S be the set of real numbers $\mathbb{R}$. Consider $G_\lambda(s)$ to be a family of plants of degree 1 given by $G_\lambda(s)=1/(s+\lambda^2)$, where $\lambda \in \mathbb{R}$. Let K_λ be a family of feedback gains given by $K_\lambda = k_1\lambda + k_2$ where $k_1, k_2, \lambda \in \mathbb{R}$. Let us now ask the following question: Does there exist some values of $k_1, k_2 \in \mathbb{R}^2$ such that the closed loop system $G_\lambda(s)[1+K_\lambda G_\lambda(s)]^{-1}$ is stable for all $\lambda \in \mathbb{R}$?

Of course it is easy to see that K_λ stabilizes $G_\lambda(s)$ for all $\lambda \in \mathbb{R}$ iff

$$\lambda^2 + k_1\lambda + k_2 > 0. \tag{1.2}$$

Clearly, for

$$k_1^2 < 4k_2, \tag{1.3}$$

the inequality (1.2) is satisfied for all $\lambda \in \mathbb{R}$. Therefore, the family of plants $G_\lambda(s)$ is stabilizable by the family of switching compensators $K_\lambda(s)$ provided (1.3) is satisfied.

Example 1.2 (*a parameterization problem*). As a continuation of Example 1.1, let $G_\lambda(s)=a/(bs+\lambda^2)$, $K_\lambda = k_1\lambda + k_2$. Let us now ask the following question: For which $a, b \in \mathbb{R}^2$ does there exist $k_1, k_2 \in \mathbb{R}^2$ such that the closed loop system $G_\lambda(s)[1+K_\lambda G_\lambda(s)]^{-1}$ is stable for all $\lambda \in \mathbb{R}$?

Once again, K_λ stabilizes $G_\lambda(s)$ for all $\lambda \in \mathbb{R}$ iff

$$b(\lambda^2 + ak_1\lambda + ak_2) > 0. \tag{1.4}$$

Eliminating the variables $k_1, k_2 \in \mathbb{R}^2$ from the above inequation, we may check that the set of $a, b, \lambda \in \mathbb{R}^3$ for which there exists some $k_1, k_2 \in \mathbb{R}^2$ satisfying (1.4) is given by

$$(b>0 \text{ and } a=0 \text{ and } \lambda \neq 0) \quad \text{or} \quad (a \neq 0). \tag{1.5}$$

The set of $(a, b) \in \mathbb{R}^2$ for which (1.5) is satisfied for all $\lambda \in \mathbb{R}$ is given by

$$\{(a, b) | a \neq 0\}, \tag{1.6}$$

which is also the required solution to the parameterization problem.

As pointed out in [15], the compensator problem 1.1 and the parameterization Problem 1.2 is encountered in reliability studies. One frequently encounters situations when it is desirable to stabilize a plant with multiple modes of operation. For example, if $G_1(s)$ models a plant in its nominal mode, one might consider $G_2(s), \cdots, G_r(s)$ as the models of the plant in the faulted mode. It might be desirable to construct a feedback nonswitching compensator that simultaneously stabilizes $G_1(s), \cdots, G_r(s)$. In another situation, for example in considering an adaptive control problem, one considers a parameterized family of plants $G_\lambda(s)$ and wishes to construct a stabilizing compensator $K_\lambda(s)$ such that $(G_\lambda(s), K_\lambda(s))$ is stable for all λ in a parameter set. If $K_\lambda(s)$ is independent of λ, then the problem reduces to the blending problem, i.e. of constructing a fixed compensator for a family of plants. For details on the motivation and other references we refer to [1], [12].

The main idea of this paper is now summarized. First of all, the space of proper plants of McMillan degree n and the space of proper compensators of McMillan degree q have been described as a quasi-affine variety. In particular the plants and the compensators under consideration are parameterized as semialgebraic subsets of the affine spaces $\mathbb{R}^N$ and $\mathbb{R}^M$, respectively. By using the Routh-Hurwitz criterion [11], the set of plants and the set of compensators that correspond to a stable system in the closed loop is described by a set of semialgebraic conditions in the product space $\mathbb{R}^N \times \mathbb{R}^M$. The stabilizable plants are now described by the application of the decision method [2] which utilizes a rational procedure of eliminating the compensator variables from the above semialgebraic sets of conditions. By Tarski [26] and Seidenberg [22]

(see also Cohen [7]) this results in a semi-algebraic parameterization of the set of stabilizable plants in $\mathbb{R}^N$.

Of course the above argument continues to hold if one considers the pole-placement problem instead of the stabilization problem and a r-tuple of plants or more generally a rational family of plants of a fixed McMillan degree instead of a single plant. It may be noted, however, that the above parameterization may not be obtained by an efficient algorithm since it is known [10] that the Tarski-Seidenberg algorithms are computationally inefficient. However, a recent improvement by Collins [8] and by Arnon [3] have considerably improved the efficiency.

The organization of this paper is as follows. In § 2 a parameterization of the space of input output systems of a fixed McMillan degree has been described. In § 3 the set of stabilizable/pole assignable plants have been parameterized by the application of the Routh-Hurwitz condition [11] and the decision theory [2]. The parameterization problem of § 3 is generalized in § 4 to a family of plants, rather than one. Specifically, a simultaneously stabilizable family of plants has been parameterized. In § 5, the path component properties of the pole placement problem is described. In §§ 6 and 7 we restrict our attention to the case of a r-tuple of single input or single output plants (in particular $1 \times m$ plants). In § 6, an explicit solution to the parameterization problem is obtained under the hypothesis that $(q+1)(m+1) = \sum n_i + rq$ where n_i, $i = 1, \cdots, r$ and q are the McMillan degrees of the plants and the compensator, respectively. Especially when the above hypothesis is not satisfied, in § 7 we parameterize a set of unstabilizable r-tuples of plants, thereby obtaining a necessary condition to the simultaneous stabilization problem. The paper concludes in § 8 with a discussion on the possibility of parameterizing the set of stabilizable/pole assignable r-tuples of plants by a dynamic compensator of finite but a priori unbounded McMillan degree. This new problem serves to give some measure of the relative depth of the parameterization questions posed and solved in this paper.

2. A parameterization of the space of systems. Let k be either $\mathbb{R}$ or $\mathbb{C}$. We now parameterize the set of $p \times m$ proper, rational, matrix valued transfer functions over k as a subset of k^N where $N = n(m+p) + mp$. Let (A, B, C, D) be a 4-tuple of matrices in k^N of orders $n \times n$, $n \times m$, $p \times n$ and $p \times m$, respectively. Let us consider the following.

DEFINITION. A tuple of matrices (A, B, C, D) is defined to be a minimal system of degree n if the proper $p \times m$ transfer function

$$G(s) = \sum_{i=1}^{\infty} CA^{i-1}B/s^i + D \tag{2.1}$$

is of McMillan degree n.

It is well known that a tuple (A, B, C, D) is minimal iff it is observable and reachable. The space of minimal systems of degree n is denoted by $\tilde{S}^n_{m,p}$. Let us consider the following proposition.

LEMMA 2.1. *$\tilde{S}^n_{m,p}$ is an irreducible subset of k^N.*

Proof. The affine space k^N is irreducible. $\tilde{S}^n_{m,p}$ is a nonempty Zariski open subset of k^N since it contains observable and reachable 4-tuples of matrices. Hence $\tilde{S}^n_{m,p}$ is irreducible. (See Hartshorne [16].) Q.E.D.

It is quite possible that two 4-tuples of matrices (A_1, B_1, C_1, D_1) and (A_2, B_2, C_2, D_2) correspond via (2.1) to the same transfer function $G(s)$. The above is indeed the case iff there exist a nonsingular $n \times n$ matrix $g \in Gl(n, k)$ such that

$$A_2 = gA_1g^{-1}, \quad B_2 = gB_1, \quad C_2 = C_1g^{-1}, \quad D_2 = D_1. \tag{2.2}$$

Thus there exists an action of $Gl(n, k)$ on the space $\tilde{S}^n_{m,p}$ and the problem is to parameterize the moduli space $\tilde{S}^n_{m,p}/Gl(n, k)$. This is done as follows.

LEMMA 2.2. *$\tilde{S}^n_{m,p}/Gl(n, k)$ is an analytic manifold of dimension $n(m+p)+mp$. Moreover, there exists a set of local co-ordinate charts with rational co-ordinate functions.*

Remark. The proof of Lemma 2.2 is an adaptation from Clark [6], Hazewinkel and Kalman [18], Byres and Hurt [4] and Hazewinkel [17]. We choose to restate this well-known fact since the algebraic structure of the moduli space is important in what we derive later on in this paper.

Proof. Let $\tilde{\Sigma}_0$ denote the set of reachable pairs of matrices (A, B) and denote

$$\Sigma_0 = \tilde{\Sigma}_0/Gl(n, k). \tag{2.3}$$

It is well known (see [4], [18]) that Σ_0 is an analytic manifold of dimension nm which admits a rational atlas. Let $\tilde{\Sigma}^n_{m,p}$ be the observable and reachable triples (A, B, C) and define

$$\Sigma^n_{m,p} = \tilde{\Sigma}^n_{m,p}/Gl(n, k). \tag{2.4}$$

Using a result due to Byrnes and Hurt [4], $\Sigma^n_{m,p} \to \Sigma_0$ is canonically an algebraic vector bundle of rank pn. Hence $\Sigma^n_{m,p}$ is a $n(m+p)$-dimensional analytic manifold having rational coordinate functions. Finally the proof of this lemma follows from the observation that

$$\tilde{S}^n_{m,p}/Gl(n, k) = \Sigma^n_{m,p} \times k^{mp}. \tag{2.5}$$

Q.E.D.

Alternatively, it is also possible to parameterize $\tilde{S}^n_{m,p}/Gl(n, k)$ in the following way.

Let $H^n_{m,p}$ be the affine $(2n-1)mp$-dimensional space of $p \times m$ block Hankel matrices of the type

$$H = \begin{bmatrix} H_1 & H_2 & \cdots & H_n \\ H_2 & H_3 & \cdots & H_{n+1} \\ \vdots & \vdots & \ddots & \vdots \\ H_n & H_{n+1} & & H_{2n-1} \end{bmatrix} \tag{2.6}$$

where H_i, $i = 1, \cdots, 2n-1$ are $p \times m$ matrices. Consider the affine space

$$H^n_{m,p} \times k^{mp} \times k^{mp}, \tag{2.7}$$

and define the subset

$$S^n_{m,p} \subset H^n_{m,p} \times k^{mp} \times k^{mp} \tag{2.8}$$

given by

$$\begin{aligned} S^n_{m,p} &= \{(H, H_{2n}, H_0) | \text{rank } H = n \text{ and col } (H_{n+1}, \cdots, H_{2n}) \\ &= \sum_{i=1}^{n} [\alpha_i I_p] \text{ col } (H_i, \cdots, H_{n+i-1}) \text{ for some } \alpha_1, \cdots, \alpha_n \in k\}. \end{aligned} \tag{2.9}$$

There is an algebraic map

$$\Phi: \tilde{S}^n_{m,p} \to S^n_{mp} \tag{2.10}$$

defined by

$$\Phi(A, B, C, D) = \left(\begin{bmatrix} CB & \cdots & CA^{n-1}B \\ CA^{n-1}B & \cdots & CA^{2n-2}B \end{bmatrix}, CA^{2n-1}B, D \right). \tag{2.11}$$

We now consider the following lemma.

LEMMA 2.3. *$S^n_{m,p}$ is a quasi affine algebraic variety in the affine space $k^{(2n+1)mp}$.*

Proof. By (2.9), $S^n_{m,p}$ is an open subset of a closed algebraic subset in $k^{(2n+1)mp}$. Moreover, $S^n_{m,p}$ is irreducible, since $\tilde{S}^n_{m,p}$ is, and the map Φ is algebraic (see Shaferavich [23]). Q.E.D.

A topology on $S^n_{m,p}$ is induced from the Zariski topology [16] on $k^{(2n+1)mp}$. It is well known, by realization theory [19], that every element of $S^n_{m,p}$ corresponds with a $p \times m$ proper plant of McMillan degree n. Hence $S^n_{m,p}$ is isomorphic to the moduli space $\tilde{S}^n_{m,p}/Gl(n, k)$.

3. A parameterization of the space of stabilizable/pole assignable systems. Let k be the real field $\mathbb{R}$. To begin with, we consider the product space of $p \times m$ proper plants of degree n and $m \times p$ proper compensators of degree q, given by

$$S^n_{m,p} \times S^q_{p,m}, \tag{3.1}$$

which is of course a quasi affine variety. (See [12]). We now consider the following two problems

Problem 3.1. Describe the set of plants $G(s)$ in $S^n_{m,p}$ and compensators $K(s)$ in $S^q_{p,m}$ such that the pair $(G(s), K(s))$ is stable in the closed loop.

Problem 3.2. Describe the set of plants $G(s)$ in $S^n_{m,p}$ and compensators $K(s)$ in $S^q_{p,m}$ such that the closed loop system $G(s)[I+K(s)G(s)]^{-1}$ has poles in a prescribed $n+q$ tuple of self-conjugate complex points $s_1, \cdots, s_{n+q}$.

We now show that there exists a semialgebraic parameterization to the set of plants and compensators satisfying the properties stipulated in Problems 3.1 and 3.2.

THEOREM 3.1. *The subset U_1 of* (3.1) *given by*

$$U_1 = \{(G(s), K(s)) | K(s) \text{ stabilizes } G(s)\}, \tag{3.2}$$

is semialgebraic.

In order to prove Theorem 3.1 we need the following lemma.

LEMMA 3.1. *Let Φ be the map*

$$\Phi: S^n_{m,p} \times S^q_{p,m} \to \mathbb{R}^{n+q} \tag{3.3}$$

defined by

$$\begin{aligned} \Phi(G(s), K(s)) = {} & \text{coefficients of the monic characteristic polynomial} \\ & \pi(s) \text{ of the closed loop system } G(s)[I+K(s)G(s)]^{-1}. \end{aligned} \tag{3.4}$$

Then Φ is rational.

Proof. Consider the affine spaces

$$H_n = H^n_{m,p} \times \mathbb{R}^{mp} \times \mathbb{R}^{mp} \tag{3.5}$$

and

$$H_q = H^q_{p,m} \times \mathbb{R}^{mp} \times \mathbb{R}^{mp} \tag{3.6}$$

associated with $S^n_{m,p}$ and $S^q_{p,m}$, respectively. By realization theory [19], every point in the quasi-affine algebraic variety $S^n_{m,p} \times S^q_{p,m}$ of $H_n \times H_q$ corresponds to a plant-compensator pair $(G(s), K(s))$, and therefore corresponds to the closed loop plant $\underline{G}(s) = G(s)[I+K(s)G(s)]^{-1}$. Moreover, the coefficients of the characteristic polynomial of $\underline{G}(s)$ is rational in the parameters of $H_n \times H_q$ since $G(s)$ and $K(s)$ are of fixed degrees n and q, respectively. By restriction to $S^n_{m,p} \times S^q_{p,m}$, Φ is clearly rational. Q.E.D.

LEMMA 3.2. *Consider the real $n+q(th)$ degree polynomial*

$$p(s) = s^{n+q} + a_{n+q-1}s^{n+q-1} + \cdots + a_0 \tag{3.7}$$

parameterized as points in $\mathbb{R}^{n+q}$. *Then the set*

$$S=\{(a_0,\cdots,a_{n+q-1})|p(s)\ \textit{is stable}\} \tag{3.8}$$

is semialgebraic in $\mathbb{R}^{n+q}$.

Proof. The proof is omitted as it is the well-known Routh–Hurwitz condition [11]. Q.E.D.

Theorem 3.1 now follows trivially from the Lemmas 3.1 and 3.2. Note that the compensator $K(s)$ stabilizes the plant $G(s)$ just in case the associated characteristic polynomial $\pi(s)$ is stable. Note also that the coefficients of $\pi(s)$ are rational in the plant and the compensator parameters. Q.E.D.

The proof of the following theorem is analogous and is omitted.

THEOREM 3.2. *The subset* U_2 *of* (3.1) *given by*

$$U_2=\{(G(s),K(s))|\textit{the poles of } G(s)(I+K(s)G(s))^{-1} \textit{ are at a given set of self conjugate complex points } s_1,\cdots,s_{n+q}\} \tag{3.9}$$

is semialgebraic.

We now state the stabilizability problem as follows.

Problem 3.3. Describe the set of plants in $S^n_{m,p}$ for which there exists a compensator $K(s)$ in $S^q_{p,m}$ such that the closed loop system $G(s)[I+K(s)G(s)]^{-1}$ is stable.

We also state the pole assignability problem as follows.

Problem 3.4. Describe the set of plants in $S^n_{m,p}$ for which there exists a compensator $K(s)$ in $S^q_{p,m}$ such that the closed loop system $G(s)[I+K(s)G(s)]^{-1}$ has poles in an arbitrary $n+q$ set of self-conjugate complex points.

The following theorem reveals the semialgebraic nature of the sets described in Problems 3.3 and 3.4.

THEOREM 3.3. *The subset* U_s *of* U_1 *given by*

$$U_s=\{G(s)|\exists K(s)\in S^q_{p,m} \textit{ and } G(s)[I+K(s)G(s)]^{-1} \textit{ is stable}\} \tag{3.10}$$

is semialgebraic. The subset U_p *of* U_2 *given by*

$$U_p=\{G(s)|\textit{for all self conjugate set } s_1,\cdots,s_{n+q} \textit{ of complex numbers, } \exists K(s)\in S^q_{p,m} \textit{ and } G(s)[I+K(s)G(s)]^{-1} \textit{ has poles at } s_1,\cdots,s_{n+q}\} \tag{3.11}$$

is semialgebraic.

Proof. Consider the product space (3.1) and consider the projection

$$\text{proj: } S^n_{m,p}\times S^q_{p,m}\to S^n_{m,p}. \tag{3.12}$$

It is clear that

$$U_s=\text{proj } U_1. \tag{3.13}$$

By Theorem 3.1, U_1 is semialgebraic and proj is a rational map. Thus by the Tarski [26], Seidenberg [22] theory of elimination over $\mathbb{R}$, U_s is semialgebraic. In order to show that U_p is semialgebraic consider the product space

$$S^n_{m,p}\times S^q_{p,m}\times\mathbb{R}^{n+q}, \tag{3.14}$$

and its subset U_2' given by

$$U_2'=\{(G(s),K(s),(c_0,\cdots,c_{n+q-1}))|\text{ the poles of } G(s)[I+K(s)G(s)]^{-1} \text{ are at the zeros of } \pi(s)=s^n+c_{n+q-1}s^{n+q-1}+\cdots+c_0\}. \tag{3.15}$$

Consider now the following two projections:

$$\text{proj}_1\text{: } S^n_{m,p}\times S^q_{p,m}\times\mathbb{R}^{n+q}\to S^n_{m,p}\times\mathbb{R}^{n+q} \tag{3.16}$$

and

$$\text{proj}_2 : S^n_{m,p} \times \mathbb{R}^{n+q} \to S^n_{m,p}. \tag{3.17}$$

It is easy to see that

$$U_p = \text{proj}_2 [\overline{\text{proj}_1 (U_2')}], \tag{3.18}$$

where $\bar{\Omega}$ denotes the complement of Ω in the respective ambient space. Since the complement of a semialgebraic set is semialgebraic, by Tarski-Seidenberg [26], [22], U_p is semialgebraic. Q.E.D.

Example 3.1. Consider the single input single output plant $g(s)$ and the compensator $k(s)$ given by

$$g(s) = 1/(s^2 + bs + c), \tag{3.19}$$

$$k(s) = k/(s + \alpha). \tag{3.20}$$

The characteristic polynomial is given by

$$(s^2 + bs + c)(s + \alpha) + k, \tag{3.21}$$

which vanishes in the left half-plane iff

$$\alpha + b > 0 \text{ and } \alpha c + k > 0 \text{ and } (\alpha + b)(\alpha b + c) - (\alpha c + k) > 0. \tag{3.22}$$

The above inequalities (3.22) have been obtained by the application of the Routh-Hurwitz condition to the characteristic polynomial (3.21). In order to describe the set of stabilizable plants, we need to eliminate the variables k, α from (3.22) and obtain

$$b > 0 \text{ or } c - b^2 > 0. \tag{3.23}$$

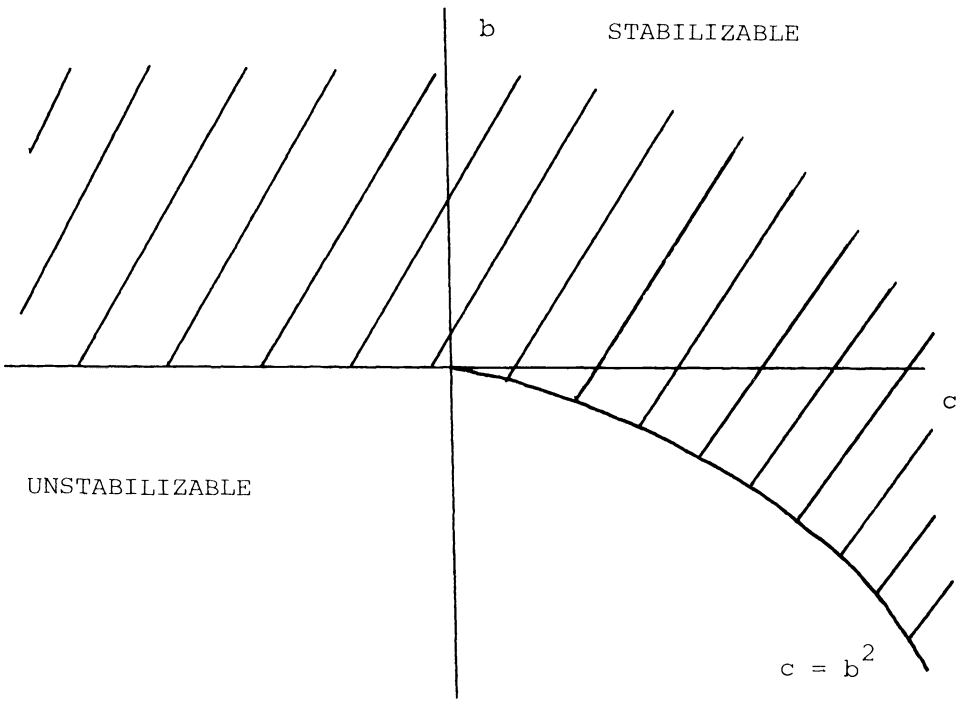


FIG. 3.1. *The set of stabilizable and unstabilizable plants of Example* 3.1.

In other words, for every b, c satisfying (3.23) there exists $k, \alpha \in \mathbb{R}$ such that (3.22) is satisfied. Thus the set of stabilizable plants are given by the shaded region in Fig. 3.1.

4. Semialgebraic path component properties of families of systems. Consider the space of r-tuples of plants

$$W = S_{m,p}^{n_1} \times \cdots \times S_{m,p}^{n_r} \tag{4.1}$$

of McMillan degrees of $n_1, \cdots, n_r$ respectively. We now define a family of r-tuples of plants and a family of compensators as follows. Let $\Lambda_1 \subset \mathbb{R}^{s_1}$, $\Lambda_2 \subset \mathbb{R}^{s_2}$ be a pair of semialgebraic subsets. A collection of families of r-tuple of plants in W, parameterized by a semialgebraic subset W_1 of W is given by an algebraic function

$$\phi_1 : \Lambda_1 \times W_1 \to W. \tag{4.2}$$

Similarly, a collection of families of compensators in $S_{p,m}^q$ parameterized by a semialgebraic subset W_2 of W is given by the algebraic function

$$\phi_2 : \Lambda_2 \times W_2 \to S_{p,m}^q. \tag{4.3}$$

Note that a given element $w_1 \in W_1$ indeed defines a family of r-tuples of plants given by

$$\phi_1(\cdot, w_1) : \Lambda_1 \to W. \tag{4.4}$$

Similarly,

$$\phi_2(\cdot, w_2) : \Lambda_2 \to S_{p,m}^q \tag{4.5}$$

for $w_2 \in W_2$, defines a family ofcompensators. For notational simplicity, we would denote by $w_1 \in W_1$, the family of r-tuples of plants given by $\phi_1(\cdot, w_1)$. Similarly, $w_2 \in W_2$ will be used to denote the family of compensators $\phi_2(\cdot, w_2)$. We now define the notion of σ-stabilizability. Let h be a fixed algebraic function

$$h : \Lambda_1 \to \Lambda_2. \tag{4.6}$$

Let f_i, $i = 1, \cdots, r$ respectively, be the projections

$$f_i : W \to S_{m,p}^{n_i}. \tag{4.7}$$

DEFINITION (σ-*stabilizability*). The family of r-tuples of plants $w_1 \in W_1$, is said to be σ-stabilizable by a family of compensators $w_2 \in W_2$, if the closed loop systems

$$f_i \phi_1(\lambda, w_1)[I + \phi_2(h(\lambda), w_2) f_i \phi_1(\lambda, w_1)]^{-1} \tag{4.8}$$

have poles with real part less than $-\sigma$, for all $\lambda \in \Lambda_1$ and for all $i = 1, \cdots, r$.

In order to introduce the main result of this section we consider the following problem.

Problem. Describe the subset W_1' of W_1 given by

$$W_1' = \{w_1 \in W_1 | \exists \sigma > 0, w_2 \in W_2 \text{ and } w_1 \text{ is } \sigma\text{-stabilizable by } w_2\}. \tag{4.9}$$

The main result is now summarized as follows.

THEOREM 4.1. *For a fixed algebraic function h described by* (4.6), *the set W_1' of σ-stabilizable families of r-tuple of plants for $\sigma > 0$, is open, semialgebraic. There exists a semialgebraic subset X of a proper algebraic set, where X consists of non-σ stabilizable families of r-tuples of plants in W_1, such that $W_1 - X$ has finitely many path components with the property that—if w_1 and w_1' are in the same path component of $W_1 - X$ then the*

family w_1 is σ-stabilizable iff the family w_1' is σ-stabilizable by some family of compensators $w_2 \in W_2$.

Proof. To see that W_1' is open, let $w_1 \in W_1'$. By assumption, there exists a $\sigma_0 > 0$ such that w_1 is σ_0-stabilizable by some $w_2 \in W_2$. Equivalently, the closed loop systems $f_i\phi_1(\lambda, w_1)[I + \phi_2(h(\lambda), w_2), f_i(\phi_1(\lambda, w_1))]^{-1}$ have poles with real part less than $-\sigma_0$, for all $\lambda \in \Lambda_1$ and for all $i = 1, \cdots, r$. Since $\sigma_0 > 0$, every element w_1' sufficiently close to w_1 in W_1', is σ_0'-stabilizable for σ_0' sufficiently close to σ_0. Hence $\sigma_0' > 0$. Hence W_1' is open.

To see that W_1' is semialgebraic, consider the space

$$W_1 \times W_2 \times \Lambda_1 \times \mathbb{R}^-, \tag{4.10}$$

where $\mathbb{R}^-$ is the open negative real axis. Consider the subset U in (4.10) given by

$$U = \{(w_1, w_2, \lambda_1, \sigma) | f_i(\phi_1(\lambda_1, w_1)) \text{ is } \sigma\text{-stabilizable by } \phi_2(h(\lambda_1), w_2)\}. \tag{4.11}$$

By the Routh–Hurwitz condition [11], U is semialgebraic. Let p be the projection

$$\text{proj}_1 \colon W_1 \times W_2 \times \Lambda_1 \times \mathbb{R}^- \to W_1 \times \Lambda_1 \tag{4.12}$$

by the Tarski–Seidenberg [26], [22] theory of elimination over R, $U_1 = \text{proj}_1 U$ is semialgebraic in $W_1 \times \Lambda_1$. Moreover $\bar{U}_1$, the complement of U_1 in $W_1 \times \Lambda_1$ is also semialgebraic. Consider the projection

$$\text{proj}_2 \colon W_1 \times \Lambda_1 \to W_1. \tag{4.13}$$

The set $U_2 = \overline{\text{proj}_2(\bar{U}_1)}$ is semialgebraic in W_1. Moreover, U_2 precisely corresponds to those families of r tuples of plants that are σ-stabilizable. Hence $W_1' = U_2$ is semialgebraic.

Thus W_1' is open, semialgebraic and by Delzell's theorem [9], is described by conjunction and disjunction of strict inequalities.

$$h_j(\cdot) > 0 \tag{4.14}$$

for $j = 1, \cdots, t$, where t is some positive integer. Let us now define a proper, algebraic subset X_1 of W_1 given by

$$X_1 = \bigcup_j \{h_j(\cdot) = 0\} \tag{4.15}$$

for $j = 1, \cdots, t$. Since X_1 is algebraic, by Whitney [20], $W_1 - X_1$ has finitely many path components. Let X be defined by

$$\begin{aligned} X &= \bigcup_j [\{h_j(\cdot) = 0 \cap \bar{W}_1'] \\ &= X_1 \cap \bar{W}_1'. \end{aligned} \tag{4.16}$$

Clearly X is semialgebraic and is a subset of the proper algebraic set X_1. Moreover

$$W_1 - X = (W_1 - X_1) \cup (X_1 - \bar{W}_1'). \tag{4.17}$$

Since $W_1 - X_1$ has finitely many path components with the property that every component either contains simultaneously σ-stabilizable family of plants for some $\sigma > 0$ or simultaneously σ-unstabilizable family of plants, it is clear $W_1 - X$ also has the same property. This is because a σ-stabilizable component and a σ-unstabilizable component cannot be separated by a boundary of σ-stabilizable family of plants, σ-stabilizability being an open condition. Q.E.D.

Let us now consider a refinement of Theorem 4.1 under the hypothesis that Λ_1 is compact.

THEOREM 4.2. *Let the parameter space Λ_1 be compact. For a fixed algebraic function h described by* (4.6), *the set W_1' of stabilizable families of r-tuple of plants is open, semialgebraic. There exists a semialgebraic subset X of a proper algebraic set, where X consists of nonstabilizable families of r-tuples of plants in W_1, such that $W_1 - X$ has finitely many path components with the property that—if w_1 and w_1' are in the same path components of $W_1 - X$, then the family w_1 is stabilizable iff the family w_1' is stabilizable by some family of compensators $w_2 \in W_2$.*

Remark. Note first of all that in Theorem 4.2 we are discussing stabilizability (or zero stabilizability) as opposed to Theorem 4.1 where we have σ-stabilizability, for $\sigma > 0$.

An important corollary of Theorem 4.2, follows in the special case when $\Lambda_1 = \{1, 2, \cdots, r\}$ and h is a constant function.

COROLLARY 4.1. *There exists a semi-algebraic subset of a proper algebraic set X, consisting of nonstabilizable r-tuples of plants such that $W_1 - X$ has finitely many components with the property that—if $(G_1(s), \cdots, G_r(s))$ and $(G_1'(s), \cdots, G_r'(s))$ are in the same path component then $(G_1(s), \cdots, G_r(s))$ is simultaneously stabilizable by a fixed nonswitching compensator iff $(G_1'(s), \cdots, G_r'(s))$ is so.*

If on the other hand Λ_2 the compensator parameter space is a single point, so that the algebraic function h described by (4.6) is a constant function, we have the simultaneous version of the "blending problem" originally addressed by Tannenbaum [24], [25].

Finally we come to the proof of Theorem 4.2.

Proof. The proof of this theorem is analogous to Theorem 4.1. However, one has to show that the set W_1' given by (4.18) is open.

$$W_1' = \{w_1 \in W_1 | \exists w_2 \in W_2 \text{ and } w_1 \text{ is stabilizable by } w_2\}. \tag{4.18}$$

Let $w_1 \in W_1'$. By hypothesis, there exists a $w_2 \in W_2$ with the following property—if $\pi_i(s, \lambda_1)$ is the characteristic polynomial of the closed loop system for $\lambda_1 \in \Lambda_1$ and $i \in \{1, \cdots, r\}$, then $\pi_i(s, \lambda_1)$ has roots in the open left half of the complex plane for all $\lambda_1 \in \Lambda_1$ and $i \in \{1, \cdots, r\}$. Let $s_{i,j}^{(\lambda_1)}$, $j = 1, \cdots, n_i + q$ be the roots of $\pi_i(s, \lambda_1)$, for a fixed $i \in \{1, \cdots, r\}$, $\lambda_1 \in \Lambda_1$. Define

$$r^{(\lambda_1)} = \max_{i,j} \operatorname{Re} s_{i,j}^{(\lambda_1)}, \tag{4.19}$$

where $r^{(\lambda_1)}$ is clearly a continuous function of λ_1. Since Λ_1 is compact, $r^{(\lambda_1)}$ attains a maximum value r where $r < 0$, since $w_1 \in W_1'$. Consider w_1' in a sufficiently small open neighborhood N_{w_1} of w_1 in W_1, together with w_2 in W_2. By repeating the above argument, there exists $r' < 0$ sufficiently close to r, such that the real parts of the roots of the associated characteristic polynomials $\pi_1'(s, \lambda_1)$ for all $i \in \{1, \cdots, r\}$ and $\lambda_1 \in \Lambda_1$ is less than or equal to r'. Hence $N_{w_1} \subset W_1'$. Q.E.D.

Remark. A semialgebraic subset of a proper algebraic set is automatically of Lebesgue measure zero. Thus, roughly speaking, one is deleting a "thin" set of nonstabilizable plants (family of plants) so that the remaining set has a "component" property. This, we remark, is conceptually significant since we now have to analyze big pieces of components rather than an individual family of plants.

5. Path component properties of the simultaneous pole placement problem. In this section we want to generalize the pole assignability Problem 3.4 to the simultaneous pole placement problem described as follows.

Problem 5.1. Parameterize the set of r-tuples of m input p output plants, $G_1(s), \cdots, G_r(s)$ each of a given McMillan degree n_i, $i=1, \cdots, r$, respectively, for which there exist a nonswitching p input m output proper compensator $K(s)$ of McMillan degree q that arbitrarily assigns all the poles of the closed loop systems $G_i(s)[I+K(s)G_i(s)]^{-1}$; $i=1, \cdots, r$.

Generalizing the technique used in the proof of Theorem 3.3, we now show the following.

THEOREM 5.1. *The set W_p of simultaneously pole assignable (by a proper compensator of Mcmillan degree q) r-tuples of plants $G_1(s), \cdots, G_r(s)$, in $W = S^{n_1}_{m,p} \times \cdots \times S^{n_r}_{m,p}$ is semialgebraic. There exists a proper algebraic set X in W such that $W-X$ has finitely many path components with the property that—if $(G_1(s), \cdots, G_r(s))$ and $(G'_1(s), \cdots, G'_r(s))$ are in the same path component, then $(G_1(s), \cdots, G_2(s))$ is simultaneously pole assignable by a proper compensator of McMillan degree q iff $(G'_1(s), \cdots, G'_r(s))$ is so.*

Remark. We do not claim in Theorem 5.1 that the set W_p is open.

Proof. The fact that W_p is semialgebraic is an immediate generalization of Theorem 3.3 (see [12] for details). Thus W_p is described by conjunction and disjunction of inequalities and equalities of the type

$$h_j(\cdot) > 0, \qquad g_k(\cdot) = 0 \tag{5.1}$$

for $j=1, \cdots, t_1$, $k=1, \cdots, t_2$ where t_1, t_2 are some positive integers. Define X to be the proper algebraic set given by

$$X = \bigcup_j \{h_j(\cdot)=0\} \bigcup_k \{g_k(\cdot)=\}. \tag{5.2}$$

It is clear that X is the proper algebraic set with the required path component property. Q.E.D.

For the rest of this section, our aim is to refine Theorem 5.1. First of all we claim that the set W_p, of pole assignable r-tuple of plants is not open in general. This we show by considering the following example.

Example 5.1. Consider a single input single output plant of McMillan degree 1 given by $(p_1 s+p_2)/(s+p_3)$ parameterized in $\mathbb{R}^3$ by the point (p_1, p_2, p_3). Consider the gain feedback c_1/c_2. To say that the gain places the poles of the system at $s=-\alpha$ is to say that the following matrix equation

$$[c_1 \quad c_2]\begin{bmatrix} p_1 & p_2 \\ 1 & p_3 \end{bmatrix} = [1 \quad \alpha] \tag{5.3}$$

has a solution. The set of all p_1, p_2, p_3 for which (5.3) has a solution is given by $\mathcal{S}_1 \cup \mathcal{S}_2$ where

$$\mathcal{S}_1 = \{(p_1, p_2, p_3) | p_2 \neq p_1 p_3\},$$

$$\mathcal{S}_2 = \{(p_1, p_2, p_3) | p_2 = p_1 p_3, p_3 = \alpha\}.$$

Note that although $\mathcal{S}_1$ is an open set, $\mathcal{S}_2$ is a proper Zariski closed subset of $\mathbb{R}^3$. Note also that the set

$$\mathbb{R}^2 - \{(p_1, p_2, p_3) | p_2 = p_1 p_3\} \tag{5.4}$$

has finitely many components that are pole assignable.

6. An explicit solution of the parameterization problem. In this section, we consider an r-tuple of $p \times m$, $\min(m, p) = 1$ strictly proper plants of McMillan degrees n_i, $i=$

$1, \cdots, r$, respectively. The problem is to parameterize explicitly the set of r-tuples of plants that can be stabilized by a proper compensator of McMillan degree q. In this section we obtain the parameterization under the assumption $(q+1)(m+1)=\Sigma n_i + rq$. Without any loss of generality we can assume that $m \geqq p$, for if $K(s)$ stabilizes $G_i^T(s)$, then $K^T(s)$ stabilizes $G_i(s)$. A given set of r, m input 1 output plants of McMillan degree $\leqq n_i$ may be represented as

$$\left[\sum_{j=0}^{n_i} p_{m+1,j}^i s^j\right]^{-1}\left[\sum_{j=0}^{n_i} p_{1,j}^i s^j, \cdots, \sum_{j=0}^{n_i} p_{m,j}^i s^j\right] \tag{6.1}$$

for $i=1,2,\cdots,r$, where $p_{m+1,n_i}^i=1$, $i=1,\cdots,r$ and $p_{k,n_i}^i=0$, $k=1,\cdots,m$, $i=1,\cdots,r$. Similarly, a 1 input m output proper compensator of McMillan degree $\leqq q$ is represented as

$$\left[\sum_{j=0}^{q} a_{1,j}s^j, \cdots, \sum_{j=0}^{q} a_{m,j}s^j\right]^T\left[\sum_{j=0}^{q} a_{m+1,j}s^j\right]^{-1}. \tag{6.2}$$

The associated return difference equation, $\det[I+K(s)G_p(s)]=0$ is given by

$$\begin{aligned}\pi_i(s) &= \sum_{k=1}^{m+p}\left[\sum_{j=0}^{n_i} p_{k,j}^i s^j\right]\left[\sum_{j=0}^{q} a_{k,j}s^j\right] \\ &\triangleq \sum_{j=0}^{n_i+q} c_{i,j}s^j \quad \forall i=1,\cdots,r,\end{aligned} \tag{6.3}$$

where $c_{i,n_i+q}=1\ \forall i=1,\cdots,r$. A generic r-tuple of plants defines a linear mapping χ, via (6.3), between the compensator parameters and the coefficient of the return difference polynomials given by

$$\chi: \mathbb{R}^{q(m+1)+m+1} \to \mathbb{R}^{\Sigma n_i + rq + 1}, \tag{6.4}$$

where χ may be defined as

$$\chi(\underline{a}) = \underline{a}S = (c_{1,0}, \cdots, c_{1,n_1+q-1}, \cdots, c_{r,n_r+q-1}, c_{1,n_1+q}), \tag{6.5}$$

where $\underline{a}$ is the $q(m+1)+m+1$ compensator parameters $a_{i,j}$ and S is the associated Sylvestor's matrix (see [12], [15] for details). It is known [12] that for a generic r-tuple of plants, the rows of S are independent. Thus the image of χ given by (6.4) is a subspace of codimension 1 in $\mathbb{R}^{\Sigma n_i + rq + 1}$. Let us now define

$$\begin{aligned}\mathcal{D} = \Bigg\{(c_{i,j}, j=0,\cdots,n_i+q;\ i=1,\cdots,r) \Bigg| &\sum_{j=0}^{n_i+q} c_{i,j}s^j \text{ has roots in the open} \\ &\text{left half of the complex plane, for } i=1,\cdots,r\Bigg\}\end{aligned} \tag{6.6}$$

and its convex hull $\Omega(\mathcal{D}) \subset R^{n_1+q} \times \cdots \times \mathbb{R}^{n_r+q}$. It was shown by Chen [5] that

LEMMA 6.1 (Chen [5]). *The convex hull of $\mathcal{D}$ is given by*

$$\Omega(\mathcal{D}) = \{(c_{i,j}) | c_{i,j} > 0\}. \tag{6.7}$$

Moreover if image χ is an affine hyperplane then

$$\text{image}\,(\chi) \cap \Omega(\mathcal{D}) \neq \varnothing \text{ iff } \text{image}\,(\chi) \cap \mathcal{D} \neq \varnothing. \tag{6.8}$$

Thus the stabilizability of the r-tuple of plants (6.1) is equivalent to solving (6.5) for some $c_{i,j} > 0\ \forall i,j$. We now prove the following lemma.

LEMMA 6.2. *Assume* $(q+1)(m+1)=\sum n_i+rq$. *For a given r-tuple of plants (chosen generically) let* $\underline{\alpha}$ *be a vector orthogonal to image* χ. *Then the r-tuple of plants is simultaneously unstabilizable iff* $\underline{\alpha}\in\Omega(\mathscr{D})$.

Proof. (if) By assumption $\underline{\alpha}\in\Omega(\mathscr{D})$ so that

$$\text{image}\,(\chi)\cap\Omega(\mathscr{D})=\varnothing. \tag{6.9}$$

Thus by Lemma 6.1,

$$\text{image}\,\chi\cap\mathscr{D}=\varnothing. \tag{6.10}$$

(only if) Assume that $\underline{\alpha}\notin\Omega(\mathscr{D})$. Then clearly there exists $\underline{\alpha}_1\in\Omega(\mathscr{D})$ such that $\underline{\alpha}\perp\underline{\alpha}_1$. Hence $\underline{\alpha}_1\in$ image χ so that

$$\text{image}\,\chi\cap\Omega(\mathscr{D})\neq\varnothing$$

or equivalently by Lemma 6.1,

$$\text{image}\,\chi\cap\mathscr{D}\neq\varnothing. \qquad \text{Q.E.D.}$$

We now use Lemma 6.2 to obtain a parameterization of the set of unstabilizable r-tuples of plants. Let us represent the Sylvestor matrix S in (6.5) as

$$S=\begin{bmatrix} s_{11} & \cdots & s_{1,t} \\ s_{t-1,1} & \cdots & s_{t-1,t}\end{bmatrix}, \tag{6.11}$$

where $t=\sum n_i+rq+1$. Let

$$v_1\bar{i}_1+v_2\bar{i}_2+\cdots+v_t\bar{i}_t \tag{6.12}$$

be a vector orthogonal to image χ, where $\bar{i}_j, j=1,\cdots,t$ is the standard basis of $\mathbb{R}^t$. Then clearly

$$v_i=(-1)^{i+1}\det S_i,\ i=1,\cdots,t, \tag{6.13}$$

where S_i is obtained from S by deleting its ith column. The set of stabilizable r-tuples of plants is now obtained by the semialgebraic condition

$$(v_1<0 \text{ or } \cdots \text{ or } v_t<0) \tag{6.14}$$

and

$$(v_1>0 \text{ or } \cdots \text{ or } v_t>0).$$

Example 6.1. Assume $q=0$, $r=2$, $m=2$, $p=1$, $n_1=1$, $n_2=2$. The Sylvestor matrix S is given by

$$S=\begin{bmatrix} p^1_{1,0} & p^2_{1,0} & p^2_{1,1} & 0 \\ p^1_{2,0} & p^2_{2,0} & p^2_{2,1} & 0 \\ p^1_{3,0} & p^2_{3,0} & p^2_{3,1} & 1\end{bmatrix}. \tag{6.15}$$

Thus,

$$\begin{aligned} v_1&=p^2_{1,0}p^2_{2,1}-p^2_{2,0}p^2_{1,1}\\ v_2&=p^1_{2,0}p^2_{1,1}-p^1_{1,0}p^2_{2,1},\\ v_3&=p^1_{1,0}p^2_{2,0}-p^1_{2,0}p^2_{1,0},\\ v_4&=\det\begin{bmatrix} p^1_{1,0} & p^2_{1,0} & p^2_{1,1} \\ p^1_{2,0} & p^2_{2,0} & p^2_{2,1} \\ p^1_{3,0} & p^2_{3,0} & p^2_{3,1}\end{bmatrix}.\end{aligned} \tag{6.16}$$

The simultaneously unstabilizable plants are given by

$$v_i v_j > 0 \quad \text{for } i, j \in \{1, 2, 3, 4\}. \tag{6.17}$$

7. A necessary condition for simultaneous stabilization of min $(m, p) = 1$ plants. We begin this section with the remark that the process of parameterizing the simultaneously stabilizable/unstabilizable r-tuples of plants (as described in §§ 2, 3 and 4) is computationally inefficient since it involves eliminating the set of compensator coefficients using decision algebra [2]. Thus we view §§ 2, 3 and 4 to be qualitative.

In [15], Ghosh and Byrnes have obtained sufficient conditions for the generic simultaneous stabilization of a r-tuple of plants. Furthermore the techniques described in [15] can be used to construct a simultaneously stabilizing compensator. In this section we obtain a necessary condition to the following simultaneous stabilization problem.

Problem 7.1. Given an r-tuple of $1 \times m$ proper plants $G_1(s), \cdots, G_r(s)$ of McMillan degrees n_i, $i = 1, \cdots, r$, respectively, does there exist a compensator $K(s)$ of McMillan degree q such that the closed-loop systems $G_i(s)[I + K(s)G_i(s)]^{-1}$, $i = 1, \cdots, r$ are (internally) stable?

Note that in the above Problem 7.1, the McMillan degree q of the compensator is held fixed. In order to illustrate the main idea of this section we consider the following example.

Example 7.1. Consider a triplet of single input single output plants $p_i(s)/q_i(s)$, $i = 1, 2, 3$ of McMillan degrees n_1, n_2, n_3 respectively. Let $c(s)/d(s)$ be the corresponding stabilizing compensators. Thus there exist Hurwitz polynomials $\Delta_1(s)$, $\Delta_2(s)$, $\Delta_3(s)$ such that

$$p_i(s)c(s) + q_i(s)d(s) = \Delta_i(s), \qquad i = 1, 2, 3, \tag{7.1}$$

or equivalently,

$$\begin{bmatrix} p_1(s) & q_1(s) \\ p_2(s) & q_2(s) \\ p_3(s) & q_3(s) \end{bmatrix} \begin{bmatrix} c(s) \\ d(s) \end{bmatrix} = \begin{bmatrix} \Delta_1(s) \\ \Delta_2(s) \\ \Delta_3(s) \end{bmatrix}. \tag{7.2}$$

A necessary condition for (7.2) to be satisfied is given by

$$\Delta_1(s)\eta_{32}(s) + \Delta_2(s)\eta_{13}(s) + \Delta_3(s)\eta_{21}(s) = 0, \tag{7.3}$$

where $\eta_{ij}(s) = p_i(s)q_j(s) - q_i(s)p_j(s)$.

In fact if the triplet of plants is chosen generically, then (7.3) is also a sufficient condition for the existence of $c(s)$, $d(s)$ satisfying 7.2 (see [13] for details). Now if we assume $n_1 = n_2 = n_3 = 1$ and $q = 0$, then generically we may write

$$\Delta_i(s) = \Delta_{i0} + \Delta_{i1}s, \tag{7.4}$$

and

$$\begin{aligned} \eta_{13}(s) &= \xi_{10} + \xi_{11}s + \xi_{12}s^2, \\ \eta_{13}(s) &= \xi_{20} + \xi_{21}s + \xi_{22}s^2, \\ \eta_{21}(s) &= \xi_{30} + \xi_{31}s + \xi_{32}s^2, \end{aligned} \tag{7.5}$$

so that (7.3) can be written as

$$[\Delta_{10}\,\Delta_{11}\,\Delta_{20}\,\Delta_{21}\,\Delta_{30}\,\Delta_{31}]\begin{bmatrix}\xi_{10} & \xi_{11} & \xi_{12} & 0\\ 0 & \xi_{10} & \xi_{11} & \xi_{12}\\ \xi_{20} & \xi_{21} & \xi_{22} & 0\\ 0 & \xi_{20} & \xi_{21} & \xi_{22}\\ \xi_{30} & \xi_{31} & \xi_{32} & 0\\ 0 & \xi_{30} & \xi_{31} & \xi_{32}\end{bmatrix} = [0\,0\,0\,0]. \tag{7.6}$$

A necessary and sufficient condition that $\Delta_i(s)$ defined in (7.4) is stable is given by the following:

$$\Delta_{i0} \text{ and } \Delta_{i1} \text{ have the same sign, for all } i = 1, 2, 3.$$

Assume for the purpose of illustration that every entry of the vector

$$[\Delta_{10}\,\Delta_{11}\,\Delta_{20}\,\Delta_{21}\,\Delta_{30}\,\Delta_{31}] \tag{7.7}$$

is positive. Thus to say that (7.6) is satisfied is to say that the polytope generated by the 6 points $p_1, p_2, p_3, p_4, p_5, p_6$ contain the origin, where

$$\begin{aligned} p_1 &= (\xi_{10}\,\xi_{11}\,\xi_{12}\,0), & p_2 &= (0\,\xi_{10}\,\xi_{11}\,\xi_{12}),\\ p_3 &= (\xi_{20}\,\xi_{21}\,\xi_{22}\,0), & p_4 &= (0\,\xi_{20}\,\xi_{21}\,\xi_{22}),\\ p_5 &= (\xi_{30}\,\xi_{31}\,\xi_{32}\,0), & p_6 &= (0\,\xi_{30}\,\xi_{31}\,\xi_{32}). \end{aligned} \tag{7.8}$$

Thus a necessary and sufficient condition that the triplet of single input single output plants of McMillan degree 1 is stabilizable by a feedback gain is that the polytope generated by

$$\begin{aligned} &p_1, p_2, p_3, p_4, p_5, p_6 \quad \text{or} \quad p_1, p_2, p_3, p_4, -p_5, -p_6 \quad \text{or}\\ &p_1, p_2, -p_3, -p_4, p_5, p_6 \quad \text{or} \quad p_1, p_2, -p_3, -p_4, -p_5, -p_6 \end{aligned} \tag{7.9}$$

contains the origin.

We would now like to generalize the argument presented in Example 7.1. Note first of all that for $n_1 = n_2 = n_3 = q = 1$ or for $n_1 = n_2 = n_3 = 2$, $q = 0$ one can mimic the argument of Example 7.1 to obtain a necessary and sufficient condition for the simultaneous stabilization of a triplet of plants.

The proof relies on the fact that a quadratic polynomial $\Delta_i(s)$, $i = 1, 2, 3$ is stable iff every coefficient of $\Delta_i(s)$ has the same sign. However, in general, a polynomial $\Delta(s)$ of degree n is stable only if every coefficient of $\Delta(s)$ is of the same sign. This idea is now applied as follows.

Consider a $m+2$ tuple of $1 \times m$ proper plants represented as

$$G_i(s) = \left[\frac{n_{i1}(s)}{n_{i,m+1}(s)} \quad \frac{n_{i2}(s)}{n_{i,m+1}(s)} \quad \cdots \quad \frac{n_{im}(s)}{n_{i,m+1}(s)}\right] \tag{7.10}$$

for $i = 1, 2, \cdots, m+2$. Consider a $m \times 1$ proper compensator represented as

$$k(s) = \left[\frac{c_1(s)}{c_{m+1}(s)} \quad \cdots \quad \frac{c_m(s)}{c_{m+1}(s)}\right]^T. \tag{7.11}$$

The compensator (7.11) stabilizes each one of the plants (7.10) iff there exists Hurwitz polynomials $\Delta_1(s), \cdots, \Delta_{m+2}(s)$ such that

$$n_{i1}(s)c_1(s) + \cdots + n_{i,m+1}(s)c_{m+1}(s) = \Delta_i(s) \tag{7.12}$$

for $i=1,\cdots,m+2$. Writing (7.12) in the matrix notation as

$$\begin{bmatrix} n_{11}(s) & \cdots & n_{1,m+1}(s) \\ n_{m+1,1}(s) & \cdots & n_{m+1,m+1}(s) \\ n_{m+2,1}(s) & \cdots & n_{m+2,m+1}(s) \end{bmatrix} \begin{bmatrix} c_1(s) \\ c_{m+1}(s) \end{bmatrix} = \begin{bmatrix} \Delta_1(s) \\ \Delta_{m+1}(s) \\ \Delta_{m+2}(s) \end{bmatrix}, \tag{7.13}$$

a necessary condition for the existence of $c_1(s),\cdots,c_{m+1}(s)$ that satisfy (7.13) is given by the condition

$$\det \begin{bmatrix} n_{11}(s) & \cdots & n_{1,m+1}(s) & \Delta_1(s) \\ n_{m+2,1}(s) & \cdots & n_{m+2,m+1}(s) & \Delta_{m+2}(s) \end{bmatrix} = 0. \tag{7.14}$$

The equation (7.14) can be viewed as a generalization of (7.3). Using the necessary condition for the stability of a polynomial, viz that every coefficient is of the same sign, one obtains the required necessary condition.

8. Conclusion and future developments. In this paper we describe an approach to simultaneous system design using semialgebraic geometric methods. The procedure described in this paper relies heavily on the fact that the McMillan degree of the compensator under consideration is bounded. Although this might be a reasonable assumption under most practical situations, it is also of interest to consider Problems 1.1 and 1.2 with the assumption that the McMillan degree of the compensator is not a priori bounded. We remark that the "space of compensators" under the above hypothesis is not finite-dimensional and semialgebraic geometric methods using decision algebra [2] are therefore not applicable. Using transcendental and interpolation methods, we have obtained (see [14]) semialgebraic parameterization of the simultaneously stabilizable and pole assignable collections of plants. Complete solution to the parameterization Problems 1.1 and 1.2 where the compensators are of unbounded McMillan degree is a subject of future research.

As an additional remark, it may be interesting to point out a recent research work by Richter and DeCarlo [21] on eigenvalue assignment by decentralized feedback and the fact the simultaneous system design problem may be viewed as a decentralized feedback problem with special structure.

Acknowledgments. Encouragement and constructive criticisms of Prof. Chris. I. Byrnes during the research work is gratefully acknowledged. Suggestions and comments of the referees are also gratefully acknowledged.

REFERENCES

[1] J. ACKERMANN AND S. TURK, *A common controller for a family of plant models*, 21st IEEE Conference on Decision and Control, 1982, pp. 240-244.

[2] B. D. O. ANDERSON, N. K. BOSE AND E. I. JURY, *Output feedback stabilization and related problems-solution via decision methods*, IEEE Trans. Automat. Control, AC-20 (1975), pp. 53-66.

[3] D. S. ARNON, *Algorithms for the geometry of semialgebraic sets*, Technical Report Number 436, Computer Science Dept., Univ. Wisconsin, Madison, 1981.

[4] C. I. BYRNES AND N. E. HURT, *On the moduli of linear dynamical systems*, Adv. in Math., Suppl. Series, 4 (1978), pp. 83-122; also in Modern Mathematical Systems Theory, MIR Press, Moscow, 1978. (In Russian.)

[5] R. CHEN, Ph.D. Dissertation, Univ. Florida, Gainesville, 1979.

[6] J. M. C. CLARK, *The consistent selection of local co-ordinates in linear system identification*, Proc. J.A.C.C., Purdue, 1976, pp. 576-580.

[7] P. COHEN, *Decision procedures for real and p-adic fields*, Comm. Pure Appl. Math., 22 (1969), pp. 131-151.

[8] G. E. COLLINS, *Quantifier elimination for real closed fields by cylindrical algebraic decomposition*, Second G.I. Conference, Kaiserslautern, 1975, Lecture Notes in Computer Science 33, Springer-Verlag, New York (1975), pp. 134–183.

[9] C. M. DELZELL, *A constructive, continuous solution to Hilbert's 17th problem, and other results in semialgebraic geometry*, Ph.D. Dissertation, Stanford Univ., Stanford, CA, 1980.

[10] M. J. FISCHER AND M. O. RABIN, *Superexponential complexity of presburger arithmetic*, MIT, MAC Tech. Memo. 43, Feb. 1974, also in Complexity of Computation. Proc. Sympos., New York, 1973, pp. 27–41. SIAM-AMS Proceedings, Vol. VII, American Mathematical Society, Providence, RI, 1974.

[11] F. R. GANMACHER, *The Theory of Matrices*, Chelsea, New York.

[12] B. K. GHOSH, *Simultaneous stabilization and pole placement of a multimode linear dynamic system*, Ph.D. Dissertation, Harvard Univ., Cambridge, MA, 1983.

[13] ———, *Simultaneous partial pole placement—a new approach to multimode system design*, IEEE Trans. Automat. Control, submitted.

[14] ———, *Transcendental and interpolation methods in simultaneous stabilization and simultaneous partial pole placement problems*, this Journal, submitted.

[15] B. K. GHOSH AND C. I. BYRNES, *Simultaneous stabilization and simultaneous pole placement by non-switching dynamic compensation*, IEEE Trans. Automat. Control, AC-28 (1983), pp. 735–741.

[16] R. HARTSHORNE, *Algebraic Geometry*, Springer-Verlag, New York, 1977.

[17] M. HAZEWINKEL, *Moduli and canonical forms for linear dynamical systems* II: *The topological case*, Math. System Theory, 10 (1977), pp. 363–385.

[18] M. HAZEWINKEL AND R. E. KALMAN, *On invariants, canonical forms and moduli for linear constant finite-dimensional dynamical systems*, in Lecture Notes in Economic and Mathematical System Theory 131, Springer-Verlag, Berlin, 1976, pp. 48–60.

[19] R. E. KALMAN, M. ARBIB AND P. FALB, *Topics in Mathematical System Theory*, McGraw-Hill, New York, 1965.

[20] J. W. MILNOR, *Singular points of complex hypersurfaces*, Ann. Mathematics Studies 61, Princeton Univ. Press, Princeton, NJ, 1974.

[21] S. RICHTER AND R. DECARLO, *A homotopy method for eigenvalue assignment using decentralized state feedback*, IEEE Trans. Automat. Control, Vol. AC-29 (1984), pp. 148–158.

[22] A. SEIDENBERG, *A new decision method for elementary algebra*, Ann. Math., 60 (1954), pp. 365–374.

[23] I. R. SHAFERAVICH, *Basic Algebraic Geometry*, Springer-Verlag, New York, 1974.

[24] A. TANNENBAUM, *Feedback stabilization of linear dynamical plants with uncertainty in the gain factor*, Int. J. Control, 32 (1980), pp. 1–16.

[25] ———, *Invariance and System Theory: Algebraic and Geometric Aspects*, Springer-Verlag, Berlin, 1981.

[26] A. TARSKI, *A Decision Method for Elementary Algebra and Geometry*, Univ. California Press, Berkeley, 1951.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 3, May 1986

010

STATE ESTIMATION AND CONTROL OF CONDITIONALLY LINEAR SYSTEMS*

WOJCIECH J. KOLODZIEJ† AND RONALD R. MOHLER†‡

Abstract. The filtering problem for a partially observable stochastic system, with linear-in-unobservable state dynamics and non-Gaussian initial conditions is studied here. It is shown that the conditional expected value of the unobservable states, given the past and present observations, can be expressed in terms of a finite dimensional set of statistics. This result, which generalizes the conditionally Gaussian filter, is used to derive a separation principle for a linear-quadratic control problem.

Key words. optimal filtering, stochastic control, non-Gaussian stochastic systems

Introduction. Stochastic, partially observable systems, with linear-in-unobservable state dynamics are termed *conditionally linear systems* here. It is well-known that the solution of a state estimation problem for a conditionally linear system with Gaussian distribution of the initial state is given in terms of two sets of sufficient statistics, satisfying stochastic differential equations [4].

Discussed here is the state estimation problem which generalizes the above result for the case of an arbitrary a priori distribution of the initial state. The method applied in this study is based on the derivation of an explicit formula for the conditional characteristic function of the state, given the past and present observations. This approach, as opposed to the one presented in [1], avoids the use of Zakai's equation.

It is shown here that the conditional characteristic function of the present and past states, given the present and past observations, is parametrically determined by a finite number of sufficient statistics. This result leads to the derivation of a filter, in the form of a finite set of stochastic differential equations which extends the result of [1] in a similar manner as a conditionally Gaussian filter generalizes a Kalman filter.

Also discussed here and illustrated by the examples, is the suitability of the filter structure for the study of stochastic control and parameter estimation.

1. Problem formulation and the main result. Given the following system of stochastic differential equations

$$dx_t = (f_0(t, y) + f_1(t, y)x_t)\,dt + g_0(t, y)\,dw_t + q_0(t, y)\,dv_t, \tag{1.1}$$

$$dy_t = (h_0(t, y) + h_1(t, y)x_t)\,dt + dv_t, \qquad 0 \leqq t \leqq T, \tag{1.2}$$

where f_0, f_1, g_0, q_0, h_0, h_1 are $\mathbf{Y}_t$ measurable functionals of y, with $\mathbf{Y}_t = \sigma\text{-alg}\{y_s, 0 \leqq s \leqq t\}$, and w_t, v_t are independent Wiener processes.

The objective is to find $\hat{x}_t = \mathbf{E}(x_t|\mathbf{Y}_t)$, assuming that x_t, y_t satisfy (1.1) and (1.2), and that the conditional distribution of the initial states $F(a) = P(x_0 \leqq a|y_0)$ is given.

The organization of this section starts with Lemma 1, whereby it is shown that the conditional characteristic function of $(x_{t_0}, x_{t_1}, \cdots, x_{t_n})|\mathbf{Y}_t$, for arbitrary partition $0 \leqq t_0 < t_1 < \cdots < t_n \leqq t \leqq T$, of the interval $[0, T]$ is of a particular form. Results from the theory of conditionally Gaussian processes are used here.

* Received by the editors April 24, 1984, and in revised form February 1, 1985. This research was sponsored by the Office of Naval Research under contract N00014-81-K0814.

† Department of Electrical and Computer Engineering, Oregon State University, Corvallis, Oregon 97331.

‡ During 1983-84, NAVALEX Professor, Department of Electrical and Computer Engineering, Naval Postgraduate School, Monterey, California 93943.

Next, in Lemma 2, the explicit formula for the characteristic function of $x_t|\mathbf{Y}_t$ is derived, and finally, in Lemma 3, all the results are organized to yield the recursive, finite-dimensional set of filter equations.

The assumptions used in the proof of Lemma 1 and 2 are listed below: Let C_T denote the space of continuous functions $\eta = \{\eta_t, 0 \leqq t \leqq T\}$. It is assumed that for each $\eta \in C_T$

$$\int_0^T \left(\sum_{k=0}^{1} (|f_k(t, \eta)| + |h_k(t, \eta)|) + |g_0(t, \eta)|^2 + |q_0(t, \eta)|^2 \right) dt < \infty. \tag{1.3}$$

The above assumption assures the existence of the Ito integrals in (1.1) and (1.2) [3]. In order to use the results for conditionally Gaussian processes it is also assumed that [4]:

$$\text{for all } \eta \in C_T, \quad t \in [0, T], \quad |f_1(t, \eta)| + |h_1(t, \eta)| \leqq \text{const}, \tag{1.4}$$

and

$$\int_0^T \mathbf{E}(|f_0(t, y)|^4 + |g_0(t, y)|^4 + |q_0(t, y)|^4)\, dt < \infty, \mathbf{E}(|x_0|^4) < \infty. \tag{1.5}$$

LEMMA 1. *Let*

$$\phi_t = \exp\left(i \sum_{k=0}^{n} z_k x_{t_k} \right), \quad z = \begin{bmatrix} z_1 \\ \vdots \\ z_n \end{bmatrix} \in R^n, \quad 0 \leqq t_0 < t_1 < \cdots < t_n \leqq t \leqq T.$$

Then the conditional characteristic function of $(x_{t_0}, x_{t_1}, \cdots, x_{t_n})|\mathbf{Y}_t$ *is given by*

$$\mathbf{e}_t(z) = \mathbf{E}(\phi_t|\mathbf{Y}_t) = \int_{-\infty}^{\infty} \exp\,(Q(t, a, z, y))\, dF(a) \tag{1.6}$$

where $Q(t, a, z, y)$ *is quadratic in the variables* a *and* z.

Proof of Lemma 1. First notice that (1.1) solves as

$$x_t = \Phi_t \left(x_0 + \int_0^t \Phi_s^{-1}(f_0 - q_0 h_0)\, ds + \int_0^t \Phi_s^{-1} q_0\, dy_s + \int_0^t \Phi_s^{-1} g_0\, dw_s \right) \tag{1.7}$$

where

$$\Phi_t = \exp\left(\int_0^t (f_1 - q_0 h_1)\, ds \right).$$

Rewrite (1.7) in the symbolic way as

$$x_t = \psi_t(x_0, w, y). \tag{1.8}$$

Now, the following version of the Bayes formula will be used [2, p. 8]: Let $\phi_t(x_0, w, y)$ be a nonanticipative functional of its arguments with $\mathbf{E}(|\phi_t|) < \infty$ for all $t \in [0, T]$. Then

$$\mathbf{E}(\phi_t|\mathbf{Y}_t) = \int_{-\infty}^{\infty} \int_{C_T} \phi_t(a, \eta, y) \rho_t(a, \eta, y)\, d\mu_w(\eta)\, dF(a) \tag{1.9}$$

where μ_w is a Wiener measure on the measurable space of continuous functions η on $[0, T]$,

$$\rho_t(a, \eta, y) = \exp\left(\int_0^t h_1(\psi_s(a, \eta, y) - \hat{x}_s(y))\, d\nu_s \right. \tag{1.10}$$

$$\left. - \frac{1}{2} \int_0^t h_1^2(\psi_s(a, \eta, y) - \hat{x}_s(y))^2\, ds \right)$$

with $d\nu_s = dy_s - (h_0 + h_1\hat{x}_s)\,ds$, and $\psi_s(a, \eta, y)$ defined by (1.8). The innovations process ν_t can be represented by

$$\nu_t = \int_0^t (dy_s - (h_0(s, y) + h_1(s, y)\hat{x}_s(y))\,ds) = v_t + \int_0^t h_1(s, y)(x_s - \hat{x}_s(y))\,ds.$$

Now using the Ito formula we have

$$e^{iz\nu_t} = e^{iz\nu_s} + iz\int_s^t h_1(\tau, y)\, e^{iz\nu_\tau}(x_\tau - \hat{x}_\tau(y))\,d\tau$$
$$+ iz\int_s^t e^{iz\nu_\tau}\,dv_\tau - \frac{z^2}{2}\int_s^t e^{iz\nu_\tau}\,d\tau.$$

Multiplying both sides of the above equation by $e^{iz\nu_s}$ and taking the conditional expectation $\mathbf{E}(\cdot|\mathbf{Y}_s)$ gives

$$\mathbf{E}(e^{iz(\nu_t - \nu_s)}|\mathbf{Y}_s) = 1 - \frac{z^2}{2}\int_s^t \mathbf{E}(e^{iz(\nu_\tau - \nu_s)}|\mathbf{Y}_s)\,d\tau.$$

Solving the last equation yields

$$\mathbf{E}(e^{iz(\nu_t - \nu_s)}|\mathbf{Y}_s) = e^{-z^2(t-s)/2}, \tag{1.11}$$

which shows that $(\nu_t, \mathbf{Y}_t)$ is a Wiener process. Now rewrite $\rho_t(a, \eta, y)$ in a more convenient form. To this end introduce the following notation:

$$A_1(t, y) = h_1\left(\Phi_t\left(\int_0^t \Phi_s^{-1}(f_0 - q_0h_0)\,ds + \int_0^t \Phi_s^{-1}q_0\,dy_s\right) - \hat{x}_t\right),$$

$$A_2(t, y) = h_1\Phi_t,$$

$$A_3(t, y) = \Phi_t^{-1}g_0,$$

$$C_1(t, y) = \int_0^t A_1(s, y)\,d\nu_s - \tfrac{1}{2}\int_0^t A_1^2(s, y)\,ds,$$

$$C_2(t, y) = \int_0^t A_2(s, y)\,d\nu_s - \int_0^t A_1(s, y)A_2(s, y)\,ds,$$

$$C_3(t, y) = \int_0^t A_2^2(s, y)\,ds,$$

$$C_4(t, y, w) = \int_0^t A_2(s, y)\int_0^s A_3(\tau, y)\,dw_\tau\,d\nu_s$$
$$- \int_0^t A_1(s, y)A_2(s, y)\int_0^s A_3(\tau, y)\,dw_\tau\,ds,$$

$$C_5(t, y, w) = -\int_0^t A_2^2(s, y)\int_0^s A_3(s, y)\,dw_\tau\,ds.$$

Note also that $C_4(t, y, w)$ and $C_5(t, y, w)$ can be rewritten with the use of the Ito formula as follows:

$$C_4(t, y, w) = \int_0^t A_4(t, s, y)\,dw_s,$$

$$C_5(t, y, w) = \int_0^t A_5(t, s, y)\,dw_s,$$

where

$$A_4(t, s, y) = \left(\int_s^t A_2(\tau, y)\, d\nu_\tau - \int_s^t A_1(\tau, y) A_2(\tau, y)\, d\tau \right) A_3(s, y),$$

$$A_5(t, s, y) = -\left(\int_s^t A_2^2(\tau, y)\, d\tau \right) A_3(s, y).$$

Now, using the above notation, we have from (1.8) and (1.10)

$$\begin{aligned} \rho_t(a, w, y) &= \exp\left(C_1 + a(C_2 + C_5) + C_4 - \frac{a^2}{2} C_3 - \frac{1}{2} \int_0^t A_2^2 \left(\int_0^s A_3\, dw_\tau \right)^2 ds \right) \\ &= \exp\left(C_1 + aC_2 - \frac{a^2}{2} C_3 + \int_0^t (aA_5 + A_4)\, dw_s - \frac{1}{2} \int_0^t A_2^2 \left(\int_0^s A_3\, dw_\tau \right)^2 ds \right). \end{aligned} \tag{1.12}$$

The arguments in (1.12) were omitted for brevity. From (1.8) it follows that

$$x_t = \psi_t(x_0, w, y) = \Phi_t \left(x_0 + A_6(t, y) + \int_0^t A_3(s, y)\, dw_s \right)$$

where

$$A_6(t, y) = \int_0^t \Phi_s^{-1}(f_0 - q_0 h_0)\, ds + \int_0^t \Phi_s^{-1} q_0\, dy_s.$$

Combining (1.12) and the above,

$$\begin{aligned} \exp(Q(t, a, z, y)) &= \int_{C_T} \phi_t(a, \eta, y) \rho_t(a, \eta, y)\, d\mu_w(\eta) \\ &= \exp\left(C_1 + aC_2 - \frac{a^2}{2} C_3 + a \left(\sum_{k=1}^n \Phi_{t_k} i z_k \right) + \sum_{k=1}^n \Phi_{t_k} A_6(t_k, y) i z_k \right) \\ &\quad \cdot \int_{C_T} \exp\left(\int_0^t (aA_5 + A_4)\, d\eta_s + \sum_{k=1}^n i z_k \Phi_{t_k} \int_0^{t_k} A_3\, d\eta_s \right. \\ &\qquad \left. - \tfrac{1}{2} \int_0^t A_2^2 \left(\int_0^s A_3\, d\eta_\tau \right)^2 ds \right) d\mu_w(\eta). \end{aligned} \tag{1.13}$$

In order to evaluate the integral in (1.13), the following results will be used:

(i) Since the above integral represents a conditional expected value of its integrand, under the condition that y_s, $s \in [0, t]$ and $x_0 = a$ are given, the resulting distributions are of conditionally Gaussian type [4]. Note that this fact does *not* depend on the $F(a)$.

(ii) With all the variables in (1.13) being conditionally Gaussian we can use a convenient theorem:

THEOREM [4, pp. 12–13]. *Let w_t, $t \in [0, T]$ be a Wiener process and let $R(t)$, $G(t)$, and $H(t) \geqq 0$ be such that*

$$\int_0^T (|R(t)| + G(t)^2 + H(t))\, dt < \infty.$$

Then for all $t \in [0, T]$

$$\begin{aligned} &\mathbf{E}\left(\exp\left(\int_0^t R(s) G(s)\, dw_s - \int_0^t H(s) \left(\int_0^s G(\tau)\, dw_\tau \right)^2 ds \right) \right) \\ &\qquad = \exp\left(\frac{1}{2} D(t) + \frac{1}{2} \int_0^t G(s)^2 \Gamma(s)\, ds \right) \end{aligned} \tag{1.14}$$

where

$$d\Gamma(s) = (2H(s) - \Gamma(s)^2 G(s)^2)\, ds, \qquad \Gamma(t) = 0,$$

and $D(t)$ *is the covariance of* $\int_0^t R(s)\, d\xi_s$, *where*

$$d\xi_s = G(s)^2 \Gamma(s) \xi_s\, ds + G(s)\, dw_s, \qquad \xi_0 = 0.$$

Comparing the last integral in (1.13) with the equation given by (1.14), we note that the corresponding $R(t)$ is a linear function of a and z. Now (1.9), (1.13), (1.14), and the definition of $D(t)$ conclude the proof of Lemma 1.

From Lemma 1 it follows in particular that for $z \in R$, the characteristic function of $x_t|\mathbf{Y}_t$ is given by

$$\begin{aligned}\mathbf{e}_t(z) = C(t, y) \int_{-\infty}^{\infty} \exp\,(a^2 F_1(t, y) + aF_2(t, y) \\ + izaF_3(t, y) + izF_4(t, y) + z^2 F_5(t, y))\, dF(a),\end{aligned} \tag{1.15}$$

where F_1, F_2, F_3, F_4, F_5 do not depend on $F(a)$. Normalizing $\mathbf{e}_t(z)$ (i.e., requiring that $\mathbf{e}_t(0) = 1$) yields

$$\mathbf{e}_t(z) = \exp\,(izF_4 + z^2 F_5) \frac{\int_{-\infty}^{\infty} \exp\,(a^2 F_1 + a(F_2 + izF_3))\, dF(a)}{\int_{-\infty}^{\infty} \exp\,(a^2 F_1 + aF_2)\, dF(a)}. \tag{1.16}$$

Then from the general properties of the characteristic function, it follows that

$$\frac{1}{i} \frac{d\mathbf{e}_t(z)}{dz}\bigg|_{z=0} = \hat{x}_t, \qquad \left(\frac{1}{i}\right)^2 \frac{d^2 \mathbf{e}_t(z)}{dz^2}\bigg|_{z=0} = P_t + \hat{x}_t^2,$$

where $P_t = \mathbf{E}((x_t - \hat{x}_t)^2|\mathbf{Y}_t)$ i.e., the conditional variance of $x_t|\mathbf{Y}_t$. From the above and (1.16)

$$\hat{x}_t = F_3 I_t(1) + F_4, \tag{1.17}$$

$$P_t = -2F_5 + F_3^2 (I_t(2) - I_t^2(1)), \tag{1.18}$$

where

$$I_t(n) = \frac{\int_{-\infty}^{\infty} a^n \exp\,(a^2 F_1 + aF_2)\, dF(a)}{\int_{-\infty}^{\infty} \exp\,(a^2 F_1 + aF_2)\, dF(a)}, \qquad n = 1, 2. \tag{1.19}$$

The following Lemma defines F_i, $i = 1, 2, 3, 4, 5$ in (1.16).

LEMMA 2. *The characteristic function of* $x_t|\mathbf{Y}_t$ *is given by*

$$\mathbf{e}_t(z) = \exp\,(-\tfrac{1}{2} z^2 \bar{P}_t(0)) \frac{\int_{-\infty}^{\infty} \exp\,(a^2 F_1 + aF_2 + iz\bar{x}_t(a, 0))\, dF(a)}{\int_{-\infty}^{\infty} \exp\,(a^2 F_1 + aF_2)\, dF(a)}, \tag{1.20}$$

where $\bar{x}_t(a, 0)$, $\bar{P}_t(0)$ *are given as the solutions to the following set of differential equations with* $\sigma = 0$

$$\begin{aligned}&d\bar{x}_t(a, \sigma) = (f_0 + f_1 \bar{x}_t(a, \sigma))\, dt + (q_0 + \bar{P}_t(\sigma) h_1)(dy_t - (h_0 + h_1 \bar{x}_t(a, \sigma))\, dt), \\ &\bar{x}_0(a, \sigma) = a,\end{aligned} \tag{1.21}$$

$$d\bar{P}_t(\sigma) = (2f_1 \bar{P}_t(\sigma) + g_0^2 + q_0^2 - (q_0 + \bar{P}_t(\sigma) h_1)^2\, dt, \qquad \bar{P}_0(\sigma) = \sigma^2, \tag{1.22}$$

and

$$F_1 = -\frac{1}{2} \int_0^t h_1^2 \phi_s^2\, ds, \tag{1.23}$$

$$F_2 = \int_0^t \phi_s h_1 (d\nu_s + h_1 \phi_s I_s(1)\, ds), \tag{1.24}$$

$$\phi_t = \exp\left(\int_0^t (f_1 - h_1(q_0 + \bar{P}_s(0) h_1))\, ds \right). \tag{1.25}$$

Proof of Lemma 2. Since the F_i do not depend on $F(a)$ (see Lemma 1), take

$$dF(a) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(a-m)^2}{2\sigma^2} \right) da, \qquad \sigma > 0. \tag{1.26}$$

In this case the resulting conditionally Gaussian distribution allows for explicit $\mathbf{e}_t(z)$ calculation [4]. Accordingly,

$$\mathbf{e}_t(z) = \exp\left(iz\bar{x}_t(m, \sigma) - \tfrac{1}{2} z^2 \bar{P}_t(\sigma) \right), \tag{1.27}$$

where $\bar{x}_t(m, \sigma)$ and $\bar{P}_t(\sigma)$ satisfy (1.21) and (1.22) respectively. With $F(a)$ given by (1.26) it follows from (1.16) that

$$\mathbf{e}_t(z) = \exp\left(iz\left(F_4 + \hat{\sigma}^2\left(F_2 + \frac{m}{\sigma^2} \right) F_3 \right) + z^2\left(F_5 - \frac{1}{2}\hat{\sigma}^2 F_3^2 \right) \right), \tag{1.28}$$

where

$$\hat{\sigma}^{-2} = \sigma^{-2} - 2F_1.$$

Comparing (1.27) and (1.28), we have

$$\bar{x}_t(m, \sigma) = F_4 + \hat{\sigma}^2\left(F_2 + \frac{m}{\sigma^2} \right) F_3, \tag{1.29}$$

and

$$\bar{P}_t(\sigma) = \hat{\sigma}^2 F_3^2 - 2F_5. \tag{1.30}$$

Letting now $\sigma \to 0$ in (1.29) and (1.30), it follows that

$$F_4 + mF_3 = \bar{x}_t(m, 0), \tag{1.31}$$

and

$$F_5 = -\tfrac{1}{2}\bar{P}_t(0). \tag{1.32}$$

The above allows $\mathbf{e}_t(z)$ to be of the form of (1.20) with F_1 and F_2 yet to be defined. Using now (1.17) and (1.18) and explicitly calculating $I_t(n)$, $n = 1, 2$, we have

$$\Delta_t(\sigma^{-2} - 2F_1) = F_3^2 \tag{1.33}$$

and

$$\Delta_t\left(\frac{m}{\sigma^2} + F_2 \right) = F_3(\hat{x}_t - F_4), \tag{1.34}$$

with

$$\Delta_t = P_t - \bar{P}_t(0).$$

The formulae for F_1 and F_2 will be obtained by differentiating (1.33) and (1.34). However, before this is done, recall from the theory of nonlinear filtering [3] that in

general, for x_t, y_t given as a solution to (1.1) and (1.2) $\hat{x}_t$, P_t satisfy

$$d\hat{x}_t = (f_0 + f_1\hat{x}_t)\,dt + (q_0 + P_t h_1)\,d\nu_t, \qquad \hat{x}_0 = \int_{-\infty}^{\infty} a\,dF(a), \tag{1.35}$$

$$\begin{aligned} dP_t &= (2f_1P_t + g_0^2 + q_0^2 - (q_0 + P_t h_1)^2)\,dt + h_1 R_t\,d\nu_t \\ P_0 &= \int_{-\infty}^{\infty} (a - \hat{x}_0)^2\,dF(a), \end{aligned} \tag{1.36}$$

where

$$R_t = \mathbf{E}((x_t - \hat{x}_t)^3 | \mathbf{Y}_t).$$

Remark. Direct application of (1.35) and (1.36) meets the difficulty of infinite coupling between the subsequent moments. From (1.22), (1.35), and (1.36), and the fact that for conditionally Gaussian processes $R_t = 0$,

$$d\Delta_t = \Delta_t(2f_1 - h_1(2q_0 + h_1(P_t + \bar{P}_t(0))))\,dt, \qquad \Delta_0 = \sigma^2. \tag{1.37}$$

Now from (1.33) and (1.34) (upon differentiation) and using (1.35), (1.37), it follows that

$$dF_1 = -\tfrac{1}{2}h_1^2 F_3^2\,dt, \qquad F_1(0) = 0, \tag{1.38}$$

and

$$dF_2 = F_3 h_1(d\nu_t + h_1 F_3 I_t(1)\,dt), \qquad F_2(0) = 0. \tag{1.39}$$

To define F_3, notice that (1.21) solves as

$$\begin{aligned} \bar{x}_t(a, \sigma) = \phi_t\Big(a + \int_0^t \phi_s^{-1}(f_0 - h_0(q_0 + \bar{P}_s(\sigma)h_1))\Big)\,ds \\ + \int_0^t \phi_s^{-1}(q_0 + \bar{P}_s(\sigma)h_1)\,dy_s), \end{aligned} \tag{1.40}$$

where

$$\phi_t = \exp\Big(\int_0^t (f_1 - h_1(q_0 + \bar{P}_s(\sigma)h_1))\Big)\,ds. \tag{1.41}$$

Comparing the above with (1.31) shows that $F_3 = \phi_t$ for $\sigma = 0$, which ends the proof of Lemma 2.

Lemma 3 below merely organizes all the results into the filter equations and the final form of the conditional characteristic function.

LEMMA 3. *Given the system* (1.1) *and* (1.2) *together with the a priori distribution* $F(a) = P(x_0 \leqq a | y_0)$. *The following are the* <u>*filter*</u> *equations* (*i.e., formulae of the recursive type, which calculate* $\hat{x}_t = \mathbf{E}(x_t | \mathbf{Y}_t)$).

$$d\hat{x}_t = (f_0 + f_1\hat{x}_t)\,dt + (q_0 + P_t h_1)\,d\nu_t, \; \hat{x}_0 = \int_{-\infty}^{\infty} a\,dF(a), \tag{1.42}$$

$$d\nu_t = dy_t - (h_0 + h_1\hat{x}_t)\,dt, \tag{1.43}$$

$$P_t = \bar{P}_t + \phi_t^2(I_t(2) - I_t^2(1)), \tag{1.44}$$

$$d\bar{P}_t = (2f_1\bar{P}_t + g_0^2 + q_0^2 - (q_0 + \bar{P}_t h_1)^2)\,dt, \qquad \bar{P}_0 = 0, \tag{1.45}$$

$$I_t(n) = \frac{\int_{-\infty}^{\infty} a^n \exp(a^2F_1 + aF_2)\,dF(a)}{\int_{-\infty}^{\infty} \exp(a^2F_1 + aF_2)\,dF(a)}, \qquad n = 1, 2, \tag{1.46}$$

$$dF_1 = -\tfrac{1}{2}h_1^2\phi_t^2\,dt, \qquad F_1(0)=0, \tag{1.47}$$

$$dF_2 = \phi_t h_1(dv_t + \phi_t h_1 I_t(1)\,dt), \qquad F_2(0)=0, \tag{1.48}$$

$$d\phi_t = (f_1 - h_1(q_0 + \bar{P}_t h_1))\phi_t\,dt, \qquad \phi_0 = 1. \tag{1.49}$$

The characteristic function of $x_t|\mathbf{Y}_t$ is given by:

$$\mathbf{e}_t(z) = \exp\left(ix(\hat{x}_t - \phi_t I_t(1)) - \tfrac{1}{2}z^2\bar{P}_t\right)\frac{\int_{-\infty}^{\infty}\exp\left(a^2F_1 + a(F_2 + iz\phi_t)\right)dF(a)}{\int_{-\infty}^{\infty}\exp\left(a^2F_1 + aF_2\right)dF(a)}. \tag{1.50}$$

2. Special cases and control problems. Two special cases of (1.1) and (1.2) result in significant simplification of the filter equations. The first case occurs when $g_0(t, y) = 0$, $0 \leqq t \leqq T$. From (1.7) it follows then that x_t is of the form

$$x_t = A_t(y)x_0 + B_t(y).$$

Using the above equation in (1.2), we have the following estimation problem: Let x_0 be a random variable with distribution $F(a) = P(x_0 \leqq a|y_0)$. Assume that the observation process y_t, $0 \leqq t \leqq T$, admits a differential

$$dy_t = (h_0(t, y) + h_1(t, y)x_0)\,dt + dv_t,$$

where the notation stays the same as in (1.2) and h_0, h_1 satisfy (1.3) and (1.4). From Lemma 2 it follows now that the conditional characteristic function of x_0 given $\mathbf{Y}_t$ is of the form

$$\mathbf{e}_t(z) = \frac{\int_{-\infty}^{\infty}\exp\left(a^2F_1 + aF_2 + iza\right)dF(a)}{\int_{-\infty}^{\infty}\exp\left(a^2F_1 + aF_2\right)dF(a)}. \tag{2.1}$$

The above results from the fact that $dx_0 = 0$ replaces (1.1) implying $\bar{P}_t(0) = 0$ and $\bar{x}_t(a, 0) = a$, as defined by (1.21) and (1.22). Now

$$\left.\frac{d\mathbf{e}_t(z)}{dz}\right|_{z=0} = i\hat{x}_t,$$

where $\hat{x}_t = \mathbf{E}(x_0|\mathbf{Y}_t)$, combined with the general filter equations (1.42) ÷ (1.50) yields

$$\hat{x}_t = \frac{\int_{-\infty}^{\infty} a\exp\left(-\tfrac{1}{2}a^2\int_0^t h_1^2\,ds + a\int_0^t h_1(dy_s - h_0\,ds)\right)dF(a)}{\int_{-\infty}^{\infty}\exp\left(-\tfrac{1}{2}a^2\int_0^t h_1^2\,ds + a\int_0^t h_1(dy_s - h_0\,ds)\right)dF(a)}. \tag{2.2}$$

In particular if

$$dF(a) = \frac{1}{\sqrt{2\pi}\,\sigma_0}\exp\left(-\frac{(a-m_0)^2}{2\sigma_0^2}\right)da,$$

(2.2) results in

$$\hat{x}_t = \frac{m_0 + \sigma_0^2\int_0^t h_1(dy_s - h_0\,ds)}{1 + \sigma_0^2\int_0^t h_1^2\,ds}. \tag{2.3}$$

The above agrees with the result presented in [4, pp. 22-24].

The second special case for which the filter takes a simple form follows if $h_1(t, y) = 0$ (i.e., the state is not observable directly), $0 \leqq t \leqq T$. Now the filter equations (1.42) ÷

(1.50) reduce to

$$
\begin{aligned}
&d\hat{x}_t = (f_0 + f_1\hat{x}_t)\,dt + q_0(dy_t - h_0\,dt),\\
&\hat{x}_0 = \int_{-\infty}^{\infty} a\,dF(a),\\
&dP_t = (2f_1P_t + g_0^2)\,dt,\\
&P_0 = \int_{-\infty}^{\infty} (a-\hat{x}_0)^2\,dF(a).
\end{aligned}
\tag{2.4}
$$

In order to discuss a control problem using the results obtained here, assume that all the coefficients in (1.1) and (1.2), except $f_0(t, y)$ which is denoted here by $u_t(y)$, are functions of time only. If $u_t(y)$ satisfies the assumption (1.5), we say that $u = \{u_t(y), 0 \leqq t \leqq T\}$ is an admissible control and write $u \in \mathbf{U}$.

Let x_t^u, $\hat{x}_t^u$ denote the solutions to (1.1) and (1.42) respectively, for some $u \in \mathbf{U}$, and let x_t^0, $\hat{x}_t^0$ correspond to $u_t \equiv 0, 0 \leqq t \leqq T$.

Define $e_t^u = x_t^u - \hat{x}_t^u$, $e_0^u = x_0 - \hat{x}_0 = e_0^0$, $e_t^0 = x_t^0 - \hat{x}_t^0$. Subtracting (1.42) from (1.1), we have

$$
de_t^u = f_1e_t^u dt + g_0dw_t + q_0dv_t - (q_0 + P_t(e^u)h_1)(h_1e_t^u + dv_t),
\tag{2.5}
$$

where $P_t = P_t(e^u)$ shows that P_t depends only on e_s^u, $0 \leqq s \leqq t$ which is seen from (1.48) rewritten as:

$$
dF_2 = \phi_t h_1(e_t^u h_1\,dt + dv_t + \phi_t h_1 I_t(1))\,dt.
\tag{2.6}
$$

From (2.5) it follows that with probability one the values of e_t^u and e_t^0 coincide for all $u \in \mathbf{U}$. Now, since

$$
d\nu_t^u = dy_t^u - (h_0 + h_1\hat{x}_t^u)\,dt = h_1e_t^u\,dt + dv_t = h_1e_t^0\,dt + dv_t = d\nu_t^0
\tag{2.7}
$$

ν_t^u and ν_t^0 coincide with probability one. From (2.7) it follows that (1.42) can be rewritten as

$$
d\hat{x}_t^u = (f_1\hat{x}_t^u + u_t)\,dt + (q_0 + h_1P_t(e_t^0))\,d\nu_t^0, \qquad \hat{x}_0^u = \hat{x}_0 = \int_{-\infty}^{\infty} a\,dF(a).
\tag{2.8}
$$

Now let $\tilde{u}_t = F_t(\hat{x}^{\tilde{u}})$, where F_t is a nonanticipative functional of $\hat{x}_s^{\tilde{u}}$, $0 \leqq s \leqq t$, and satisfies

$$
E\left(\int_0^T |F_t|^4\,dt\right) < \infty.
\tag{2.9}
$$

From (2.8) it follows that $\tilde{u}_t$ is σ-alg $\{\nu_s^0, 0 \leqq s \leqq t\}$ measurable. Now, let $\bar{u}_t$ be any admissible control and let $y_t^{\bar{u}}$ be an observation process associated with $\bar{u}_t$. From (2.7),

$$
\mathbf{Y}_t^{\bar{u}} = \sigma\text{-alg}\,\{y_s^{\bar{u}}, 0 \leqq s \leqq t\} \supseteq \sigma\text{-alg}\,\{\nu_s^{\bar{u}}, 0 \leqq s \leqq t\} = \sigma\text{-alg}\,\{\nu_s^0, 0 \leqq s \leqq t\}.
\tag{2.10}
$$

The above shows that $\tilde{u}_t$ is $\mathbf{Y}_t^{\bar{u}}$ measurable. This fact combined with (2.9) states that $\tilde{u}_t \in \mathbf{U}$, and that we can expect the separation of the stochastic control of $\tilde{u}_t$ type and the filtering problem. As an illustration of the statement, consider the following control problem.

Linear-quadratic control problem with non-Gaussian initial distributions. The partially observable controlled process (x_t, y_t), $0 \leqq t \leqq T$, is given by the stochastic equations

$$
\begin{aligned}
&dx_t = (f_1(t)x_t + u_t)\,dt + g_0(t)\,dw_t,\\
&dy_t = h_1(t)x_t\,dt + dv_t, \qquad y_0 = 0.
\end{aligned}
\tag{2.11}
$$

The independent Wiener processes w_t and v_t entering into (2.11) are also independent of the random variable x_0 (the initial state). x_0 is assumed to have distribution function $F(a) = P(x_0 \leqq a)$ with $\int_{-\infty}^{\infty} a^4\, dF(a) < \infty$ (finite fourth order moment).

The $\mathbf{Y}_t = \sigma\text{-alg}\{y_s, 0 \leqq s \leqq t\}$ measurable, stochastic process u_t is called a control at time t and is assumed to satisfy

$$\mathbf{E}\left(\int_0^T |u_t|^4\, dt\right) < \infty.$$

For $u = \{u_t, 0 \leqq t \leqq T\}$, satisfying the above we write $u \in \mathbf{U}$, where $\mathbf{U}$ is the class of admissible controls. It is also assumed that f_1, g_0, h_1 satisfy the deterministic version of (1.3) ÷ (1.5).

Consider now the performance functional

$$J(u) = \mathbf{E}\left(x_T^2 h_T + \int_0^T (x_t^2 H(t) + u_t^2 R(t))\, dt\right) \tag{2.12}$$

where $h_T \geqq 0$, $H(t) \geqq 0$, $0 \leqq R^{-1}(t) \leqq \text{const.}$, $0 \leqq t \leqq T$. The admissible control $\hat{u} \in \mathbf{U}$ is called optimal if

$$J(\hat{u}) = \inf_{u \in \mathbf{U}} J(u).$$

LEMMA. *The optimal control for the process* (2.11) *and the performance index* (2.12) *exists and is defined by*

$$\hat{u}_t = -R^{-1}(t) Q(t) \hat{x}_t, \qquad 0 \leqq t \leqq T \tag{2.13}$$

where $Q(t) \geqq 0$ *satisfies the Riccati equation*

$$-\frac{dQ(t)}{dt} = 2f_1(t) Q(t) + H(t) - Q^2(t) R^{-1}(t), \qquad Q(T) = h_T, \tag{2.14}$$

and $\hat{x}_t$ *is defined by*

$$\begin{aligned}
&d\hat{x}_t = (f_1(t) - R^{-1}(t) Q(t)) \hat{x}_t\, dt + P_t(\nu) h_1(t)\, d\nu_t,\\
&\hat{x}_0 = \int_{-\infty}^{\infty} a\, dF(a),\\
&d\nu_t = dy_t - h_1(t) \hat{x}_t\, dt,\\
&P_t = \bar{P}(t) + \phi(t)^2 (I_t(2) - I_t^2(1)),\\
&\frac{d\bar{P}(t)}{dt} = 2f_1(t) \bar{P}(t) + g_0^2(t) - h_1^2(t) \bar{P}^2(t), \qquad \bar{P}(0) = 0,\\
&\phi(t) = \exp\left(\int_0^t (f_1(s) - h_1^2(s) \bar{P}(s))\, ds\right),\\
&F_1(t) = -\tfrac{1}{2} \int_0^t h_1^2(s) \phi^2(s)\, ds,\\
&F_2(t, \nu) = \int_0^t \phi(t) h_1(t) (d\nu_t + \phi(t) h_1(t) I_t(1)\, dt),\\
&I_t(n) = \frac{\int_{-\infty}^{\infty} a^n \exp(a^2 F_1(t) + a F_2(t, \nu))\, dF(a)}{\int_{-\infty}^{\infty} \exp(a^2 F_1(t) + a F_2(t, \nu))\, dF(a)}, \qquad n = 1, 2.
\end{aligned} \tag{2.15}$$

Remark. The structure of the optimal control law is identical with the optimal controller for *LQG* problem.

Proof of the lemma. First note that the assumptions made in the control problem statement assure well-defined filter for the system (2.11) and $u \in \mathbf{U}$. Next rewrite the performance index as follows:

$$
\begin{aligned}
J(u) &= \mathbf{E}(\mathbf{E}(x_T^2 h_T | \mathbf{Y}_T) + \int_0^T \mathbf{E}(x_t^2 H(t) + u_t^2 R(t) | \mathbf{Y}_t)\, dt) \\
&= \mathbf{E}((\hat{x}_T^u)^2 h_T + \int_0^T ((\hat{x}_t^u)^2 H(t) + u_t^2 R(t))\, dt) \qquad (2.16) \\
&\quad + \mathbf{E}(h_T P_T^u + \int_0^T P_t^u H(t)\, dt).
\end{aligned}
$$

From (2.6) we conclude that P_t^u does not depend on the control u and coincides with the function P_t^0 obtained from the filter equations for $u_t \equiv 0$, $0 \leqq t \leqq T$. The process $\hat{x}_t^u$ entering (2.15) satisfies equation (see (2.8))

$$
d\hat{x}_t^u = (f_1(t)\hat{x}_t^u + u_t)\, dt + h_1(t) P_t^0\, d\nu_t^0, \qquad (2.17)
$$

where $\nu_t^0 = \nu_t^u = \int_0^t (dy_s^u - h_1(s)\hat{x}_s^u\, ds)$, according to (2.7), and ν_t^0 is a Wiener process (see (1.11)). Introduce now the function

$$
V(t, \xi) = \xi^2 Q(t) + \int_t^T Q(\tau) h_1^2(\tau) (P_\tau^0)^2\, d\tau \qquad (2.18)
$$

where $0 \leqq t \leqq T$, $-\infty < \xi < \infty$, and $Q(t)$ satisfies (2.14). It is easy to verify that $V(t, \xi)$ satisfies the following Bellman equation:

$$
\begin{aligned}
&\xi^2 H(t) + \xi f_1(t) \frac{\partial V(t, \xi)}{\partial \xi} + \frac{1}{2} (P_t^0)^2 h_1^2(t) \frac{\partial^2 V(t, \xi)}{\partial \xi^2} + \frac{\partial V(t, \xi)}{\partial t} \\
&\qquad + \min_\eta \left(\eta^2 R(t) + \eta \frac{\partial V(t, \xi)}{\partial \xi} \right) = 0, \qquad (2.19)
\end{aligned}
$$

and that $V(T, \xi) = \xi^2 h_T$. Note that $\hat{\eta}$ which minimizes the above for positive definite $R(t)$, is given by

$$
\hat{\eta} = -R^{-1}(t) Q(t) \xi. \qquad (2.20)
$$

Calculate now, with the use of the Ito formula,

$$
\begin{aligned}
V(T, \hat{x}_T^u) - V(0, \hat{x}_0) = \int_0^T \Bigg(\Bigg(\frac{\partial V(t, \xi)}{\partial t}\Bigg|_{\xi = \hat{x}_t^u} + \frac{1}{2} h_1^2(t) (P_t^0)^2 \frac{\partial^2 V(t, \xi)}{\partial \xi^2}\Bigg|_{\xi = \hat{x}_t^u} \Bigg) dt \\
+ \frac{\partial V(t, \xi)}{\partial \xi}\Bigg|_{\xi = \hat{x}_t^u} d\hat{x}_t^u \Bigg).
\end{aligned}
$$

Taking into account (2.8) and (2.19), we obtain

$$
V(T, \hat{x}_T^u) - V(0, \hat{x}_0) \geqq - \int_0^T ((\hat{x}_t^u)^2 H(t) + u_t^2 R(t))\, dt + 2 \int_0^T \hat{x}_t^u Q(t) P_t^0 h_1(t)\, d\nu_t^0.
$$

After taking the expectation of both sides of the above inequality,

$$
V(0, \hat{x}_0) \leqq \mathbf{E}((\hat{x}_T^u)^2 h_T + \int_0^T ((\hat{x}_t^u)^2 H(t) + u_t^2 R(t))\, dt). \qquad (2.21)
$$

The equality in (2.21) holds, according to (2.20), only if

$$\tilde{u}_t = -R^{-1}(t)Q(t)\hat{x}_t^{\tilde{u}}. \tag{2.22}$$

Comparing (2.21) with (2.16),

$$J(\tilde{u}) \leqq J(u) \quad \text{for all } u \in \mathbf{U}.$$

The admissibility of $\tilde{u}$ defined by (2.22) follows from (2.10) and the fact that

$$\mathbf{E}\left(\sup_{0 \leqq t \leqq T} (\hat{x}_t^{\tilde{u}})^4\right) < \infty.$$

The above can be proven in the same way as in the derivation of a conditionally Gaussian filter [4, Lemma 12.1; pp. 18-19]. This ends the proof of the lemma.

It seems to be possible, (following e.g. [2]) to show that the separation principle holds also for nonquadratic performance functionals. Obviously, the construction of the optimal control law might, in general, entail considerable analytical difficulties.

REFERENCES

[1] V. BENES AND I. KARATZAS, *Estimation and control for linear, partially observable systems with non-Gaussian initial distribution*, Stochastic Processes and Their Applications, 14 (1983), pp. 233-248.

[2] M. H. A. DAVIS, *The separation principle in stochastic control via Girsanov solutions*, SIAM J. Control, 14 (1976), pp. 176-188.

[3] R. S. LIPTSER AND A. N. SHIRYAYEV, *Statistics of Random Processes* I—*General Theory*, Springer-Verlag, New York, 1977.

[4] ———, *Statistics of Random Processes* II—*Applications*, Springer-Verlag, New York, 1978.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 3, May 1986

011

GLOBAL REALIZATIONS OF ANALYTIC INPUT-OUTPUT MAPPINGS*

J. P. GAUTHIER† AND G. BORNARD†

Abstract. A sufficient condition is presented for the existence and the uniqueness of a global minimal analytic realization of a nonlinear analytic input-output mapping, when the latter is defined on an open subset of the semigroup of the inputs defined for positive times.

The state space of this realization is exhibited as a quotient space of a Riemann surface on a Lie group.

Key words. nonlinear systems, realization, Lie groups, sheaves

AMS(MOS) subject classifications. 93C10, 93B15, 93B20, 22E99

1. Introduction. This paper deals with the problem of finding a finite dimensional minimal analytic realization of a given analytic input-output mapping (precise definitions will be given in § 2).

Conditions of existence and uniqueness of a global realization have been given by Jakubczyck [3], in the case where the input-output mapping is defined on the whole group of the inputs, i.e. for all positive and negative times.

Here we are interested in the case where the given input-output mapping is defined only on a subset of the semigroup S of the inputs corresponding to positive times. This case seems to be of importance from a control point of view.

The following small example points out the kind of difficulties that arise when considering uncompletely defined input-output mappings. Consider the input-output mapping P generated by the following equations:

Example 1.

$$\dot{x} = \frac{1}{x+u}, \qquad x \in R, \quad u \in R^{+}*,$$

$$y = x, \qquad y \in R \quad \text{(output)}.$$

$$x(0) = 0,$$

Clearly P is defined for every piecewise constant input function and every positive time i.e., P is defined on S. Remark also that the rank defined by Jakubczyk is 1.

However, it is not possible to find an analytic manifold and a family of everywhere defined analytic vector fields on it that realizes P: on any neighborhood of 0, there exists some $u > 0$ such that $\dot{x}$ is not defined.

In the present paper, sufficient conditions are given for the existence of local and global realizations when P is defined on a subset of S.

When the input-output mapping is defined for all positive and negative times, the global condition becomes necessary and sufficient for the existence of a "right invariant" realization on Lie Group.

Section 2 is devoted to general definitions and notations, § 3 presents two technical lemmas and a local realization one. The main result is developed in § 4, and an illustrative example is detailed in § 5.

2. Definitions and notation. Let U be the set of the values of the input (U is any set, with no structure needed), and $\mathcal{G}$ be the monoïd generated by ε, the empty sequence,

* Received by the editors January 24, 1984, and in revised form April 3, 1985.

† Laboratoire d'Automatique de Grenoble, ENSIEG-INPG, CNRS LA 228, 38402 Saint-Martin d'Hères, France.

and all the sequences with elements in $U \times R$, under the operation of concatenation. $\mathcal{G}$ can be endowed with the natural topology induced by the topology of R^∞.

The notation $\underline{u}(\underline{t}) = (u_k, t_k) \cdots (u_1, t_1)$, $\underline{u} \in U^k$, $\underline{t} \in R^k$, $k \in N$, will be used throughout the paper for the elements of $\mathcal{G}$.

Let G be the group obtained as the quotient of $\mathcal{G}$ through the equivalence relation $R_{\mathcal{G}}$ defined by:

$$\left.\begin{array}{l}(v,0)R_{\mathcal{G}}\varepsilon \\ \underline{u}(\underline{r})(v,0)\underline{w}(s)R_{\mathcal{G}}\underline{u}(\underline{r})\underline{w}(\underline{s}) \\ \underline{u}(\underline{r})(v,t)(v,t')\underline{w}(\underline{s})R_{\mathcal{G}}\underline{u}(\underline{r})(v,t+t')\underline{w}(\underline{s})\end{array}\right\} \begin{array}{l}\text{for every } \underline{u}(\underline{r}) \in \mathcal{G},\ \underline{w}(\underline{s}) \in \mathcal{G}, \\ v \in U,\ t \in R,\ t' \in R.\end{array}$$

An element $\underline{u}(\underline{t})$, $\underline{u} \in U^k$, $\underline{t} \in R^k$, is said to be minimal when k is minimal among all the elements of the class of $\underline{u}(\underline{t})$ for $R_{\mathcal{G}}$. Each class has exactly one minimal element.

The notation $\underline{u}^{-1}(\underline{t}) = (u_1, -t_1) \cdots (u_k, -t_k)$ will also be used for denoting the inverse of the element $\underline{u}(\underline{t}) = (u_k, t_k) \cdots (u_1, t_1)$.

G is given the quotient topology. It can also be given a natural analytic structure (in the sense given in the remark below).

Replacing R, R^∞, $\mathcal{G}$, G, R by R^+, $R^{+\infty}$, $\mathcal{S}$, S, $R_{\mathcal{S}}$ respectively in the previous definitions, one obtains the monoïd $\mathcal{S}$, the semigroup $S = \mathcal{S}/R_{\mathcal{S}}$.

Remarks. (a) For the topology so defined on S, a subset W of S is open iff, for every $k \in N$, $\underline{u} \in U^k$, the subset $W_{\underline{u}} = \{\underline{t} \in R^{+k} | \underline{u}(\underline{t}) \in W\}$ is open in R^{+k}.

Let W be an open subset of S, Y an analytic manifold. The analytic structure of S is defined in the following way: a mapping $P: W \to Y$ is analytic iff, for every $k \in N$, $\underline{u} \in U^k$, the mapping $P_{\underline{u}}(\underline{t}) = P(\underline{u}(\underline{t})): W_{\underline{u}} \to Y$ is analytic.

(b) $\mathcal{S}$ is a subset of $\mathcal{G}$, $R_{\mathcal{S}}$ is the restriction to $\mathcal{S}$ of $R_{\mathcal{G}}$. This defines a canonical inclusion of S in G. It can be shown that the topology of S is exactly that which is induced by the topology of G through this inclusion.

From a control point of view, G and S are the group and semigroup of piecewise-constant inputs for all times and for only positive times respectively.

In the paper, an analytic input-output mapping P will be an analytic mapping from D_P to some analytic manifold Y, where D_P is either G or a connected open subset of S.

It will be useful to define the action $*$ of R^+ on $R^{+\infty}$ and S:

$$t * \underline{s} = \begin{cases} (s_1, \cdots, s_i, t - \eta_i, 0, \cdots, 0) & \text{for } \eta_i \leqq t \leqq \eta_{i+1}, \\ (s_1, \cdots, s_k) & \text{for } \eta_k \leqq t, \end{cases}$$

$$t * \underline{u}(\underline{s}) = \underline{u}(t * \underline{s}),$$

where

$$t \in R^+, \quad \underline{s} = (s_1, \cdots, s_k) \in R^{+k}, \quad \underline{u} \in U^k, \qquad k \in N,$$

$$\eta_i = \sum_{j=1}^{i} s_j, \qquad i = 1, \cdots, k.$$

A subset W of S is said to be "truncation-closed" if $R^+ * W \subset W$.

Remarks.

- In the literature, an input-output mapping is often defined in an alternate way, (as a "system" for example), which implicitly insures that D_P is truncation-closed.
- S is truncation-closed.
- A truncation-closed subset of S is connected.

• When D_P is a truncation-closed open subset of S, the mapping $P_{\underline{u}}(\underline{t}) = P(\underline{u}(\underline{t}))$ is defined on a connected neighborhood of 0 in R^{+k} for every $k \in N$, $\underline{u} \in U^k$.

Consider some analytic input-output mapping $P: D_P \to Y$. The following standard notation will be used throughout the paper:

$$a \in D_P, \quad a' \in D_P, \quad \underline{b} \in D_P^q, \quad \underline{b}' \in D_P^{q'},$$

$$a = \underline{u}(\underline{t}) \in D_P, \quad \underline{u} \in U^k, \quad \underline{t} \in R^k,$$

$$\underline{b} = (b_1, \cdots, b_q) = (\underline{v}_1(\underline{s}_1), \cdots, \underline{v}_q(\underline{s}_q)) = \underline{\underline{v}}(\underline{s}), \qquad q \in N, \underline{b} \in D_P^q,$$

$$\underline{v}_i \in U^{q_i}, \underline{s}_i \in R^{q_i}, i = 1, \cdots, q,$$

$$\underline{\underline{v}} \in U^{\bar{q}}, \underline{s} \in R^{\bar{q}}, \bar{q} = \sum_{i=1}^{q} q_i,$$

$$\Psi_{a,a'}^{\underline{b},\underline{b}'} = (\Psi_{a,a'}^{b_1,b_1'}, \Psi_{a,a'^2}^{b_1,b_2'}, \cdots, \Psi_{a,a'}^{b_q,b_q'})$$

where $\Psi_{a,a'}^{b_i,b_j'} = P(b_i a b_j' a')$.

$\Psi_{\underline{u}(\underline{t}),\underline{u}'(\underline{t}')}^{\underline{b},\underline{b}'}$ is an analytic mapping from an open subset of $R^k \times R^{k'}$ to $Y^{qq'}$, when considered as a function of $\underline{t}$ and $\underline{t}'$.

The notation $\Psi_a^{\underline{b},\underline{b}'}$ will also be used for $\Psi_{a,\varepsilon}^{\underline{b},\underline{b}'}$ and $\Psi_a^{\underline{b}}$ for $\Psi_a^{\underline{b},\varepsilon}$.

Let Ψ_0 be the family of analytic mappings:

$$\Psi_0 = \{\Psi_{\underline{u}(\underline{t})}^{\underline{b},\underline{b}'} | \underline{b}, \underline{b}' \in G^\infty, \underline{u}(\underline{t}) \in G \text{ such that } b_i \underline{u}(\underline{t}) b_j' \in D_P, \forall i, j\},$$

and Ψ_1 the family Ψ_0 restricted to $\underline{b}' = \varepsilon$. The rank n_0 of the family Ψ_0 is the maximum of the ranks of the Jacobian matrices $(\partial/\partial \underline{t})(\Psi_{\underline{u}(\underline{t})}^{b_i,b_j'})$ when the arguments cross the domain.

The rank n_1 is defined in the same way on Ψ_1. Clearly, n_1 is the rank defined by Jakubczyk [3].

This notation will also be used, depending on the context, when replacing G by S, R by R^+.

An analytic system defined on the input space U and the output analytic manifold Y is a 4-tuple:

$$Q = \{M, x_0, F = \{X_u | u \in U\}, h\}$$

where M is a finite dimensional connected manifold, $x_0 \in M$, F a family of analytic vector fields on M, h an analytic mapping from M to Y. Q defines in an obvious way, considering positive times, an input-output mapping P_Q, with domain D_Q truncation-closed and open in S.

A local analytic realization of an input-output mapping P, "after a," for some element $a \in D_P$, is an analytic system Q such that:

• $P(ba) = P_Q(b)$ for every $b \in D_Q$ such that $ba \in D_P$.

• The ranks n_1 (and n_0 if n_0 is finite) are the same for P and P_Q.

A (global) analytic realization of P is an analytic system Q such that $D_P \subset D_Q$ and that $P = P_Q | D_P$.

A (local or global) analytic realization Q is said to be minimal when Q is weakly controllable, and observable (see Sussmann [5]).

3. Lemmas. Let M be an analytic connected manifold and $F = \{X_u | u \in U\}$ a family of analytic vector fields on M. Consider the following notations:

$$M^1 = M, \quad X_u^1 = X_u, \quad u \in U, \quad F^1 = F,$$

$$M^{k+1} = M^k \times M^k, \; X_u^{k+1} = X_u^k \oplus X_u^k, \; u \in U, \; F^{k+1} = \{X_u^{k+1} | u \in U\} \text{ for } k = 2, 3, \cdots.$$

The Lie algebras $L(F)$ and $L(F^k)$ generated by F and F^k respectively, considered as subalgebras of the Lie algebras of vector fields on M and M^k, are naturally isomorphic.

Let n^k be the maximal dimension of the orbits of F^k in M^k (see Sussmann [6]). The sequence n^k is increasing, since every orbit of M^k identifies with the corresponding orbit in the diagonal of M^{k+1}.

LEMMA 1 (orbits of finite dimension). *The following two statements are equivalent.*

(a) *$L(F)$ is a Lie algebra of finite dimension.*

(b) *The sequence n^k is stationary, with limit n.*

Proof. (a) $\rightarrow$ (b). $L(F^k)$ being a Lie algebra of finite dimension n, $L(F^k)$ evaluated at $x_0 \in M^k$ can be only of dimension $n' \leqq n$. Then, because of the analyticity, the orbit through x_0 is of dimension $n' \leqq n$. The sequence n^k, being increasing and bounded is stationary. The fact that its limit is exactly n is proved later.

(b) $\rightarrow$ (a). Assume that for some k, $n^{k+1} = n^k$. Let x_0 belong to an orbit of M^k of dimension n^k, let $\{Y_i^k \in L(F^k),\ i = 1, \cdots, n^k\}$ be a set of n^k vector fields, linearly independent at x_0, and let V^k be a connected neighborhood of x_0 in M^k, such that the Y_i^k are linearly independent on V^k, and the $Y_i^{k+1} = Y_i^k \oplus Y_i^k$ are also linearly independent on the neighborhood $V^{k+1} = V^k \times V^k$ of (x_0, x_0) in M^{k+1}.

Consider Z^k, any vector field in $L(F^k)$ and the corresponding vector field $Z^{k+1} = Z^k \oplus Z^k$ in $L(F^{k+1})$. Because $n^{k+1} = n^k$, Z^{k+1} expresses linearly, by analytic functions, as a function of $(Y_i^{k+1},\ i = 1, \cdots, n^k)$. If not, there would exist an orbit of dimension $n'' > n^{k+1}$ in M^{k+1}, which contradicts the definition of n^{k+1}. Then:

$$Z^{k+1}(x, y) = \sum_{i=1}^{n^k} \gamma_i(x, y)\, Y_i^{k+1}(x, y) \quad \text{for } x \in V^k, \quad y \in V^k$$

and in particular by projections on the first factor:

$$Z^k(x) = \sum_{i=1}^{n^k} \gamma_i(x, y)\, Y_i^k(x).$$

In some local coordinate system, one obtains:

$$\frac{\partial Z(x)}{\partial y_j} = \sum_{i=1}^{n^k} \frac{\partial \gamma_i(x, y)}{\partial y_j}\, Y_i^k(x) = 0$$

and then $\partial \gamma_i(x, y)/\partial y_j = 0$ for $j = 1, \cdots, n^k$, since the vector fields $Y_i^k(x)$ are linearly independent. x and y playing the same role, it follows that the $\gamma_i(x, y)$ are constant functions. In particular, one has

$$[Y_i^k, Y_j^k] = \sum_{m=1}^{n^k} \gamma_{ij}^m\, Y_m^k$$

with constant γ_{ij}^m.

Since this is true on a neighborhood of x_0, it must be true everywhere on M^k, by analyticity and connectedness. The γ_{ij}^m are the structure constants of the Lie algebra $L(F)$ which is then of dimension n^k.

Clearly one has $n^i = n^k$ for $i > k$, and then $n^k = n$. This completes the proof of Lemma 1.

LEMMA 2 (lift on a Lie group). *Let $F = \{X_u | u \in U\}$ be a family of analytic vector fields on some connected open set $V_0 \subset R^n$, $x_0 \in V_0$ and assume that:*

- *$L(F)$ the Lie algebra generated by F has dimension n;*
- *$L(F)(x_0)$ spans $T_{R^n}(x_0)$.*

Then there exist a real simply connected Lie group $\tilde{G}$ of dimension n, with identity e, and an analytic diffeomorphism Ψ from $V_1 \subset V_0$ on some neighborhood W of e in $\tilde{G}$, such that $x_0 \in V_1$ and:

- $\Psi(x_0) = e$.
- *For every $u \in U$, $\tilde{X}_u = d\Psi X_u$ is the restriction to W of some right invariant vector field on $\tilde{G}$.*

Proof. The proof, not given here, is a consequence of the third Lie's theorem ([9, p. 551] for example).

LEMMA 3 (local realization). *Let P be an analytic input-output mapping from a truncation-closed open subset D_P of S to an analytic manifold Y. Assume that the rank n_1 of the associated family Ψ_1 is finite (see notation in § 2).*

Then there exists a local minimal analytic realization Q of P, of dimension n_1, after a, for some element $a \in D_P$.

Proof. In the following, all the mappings under consideration are analytic. Assume that n_0 is finite. By hypothesis there exist $\underline{u}_1(t_1) \in S$, $\underline{u}_2(t_2) \in S$, $\underline{v}(\underline{s}_1) \in S^{q_1}$, $\underline{v}_2(\underline{s}_2) \in S^{q_2}$, $\underline{v}'(\underline{s}'_1) \in S^{q'}$, $q_1, q_2, q' \in N$, such that:

- The rank of $\Psi^{\underline{v}_1(\underline{s}_1), \underline{v}'(\underline{s}'_1)}_{\underline{u}_1(\underline{t})}$ with respect to $\underline{t}$ is n_0 at $\underline{t} = \underline{t}_1 \in R^{+k_1}$.
- The rank of $\Psi^{\underline{v}_2(\underline{s}_2)}_{\underline{u}_2(\underline{t})}$ with respect to $\underline{t}$ is n_1 at $\underline{t} = \underline{t}_2 \in R^{+k_2}$.

Denote $\underline{u} = (\underline{u}_1, \underline{u}_2) \in U^k$, $\underline{v} = (\underline{v}_1, \underline{v}_2) \in U^{\bar{q}}$, and consider the mappings:

$$\Omega_0(\underline{s}, \underline{s}', \underline{t}', \underline{t}) = \Psi^{\underline{v}(\underline{s}), \underline{v}'(\underline{s}')}_{\underline{u}(\underline{t}'), \underline{u}(\underline{t})} : R^{+\bar{q}} \times R^{+\bar{q}'} \times R^{+k} \times R^{+k} \to Y^{qq'},$$

$$\Omega_1(\underline{s}, \underline{t}', \underline{t}) = \Omega_0(\underline{s}, 0, \underline{t}', \underline{t}) : R^{+\bar{q}} \times R^{+k} \times R^{+k} \to Y^q,$$

$$\Omega_2(\underline{s}, \underline{t}) = \Omega_1(\underline{s}, 0, \underline{t}) : R^{+\bar{q}} \times R^{+k} \to Y^q.$$

By construction, the ranks of Ω_0, Ω_1 with respect to $\underline{t}'$ and the rank of Ω_2 with respect to $\underline{t}$ are n_0, n_1, n_1 at $((\underline{s}_1, 0), \underline{s}'_1, (\underline{t}_1, 0), 0)$, $((0, \underline{s}_2), (0, \underline{t}_2), 0)$, $((0, \underline{s}_2), (0, \underline{t}_2))$ respectively.

Since D_P is truncation-closed, the domains of $\Omega_0, \Omega_1, \Omega_2$ are connected neighborhoods of 0 in the corresponding spaces. Moreover, the ranks n_0, n_1, n_1 are maximal. Then, by analyticity the ranks n_0, n_1, n_1 are reached almost everywhere on the domains of $\Omega_0, \Omega_1, \Omega_2$.

It is then possible to find $\underline{s}_0, \underline{s}'_0, \underline{t}_0$ such that:

- The rank of Ω_2 is n_1 at $(\underline{s}_0, \underline{t}_0)$.
- There exists $\underline{t}'_0$ arbitrarily small, such that the ranks of Ω_0, Ω_1 are n_0, n_1 at $(\underline{s}_0, \underline{s}'_0, \underline{t}'_0, \underline{t}_0)$, $(\underline{s}_0, \underline{t}'_0, \underline{t}_0)$ respectively.

In the case where n_0 is infinite, the same argument applies for the choice of $\underline{s}_0, \underline{t}_0$, considering only Ω_1, Ω_2 and n_1.

Denoting $\underline{b} = (\underline{v}(\underline{s}_0))$, the rank of $\Psi^{\underline{b}}_{\underline{u}(\underline{t})}$ is n_1 at $\underline{t}_0$. It is then possible to find two analytic connected submanifolds V_0, Y_0, of R^{+*k}, Y^q respectively, of dimension n_1, such that the restriction of $\Psi^{\underline{b}}_{\underline{u}(\underline{t})}$ to V_0 is a diffeomorphism of V_0 on Y_0. The canonical immersion of V_0 in R^k will be denoted $\underline{t} = \underline{t}(\underline{\tau})$.

Consider the mapping

$$\bar{\Phi}^{\underline{b}}_{\underline{u}}(\theta, \underline{\tau}) = \Psi^{\underline{b}}_{(u,\theta)\underline{u}(\underline{t}(\underline{\tau}))},$$

which is defined on some connected neighborhood $\bar{W}_0$ of $(0, V_0)$ in $R^+ \times V_0$, since D_P is open. There exists a connected neighborhood W_0 of $(0, V_0)$ in $R \times V_0$ such that

- $\bar{\Phi}^{\underline{b}}_{\underline{u}}$ extends in a unique way in $\Phi^{\underline{b}}_{\underline{u}}$ on W_0.
- $\Phi^{\underline{b}}_{\underline{u}}(\theta, \underline{\tau}) \in Y_0$ for every $(\theta, \underline{\tau}) \in W_0$, since n_1 is maximal.

Consider, on W_0, the equality

$$\Phi_u^b(0, \underline{\tau}) = \Phi_u^b(\theta, \underline{\gamma}). \tag{L1}$$

For any $\underline{\gamma}_0 \in V_0$, this equality holds for $\underline{\tau} = \underline{\gamma} = \underline{\gamma}_0$, $\theta = 0$, and the rank of $\Phi_u^b(0, \underline{\tau})$ with respect to $\underline{\tau}$, and the rank of $\Phi_u^b(\theta, \underline{\gamma})$ with respect to $(\theta, \underline{\gamma})$ are both n_1 at this point.

From the implicit function theorem, there exists a connected neighborhood W_1 of $(0, \underline{\gamma}_0)$ in W_0, and a unique analytic mapping σ: $W_1 \to V_0$ such that:

$$\Phi_u^b(0, \sigma(\theta, \underline{\gamma})) = \Phi_u^b(\theta, \underline{\gamma}) \quad \text{for every } (\theta, \underline{\gamma}) \in W_1. \tag{L2}$$

Denote $V_1 = \{\underline{\gamma} \in V_0 | (0, \underline{\gamma}) \in W_1\}$.

For any $\underline{w} \in U^p$, $p \in N$, there exist $\underline{r} \in R^{+p}$ small enough, and a connected neighborhood V_2 of $\underline{\gamma}_0$ in V_1, such that $\underline{w}(\underline{r})\underline{u}(\underline{t}(\underline{\gamma})) \in D_P$ for every $\underline{\gamma} \in V_2$. Denote $c = \underline{w}(\underline{r})$ and consider the mapping:

$$\Phi_u^c(\theta, \underline{\gamma}) = \Psi^c_{(u,\theta)\underline{u}(\underline{t}(\underline{\gamma}))}$$

which is defined on some connected neighborhood W_2 of $(0, \underline{\gamma}_0)$ in $R \times V_2$, through a small extension of $\Psi^c_{(u,\theta)\underline{u}(\underline{t}(\underline{\gamma}))}$ for $\theta < 0$. An intermediate step is to show that:

$$\Phi_u^c(0, \sigma(\theta, \underline{\gamma})) = \Phi_u^c(\theta, \underline{\gamma}), \tag{L3}$$

for $(\theta, \underline{\gamma}) \in W_3$, some neighborhood of $(0, \underline{\gamma}_0)$ in W_2.

Consider now the equality:

$$\begin{aligned} &\text{(a)} \quad \Phi_u^b(0, \underline{\tau}) = \Phi_u^b(\theta, \underline{\gamma}), \\ &\text{(b)} \quad \Phi_u^c(0, \underline{\tau}) = \Phi_u^c(\theta, \underline{\gamma}). \end{aligned} \tag{L4}$$

(L4) holds for $\underline{\tau} = \underline{\gamma} = \underline{\gamma}_0$, $\theta = 0$. Denoting $\Phi_u^{b,c}$ for (Φ_u^b, Φ_u^c), the rank of $\Phi_u^{b,c}(0, \underline{\tau})$ with respect to $\underline{\tau}$ and the rank of $\Phi_u^{b,c}(\theta, \underline{\gamma})$ with respect to $(\theta, \underline{\gamma})$ are both n_1 at this point (since n_1 is maximal).

From the implicit function theorem there exist a connected neighborhood $W_3 \subset W_2$ of $(0, \underline{\gamma}_0)$ and a unique mapping $\bar{\sigma}$: $W_3 \to V_0$ such that:

$$\Phi_u^{b,c}(0, \bar{\sigma}(\theta, \underline{\gamma})) = \Phi_u^{b,c}(\theta, \underline{\gamma}) \quad \text{for every } (\theta, \underline{\gamma}) \in W_3.$$

But the solution of (L4a), unique on $W_3 \subset W_2$, is already given by σ. Then $\bar{\sigma} = \sigma|_{W_3}$, and (L3) holds.

Considering now the case $c = b_i(u, \delta)$, $i = 1, \cdots, q$ and using the fact that $(u, \delta)(u, \theta) = (u, \delta + \theta)$, an easy calculation based upon the preceding results shows that $\sigma(\delta, \sigma(\theta, \underline{\gamma})) = \sigma(\delta + \theta, \underline{\gamma})$ for $\underline{\gamma}$ near $\underline{\gamma}_0$ and δ, θ small enough. Moreover it is clear that $\sigma(0, \underline{\gamma}) = \underline{\gamma}$.

Consequently σ defines an analytic local one parameter subgroup on a neighborhood of $\underline{\gamma}_0$ in V_0. The same process can be achieved for every $\underline{\gamma} \in V_0$, and it is straightforward to see that all these local subgroups fit conveniently and define an analytic local one parameter subgroup on V_0.

Note X_u for the corresponding vector fields on V_0. The same construction can be performed for every $u \in U$.

A family of analytic vector fields has been defined on the n_1-dimensional analytic manifold V_0. Consider the output mapping h: $V_0 \to Y$ defined by:

$$h(\underline{\tau}) = P(\underline{u}, \underline{t}(\underline{\tau})).$$

Then the system:

$$Q = \{V_0, \underline{\tau}_0, F = \{X_u | u \in U\}, h\}$$

is a candidate to realize P locally after $\underline{u}(\underline{t}_0)$.

One has to show that, for any $p \in N$, $\underline{w} \in U^p$, one has

$$\text{(L5)} \qquad h(\exp X_{w_p} r_p \circ \cdots \circ \exp X_{w_1} r_1(\tau_0)) = P((w_p, r_p) \cdots (w_1, r_1)\underline{u}(\underline{t}(\tau_0)))$$

for every $\underline{r} \in R^{+p}$ such that both members are defined.

In (L2) and (L3) one can now replace $\sigma(\theta, \gamma)$ by $\exp X_u \theta(\gamma)$, which is defined on V_0, for small θ. From (L3) with $c = (w_p, r_p) \cdots (w_2, r_2)$, and for r_1 small enough, one obtains:

$$P((u_p, r_p) \cdots (u_2, r_2)(u_1, r_1)\underline{u}(\underline{t}(\tau_0))) = P((u_p, r_p) \cdots (u_2, r_2)\underline{u}(\underline{t}(\exp X_{u_1} r_1(\tau_0)))).$$

Iterating this operation, and using the definition of h, one obtains (L5) for small times.

Consider any $\underline{r}_0 \in R^{+p}$ such that $\underline{w}(\underline{r}_0) \in D_Q$ and $\underline{w}(\underline{r}_0)\underline{u}(t_0) \in D_P$. D_P and D_Q are both truncation-closed and both members of (L5) are defined along the same truncation-closed path from $\underline{r}_0$ to 0 in R^{+p}. It follows that they are defined on a connected open subset of R^{+p} containing $\underline{r}_0$ and 0. Since (L5) holds on a neighborhood of 0 in R^{+p}, then it holds also at $\underline{r}_0$.

The rank n_1' associated to P_Q cannot be greater than n_1. Moreover, from the choice of $\underline{s}_0$ and $\underline{t}_0$ at the beginning of the proof, there exists $\underline{t}_0' \in R^{+p}$ such that $\underline{u}(\underline{t}_0') \in D_Q$ and that Ω_1 is of rank n_1 at $(\underline{s}_0, \underline{t}_0', \underline{t}_0)$. Then $n_1' = n_1$. The same applies for n_0 when it is finite.

By construction, Q is weakly controllable, observable. This ends the proof of Lemma 3.

4. Main result: global realization.

THEOREM 1. *Let P be an analytic input-output mapping from an open truncation-closed subset D_P of S to an analytic manifold Y. Assume that the rank n_0 of the family Ψ_0 of mappings associated to P is finite, and let n_1 be of the rank of the associated family Ψ_1.*

Then there exists a global minimal analytic realization of P, of dimension n_1.

Proof. The proof is achieved in three steps.

(a) *Finite dimensional orbits and lift on a Lie group.* The rank $n_1 \leqq n_0$ is finite. From Lemma 3 there exists a local realization $Q = \{V_0, x_0 = \tau_0, F = \{X_u | u \in U\}, h\}$ associated to the sequences $\underline{u}_0(\underline{t}(\tau_0))$, $\underline{b} = \underline{v}(\underline{s}_0)$, where V_0 can be restricted to be relatively compact. Consider, for the system Q, the following notation:

$$\hat{b}(x) = \exp X_{v_p} s_p \circ \cdots \circ \exp X_{v_1} s_1(x) \quad \text{for } x \in V_0, \quad b = \underline{v}(\underline{s}) \in G,$$

$$\underline{\hat{b}}(x)(\hat{b}_1(x), \cdots, \hat{b}_q(x)) \quad \text{for } \underline{b} \in G^q,$$

$$\bar{\Phi}^{\underline{x}}_{\underline{u}(\underline{t})} = (\hat{\underline{u}}(\underline{t})(x_1), \cdots, \hat{\underline{u}}(\underline{t})(x_{q'})) : R^k \to V_0^{q'} \quad \text{for} \begin{cases} k \in N, \\ \underline{u} \in U^k, \\ \underline{t} \in R^k, \\ \underline{x} \in V_0^{q'}, \end{cases}$$

$$\bar{\Psi}^{\underline{b}}(\underline{x}) = (h(\hat{b}_1(x_1)), \cdots, h(\hat{b}_q(x_1)), \cdots, h(\hat{b}_q(x_{q'}))) : V_0^{q'} \to Y^{qq'},$$

$$\bar{\Psi}^{\underline{b},\underline{x}}_{\underline{u}(\underline{t})} = \bar{\Psi}^{\underline{b}} \circ \bar{\Phi}^{\underline{x}}_{\underline{u}(\underline{t})}.$$

Clearly, one has:

$$\bar{\Psi}^{\underline{b},\underline{x}}_{\underline{u}(\underline{t})} = \Psi^{\underline{b},(\underline{u}_0(\underline{t}(x_1)), \cdots, \underline{u}_0(\underline{t}(x_{q'})))}_{\underline{u}(\underline{t})} \quad \text{for } \underline{x} \in V_0^{q'}, \quad \underline{b} = \underline{v}(\underline{s}), \quad \underline{t} \text{ and } \underline{s} \text{ small.}$$

The rank of $\Psi^{\underline{b}\underline{x}}_{\underline{u}(\underline{t})}$ with respect to $\underline{t}$ is then $\leqq n_0$.

Consider now, for $q'=2^p$, $F^{q'}$, the lift of the family F in $V_0^{q'}=V_0\times\cdots\times V_0$, and $O(\underline{x})$, the orbit of $F^{q'}$ through $\underline{x}\in V_0^{q'}$. One has:

$$O(\underline{x})=\{\underline{x}'\in V_0^{q'}|\underline{x}'=\bar{\Phi}^{\underline{x}}_{\underline{u}(t)} \text{ for } \underline{u}(t)\in G\}.$$

Denote by n the maximal dimension of the orbit $O(x)$ when $\underline{x}$ crosses $V_0^{q'}$, which has dimension $q'n_1$. This dimension is reached almost everywhere on $V_0^{q'}$, and for any $\underline{x}$ such that this dimension is reached, there exists $\underline{u}\in U^k$, $k\in N$, such that the rank of $\bar{\Phi}^{\underline{x}}_{\underline{u}(t)}$ with respect to $\underline{t}$ is n almost everywhere on a neighborhood of 0 in R^k. Meanwhile, there exists $\underline{b}\in G^\infty$ such that $\bar{\Psi}^{\underline{b}}$ has rank $q'n_1$ almost everywhere on some connected open subset of $V_0^{q'}$. Then one can find $\underline{x}\in V_0^{q'}$, k, $q\in N$, $\underline{u}\in U^k$, $\underline{t}_1\in R^k$, $\underline{b}\in G^q$, such that:

- $O(\underline{x})$ has maximal dimension n in $V_0^{q'}$.
- $\bar{\Phi}^{\underline{x}}_{\underline{u}(t)}$ has rank n (maximal) at $\underline{t}_1$ and then on some neighborhood W_1 of $\underline{t}_1$ in R^k.
- $\bar{\Psi}^{\underline{b}}$ has rank $q'n_1$.

The situation is as follows:

$$W_1 \xrightarrow{\bar{\Phi}^{\underline{x}}_{\underline{u}(t)}} V_0^{q'} \xrightarrow{\bar{\Psi}^{\underline{b}}} Y^{qq'}.$$

Sylvester's inequality applied to some Jacobian matrices, states that

$$\operatorname{rank}\bar{\Phi}^{\underline{x}}_{\underline{u}(t)}+\operatorname{rank}\bar{\Psi}^{\underline{b}}-\dim V_0^{q'}\leqq\operatorname{rank}\bar{\Psi}^{\underline{b},\underline{x}}_{\underline{u}(t)}.$$

Then $n\leqq n_0$.

Since the maximal rank n_0 of $\bar{\Psi}^{\underline{b}\underline{x}}_{\underline{u}(t)}$ is reached for some $\underline{x}_1\in V_0^{q'}$, $\underline{b}_1\in G^\infty$, $\underline{u}_1(\underline{t}_1)\in G$, the same applies necessarily for $\bar{\Phi}^{\underline{x}}_{\underline{u}(t)}$, and then $O(\underline{x}_1)$ has dimension n_0. The maximal dimension of all the orbits $O(\underline{x})$ is exactly n_0. Then from Lemma 1, the Lie algebra generated by F has finite dimension n_0.

Consider the orbit $O(\underline{x}_0)$, of dimension n_0, with $\underline{x}_0\in\Pi_1^{-1}(x_0)\cap O(\underline{x}_0)$, where Π_1 is the projection on the first factor in $V_0^{q'}$. This is possible since the system Q is weakly controllable on V_0. Clearly the system $\tilde{Q}'=(O(\underline{x}_0), \underline{x}_0, F^{q'}=\{X_u^{q'}|u\in U\}, h\circ\Pi_1)$ is a realization of P_Q, and then a local realization of P, generally not observable.

The Lie algebra generated by F has finite dimension n_0. From Lemma 2, there exist a Lie group $\tilde{G}$, simply connected, and an analytic diffeomorphism ν from some connected neighbourhood $\tilde{V}'$ of $\underline{x}_0$ in $O(\underline{x}_0)$ to a neighborhood $\tilde{V}$ of the identity e of $\tilde{G}$, such that:

- $\nu(\underline{x}_0)=e$.
- $d\nu X_u^{q'}=\tilde{X}_u$, $u\in U$, the restriction to $\tilde{V}$ of a right invariant vector field $\tilde{X}_u$ on $\tilde{G}$.

Let $\tilde{h}$ be the analytic mapping from $\tilde{V}$ to Y defined by $\tilde{h}=h\circ\Pi_1\circ\nu^{-1}$. Then $\tilde{Q}=(\tilde{V}, e, \tilde{F}=\{\tilde{X}_u|u\in U\}, \tilde{h})$ is a local realization of P.

(b) *Global realization on a Riemann surface.* Let M be the sheaf of germs of analytic mappings from $\tilde{G}$ to $\tilde{Y}$. Consider $\bar{M}$ the connected component of M containing the germs of $\tilde{h}$ at the points of $\tilde{V}$, and π the canonical projection from $\bar{M}$ to $\tilde{G}$. $\bar{M}$ is an analytic manifold, paracompact if Y is so (when Y is R, $\bar{M}$ is nothing but a Riemann surface without ramification on $\tilde{G}$), and π is a local diffeomorphism from $\bar{M}$ to $\tilde{G}$.

Then the vector fields $\tilde{X}_u$ on $\tilde{G}$ can be lifted in vector fields $\bar{X}_u$ on $\bar{M}$, not complete in general. Let $\bar{x}_0^u\in\bar{M}$ be the germ at e of $\tilde{h}$, and $\bar{h}$ be the analytic mapping from $\bar{M}$ to Y defined by $\bar{h}(\bar{x})=\tilde{f}(\tilde{x})$ where $\bar{x}$ is the germ at $\tilde{x}\in\tilde{G}$ of the mapping $\tilde{f}$. Clearly, the system $\bar{Q}=\{\bar{M}, \bar{x}_0, \bar{F}=\{\bar{X}_u|u\in U\}, \bar{h}\}$ is a local realization of P_Q.

One has now to show that, with a change of origin, $\bar{Q}$ is also a global realization of P.

Consider the following notation:

$$\tilde{\underline{u}}(\underline{t}) = (\exp \tilde{X}_{u_k} t_k \circ \cdots \circ \exp \tilde{X}_{u_1} t_1(\cdot)), \quad \text{for } k \in N, \quad \underline{u} \in N^k, \quad \underline{t} \in R^k,$$

$$\bar{\underline{u}}(\underline{t}) = (\exp \bar{X}_{u_k} t_k \circ \cdots \circ \exp \bar{X}_{u_1} t_1(\cdot)), \quad \text{on its domain.}$$

From now on, the sequence $\underline{u}_0(\underline{t}_0)$ supporting the local realization will be denoted $\underline{u}(\underline{t}_0)$, for simplicity.

Let $\underline{w}(\underline{r}_0)$, $\underline{w} \in U^p$, $p \in N$, be an arbitrary element of D_P, let $\underline{v} \in U^q$, $q \in N$, be such that $\tilde{v}(\underline{s})$ has rank n_0 almost everywhere on R^q, and consider the mapping:

$$\Omega(\underline{s}', \underline{s}, \underline{r}, \underline{t}) = P(\underline{w}(\underline{r})\underline{v}^{-1}(\underline{s}')\underline{v}(s)\underline{u}(\underline{t})).$$

Since P is defined on a truncation-closed open subset D_P of S, Ω is defined on an open connected subset W^+ of $R^{-q} \times R^{+q} \times R^{+p} \times R^{+k}$ which contains the points $(0, 0, 0, 0)$, $(0, 0, 0, \underline{t}_0)$ and $(0, 0, \underline{r}_0, 0)$. Then it can be analytically extended to some connected open subset W or $R^q \times R^q \times R^p \times R^k$ which contains the same three points. Moreover W can be restricted to be of the form: $W = W_s \times W_s \times W_{tr}$ where $W_s \subset R^q$, $W_{tr} \subset R^p \times R^k$. By abuse it will be said "$\underline{r} \in W_{tr}$" for "$\underline{r}$ is such that $(\underline{r}, \underline{t}) \in W_{tr}$ for some $\underline{t} \in R^k$."

There exists a submanifold W_λ of W_s, of dimension n_0, with the canonical immersion denoted by $\underline{s}$, such that, for some $\underline{s}_0 = \underline{s}(\underline{\lambda}_0)$, $\underline{\lambda}_0 \in W_\lambda$, the mapping:

$$\sigma(\underline{\lambda}) = \tilde{\underline{v}}^{-1}(\underline{s}_0)\tilde{\underline{v}}(\underline{s}(\underline{\lambda}))(e)$$

restricted to W_λ, is a diffeomorphism of W_λ on a neighborhood $\tilde{W}_\lambda$ of e in $\tilde{G}$.

Consider now the mapping:

$$h_{\underline{r},\underline{t}'}(\tilde{x}) = P(\underline{w}(\underline{r})\underline{v}^{-1}(\underline{s}_0)\underline{v}(\underline{s}(\sigma^{-1}(\tilde{\underline{w}}^{-1}(\underline{r})\tilde{x}\tilde{\underline{u}}(\underline{t}_0 - \underline{t}')(e))))\underline{u}(\underline{t}'))$$

$$= \Omega(\underline{s}_0, \underline{s}(\sigma^{-1}(\tilde{\underline{w}}^{-1}(\underline{r})\tilde{x}\tilde{\underline{u}}(\underline{t}_0 - \underline{t}')(e))), \underline{r}, \underline{t}').$$

Since $\sigma^{-1}(\tilde{\underline{w}}^{-1}(\underline{r})\tilde{w}(\underline{r})\tilde{\underline{u}}^{-1}(\underline{t}_0 - t')\tilde{\underline{u}}(\underline{t}_0 - \underline{t}')(e)) = \underline{\lambda}_0$, $h_{\underline{r},\underline{t}'}(\tilde{x})$ is defined along the two paths in $W_{tr} \times \tilde{G}$:

$$\zeta \in R^+, \quad \underline{t}' = \zeta * \underline{t}_0, \quad \underline{r} = 0, \quad \tilde{x} = \tilde{\underline{w}}(\underline{r})\tilde{\underline{u}}^{-1}(\underline{t}_0 - \underline{t}')(e),$$

$$\zeta \in R^+, \quad \underline{t}' = 0, \quad \underline{r} = \zeta * \underline{r}_0, \quad \tilde{x} = \tilde{\underline{w}}(\underline{r})\tilde{\underline{u}}^{-1}(\underline{t}_0 - \underline{t}')(e),$$

linking $(0, \underline{t}_0, e)$ to $(0, 0, \tilde{\underline{u}}^{-1}(\underline{t}_0)(e))$, and $(0, 0, \tilde{\underline{u}}^{-1}(\underline{t}_0)(e))$ to $(\underline{r}_0, 0, \tilde{\underline{w}}(\underline{r}_0)\tilde{u}^{-1}(\underline{t}_0)(e)$ respectively. Then $h_{\underline{r},\underline{t}'}(\tilde{\underline{x}})$ is defined on a connected open subset W_h of $W_{tr} \times \tilde{G}$ which contains the three upper mentioned points. For $\tilde{x}$ near e, $h_{\underline{r},\underline{t}'}$ fits with $\tilde{h}$, by construction and from an argument of analyticity. It follows, by analyticity and connectedness, that it does not depend locally on $\underline{r}$ and $\underline{t}'$, everywhere on W_{tr}. Then the germs of $h_{\underline{r},\underline{t}'}$, at points $\tilde{x}$ where it is defined, are elements of $\bar{M}$.

In particular, at $(\underline{r}_0, 0, \tilde{\underline{w}}(\underline{r}_0)\tilde{\underline{u}}^{-1}(\underline{t}_0)(e))$, this gives:

$$\bar{h}(\bar{\underline{w}}(\underline{r}_0)\bar{\underline{u}}^{-1}(\underline{t}_0)(\bar{x}_0)) = P(\underline{w}(\underline{r}_0)\underline{v}^{-1}(\underline{s}_0)\underline{v}(\underline{s}_0)\underline{u}(0)).$$

The mapping

$$\Lambda(\underline{s}', \underline{s}) = P(\underline{w}(\underline{r})\underline{v}^{-1}(\underline{s}')\underline{v}(\underline{s})\underline{u}(0)), \quad \text{for any } \underline{r} = \zeta * \underline{r}_0, \zeta \in R^+,$$

is defined on $W_s \times W_s$. For the open subset $\underline{s}' < 0$, $\underline{s} > 0$, one has:

$$\Lambda(\underline{s}', \underline{s}) = P(\underline{w}(\underline{r})(v_1, -s_1') \cdots (v_{q-1}, -s_{q-1}')(v_q, s_q - s_q')(v_{q-1}, s_{q-1}) \cdots (v_1, s_1)\underline{u}(0)).$$

Then this relation holds also on $W_s \times W_s$, and in particular for every $\underline{s}', \underline{s}$ such that

$s'_q = s_q$. The same process can be iterated for $q-1, \cdots, 1$. This gives:

$$\bar{h}(\bar{\underline{w}}(\underline{r})\bar{u}^{-1}(\underline{t}_0)(\bar{x}_0)) = P(\underline{w}(\underline{r})\underline{u}(0)) = P(\underline{w}(\underline{r})) \quad \text{for any } \underline{r} = \zeta * \underline{r}_0,\ \zeta \in R^+.$$

Taking the new origin in M the point $\bar{x}'_0 = \bar{u}^{-1}(\underline{t}_0)(\bar{x}_0)$, one obtains finally:

$$\bar{h}(\bar{\underline{w}}(\underline{r})(\bar{x}'_0)) = P(\underline{w}(\underline{r})) \quad \text{for every } \underline{w}(\underline{r}) \in D_P.$$

Then the system

$$\bar{Q} = (\bar{M}, \bar{x}'_0, \bar{F} = \{\bar{X}_u | u \in U\}, \bar{h})$$

is a global realization of P.

(c) *Minimality and uniqueness.* $\bar{Q}$ is an analytic system. Then from Sussmann's theory [4], [5], [6], there exists a (global) minimal realization

$$Q^m = (M^m, x_0^m, F^m = \{X_u^m | u \in U\}, h^m)$$

of $\bar{Q}$, then also of P. The minimal realization of P is not unique. However the realization Q^m which has been built is a maximal one in the sense that, if

$$Q = (M, x_0, F = \{X_u | u \in U\}, h)$$

is another minimal realization of P, then there exists a mapping $\tau: M \to M^m$ such that:

- τ is a diffeomorphism of M on its image in M^m,
- $d\tau(X_u) = X_u^m$,
- $h = h_0^m \tau$.

Consider $x \in M$, and some $\underline{u}(\underline{t})$ such that $x = \hat{\underline{u}}(\underline{t})(x_0)$. To the path $\gamma = (R^+) * \hat{u}(t)(x_0)$ between x_0 and x in M corresponds the uniquely defined path $\tilde{\gamma} = (R^+) * \hat{\underline{u}}(\underline{t})(\tilde{x}_0)$ in $\tilde{G}$, where $\tilde{x}_0$ is the point of $\tilde{G}$ corresponding to $\bar{x}'_0$ of $\bar{M}$, since the vector fields $\tilde{X}_u$ are complete in $\tilde{G}$. Applying (like in the last part of (b)) a strategy of identification of the germs of the analytic mapping along $\tilde{\gamma}$, it is clear that to γ and $\tilde{\gamma}$ corresponds the uniquely defined paths $\bar{\gamma} = (R^+) * \hat{\underline{u}}(\underline{t})(\bar{x}'_0)$ in the Riemann surface $\bar{M}$ and $\gamma^m = (R^+) * \hat{\underline{u}}(\underline{t})(x_0^m)$ in M^m. Let $\tau: M \to M^m$ be defined by: $\tau(\bar{u}(\underline{t})(x_0)) = \hat{\underline{u}}(\underline{t})(x_0^m)$. τ is well defined. Clearly x and $\tau(x)$ are indistinguishable, and since Q^m and Q are two minimal analytic realizations, τ is everywhere on M a local diffeomorphism. Moreover the indistinguishability equivalence relation on M^m restricts to the identity. Then τ is a diffeomorphism of M on its image in M^m, with the required proprieties.

This ends the proof of Theorem 1.

COROLLARY 1. *Complete case. Assume that P, analytic, is defined on $D_P = G$. Then the following two propositions are equivalent*:

(a) *The rank n_0 associated to P is finite.*

(b) *There exists a global realization of P on a Lie group $\tilde{G}$, by a family of right-invariant vector fields on $\tilde{G}$.*

Proof. $a \to b$. *n_0 is finite.* From Theorem 1, there exists a (global) realization $\bar{Q} = (\bar{M}, \bar{x}_0, \bar{F} = \{\bar{X}_u | \in U\}, \bar{h})$ where $\bar{M}$ is a Riemann surface on a Lie group $\tilde{G}$ and the $\bar{X}_u$ are the lifts of right-invariant vector fields on $\tilde{G}$.

Since $D_P = G$, the $\bar{X}_u$ are complete on $\bar{M}$. F is a transitive family of complete symmetry vector fields for the closed equivalence relation $\sim$ defined on $\bar{M}$ by: $\bar{x}_1 \sim \bar{x}_2$ when $\pi(\bar{x}_1) = \pi(\bar{x}_2)$, π being the canonical mapping from $\bar{M}$ to G. Then from Sussmann [4, Thm. 11], π defines a locally trivial fibration on $\bar{M}$, of basis $\tilde{G}$ and fiber $\pi^{-1}(\tilde{x})$. Since π is a local diffeomorphism, the fiber is discrete, and $\bar{M}$ is then a covering space of $\tilde{G}$, which is itself a simply connected Lie group. Then $\bar{M}$ and $\tilde{G}$ are essentially the same object and P realizes on $\tilde{G}$, by the right-invariant vector fields $\tilde{X}_u$.

$b \to a$. *Straightforward.* This ends the proof of Corollary 1.

Remark. When $D_P = G$, Jakubczyk's technique is equivalent to the following manipulation. Consider the subset H_1 of G:

$$H_1 = \{h_1 \in G | P(ch_1) = P(c) \text{ for every } c \in G\}.$$

H_1 is a closed subgroup of G, and Jakubczyk's rank condition expresses that H_1 has finite codimension in G. Then the quotient G/H_1, interpreted as the state space of the minimal realization, is a manifold.

Again with $D_P = G$, consider now the subset H_0 of G:

$$H_0 = \{h_0 \in G | P(ch_0 d) = P(cd) \text{ for every } c, d \in G\}.$$

H_0 is a closed subgroup of G, and moreover it is a *normal* subgroup. The quotient $G_0 = G/H_0$ is then a group. The finiteness of n_0 expresses that H_0 has finite codimension. Then G_0 is a Lie group.

This group G_0 interprets as the group of diffeomorphisms of the minimal realization of P. Moreover, the group $\tilde{G}$ which was exhibited above is nothing but the universal covering of G_0.

5. Examples.

Example 1 (*continued*). The rank r_0 of the system described in Example 1 is not finite, and Theorem 1 does not apply.

Example 2. Consider $M = R^2$ with coordinates ρ, h and the vector fields:

$$X_1 = e^{-\rho} \cosh \frac{\partial}{\partial \rho} - e^{-\rho} \sinh \frac{\partial}{\partial h},$$

$$X_2 = e^{-\rho} \sinh \frac{\partial}{\partial \rho} + e^{-\rho} \cosh \frac{\partial}{\partial h};$$

take $x_0 = (\rho_0, h_0) = (\log 2/2, 5\pi/4)$ and the function $r(\rho, h) = h$. This mapping defines a realization Q of an input-output mapping P (whose domain is an open star-shaped subset of S):

$$Q = (M, x_0, \{X_1, X_2\}, r).$$

It will be shown that this input-output mapping P so defined leads to a maximal use of the main theorem (in the sense that the conflicts appearing in analytic extensions of mappings make necessary the use of sheaves of analytic germs).

Consider R^3 with coordinates (x, y, z) and the regular imbedding ψ of R^2 into R^3 so defined:

$$\psi(\rho, h) = \begin{cases} 1 + e^{\rho} \cosh = \psi_1(\rho, h), \\ 1 + e^{\rho} \sinh = \psi_2(\rho, h), \\ h = \psi_3(\rho, h). \end{cases}$$

$\psi(\rho, h)$ is the helicoïd H represented at Fig. 1. On R^2, (imbedded in R^3 as the subspace $z = 0$), consider the two vector fields $\tilde{X}_1 = \partial/\partial x$, $\tilde{X}_2 = \partial/\partial y$, and the point $\tilde{x}_0 = (0, 0)$. R^2, with its additive structure, is a Lie group, and $\tilde{X}_1$ and $\tilde{X}_2$ are right-invariant vector fields on R^2 (as a Lie group).

Denote π for the canonical map; from R^2 to R^2:

$$\pi(\rho, h) \to \begin{cases} x = \psi_1(\rho, h), \\ y = \psi_2(\rho, h). \end{cases}$$

π is a local diffeomorphism from H to $R^2(z = 0)$.

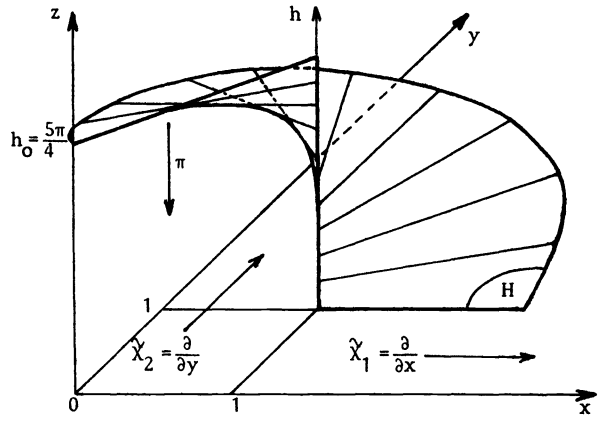


FIG. 1

It is not hard to check that X_1, X_2 are the lifts (through π) of the vector fields $\tilde{X}_1, \tilde{X}_2$ and that the point $x_0 = (\rho_0, h_0)$ belongs to $\pi^{-1}(0, 0)$.

We claim that:

(1) Q is a minimal realization of P.

(2) The Lie algebra generated by X_1 and X_2 (which is the same than the Lie algebra generated by $\tilde{X}_1, \tilde{X}_2$) is isomorphic to the (trivial) Lie algebra of the Lie group R^2 (with its additive structure).

(3) For P, $n_0 = 2$.

(4) $H = \psi(R^2)$ is a Riemann surface on the Lie group $R^2(z = 0)$ associated to the analytic mapping $\tilde{r}$ from $R^2(z = 0)$ to R,

$$\tilde{r}(x, y) = \pi - \arcsin\left(\frac{y-1}{\sqrt{(x+1)^2+(y-1)^2}}\right),$$

$\tilde{r}$ being defined for (x, y) near $(0, 0)$.

(5) π is nothing but the canonical mapping from H (as a Riemann surface on the Lie group R^2) to R^2 (as a Lie group).

(6) The realization

$$\tilde{Q} = (R^2, (0, 0), \{\tilde{X}_1, \tilde{X}_2\}, \tilde{r}(x, y)\},$$

R^2 being considered as a Lie group and $\tilde{X}_1, \tilde{X}_2$ as right-invariant vector fields, realizes locally P for small times.

There is then a maximal use of the theorem, since π is a diffeomorphism of the realizations $\tilde{Q}$, Q for small times, and is not global.

Associate to X_1 and $\tilde{X}_1$ the value u_1 for the control, and to X_2 and $\tilde{X}_2$ the value u_2, and consider the sequences of S:

$$\underline{u}(\underline{t}_0) = (u_2, 2)(u_1, 2) \quad \text{and} \quad \underline{v}(\underline{t}_0) = (u_1, 2)(u_2, 2).$$

Starting from $(x = 0, y = 0)$ with vector fields $\tilde{X}_1, \tilde{X}_2$, wth controls $\underline{u}(\underline{t}_0)$ and $\underline{v}(\underline{t}_0)$ clearly leads to the same point $(x, y) = (2, 2)$,

However, to ensure analyticity, $\tilde{r}$ cannot have the same value for these two sequences of control (along one path, $\tilde{r}$ is increasing, along other one it decreases). On the contrary, starting from x_0 in H with the vector fields X_1, X_2, (that is to say "lifting" the preceding manipulation on H) anything is all right.

REFERENCES

[1] J. P. GAUTHIER AND G. BORNARD, *Uniqueness of weakly minimal analytic rëalizations*, IEEE Trans. Automat. Control, AC-28 (1983), pp. 111-113.

[2] R. HERMANN AND A. J. KRENER, *Nonlinear controllability and observability*, IEEE Trans. Automat. Control, AC-22 (1977), pp. 728-740.

[3] B. JAKUBCZYCK, *Existence and uniqueness of realizations of nonlinear systems*, this Journal, 18 (1980), pp. 455-471.

[4] H. J. SUSSMANN, *A generalization of the closed subgroup theorem to quotients of arbitrary manifolds*, J. Differential Equations, 10 (1975), pp. 151-166.

[5] ———, *Existence and uniqueness of minimal realizations of nonlinear systems*, Math. Systems Theory, 10 (1977), pp. 263-284.

[6] ———, *On quotients of manifolds, a generalization of the closed subgroup theorem*, Bull. Amer. Math. Soc., 80, 3 (1974), pp. 573-575.

[7] ———, *Orbits of families of vector fields and integrability of distributions*, Trans. Amer. Math. Soc., 180 (1973), pp. 171-188.

[8] ———, *Some properties of vector field systems that are not altered by small perturbations*, J. Differential Equations, 20 (1976), pp. 292-315.

[9] NAIMARK STERN, *Théorie des représentations des groupes*, MIR, Moscow, French translation, 1979.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 3, May 1986

012

IDENTIFIABILITY OF SPATIALLY-VARYING CONDUCTIVITY FROM POINT OBSERVATION AS AN INVERSE STURM–LIOUVILLE PROBLEM*

COSTAS KRAVARIS† AND JOHN H. SEINFELD‡

Abstract. This paper discusses identifiability of the spatially varying parameter $\alpha(x)$ in the heat equation $u_t-(\alpha u_x)_x=f$ from measurement of u at a single point. The identifiability problem is formulated as an inverse Sturm–Liouville problem for $(\alpha y')'+\lambda y=0$. It is proved that the eigenvalues and the normalizing constants determine the above Sturm–Liouville operator uniquely. Identifiability and nonidentifiability results are obtained for three heat conduction problems.

Key words. identifiability, distributed parameter systems, inverse Sturm–Liouville problem, system identification

1. Introduction.

The partial differential equation

$$\frac{\partial u}{\partial t}-\frac{\partial}{\partial x}\left(\alpha(x,y)\frac{\partial u}{\partial x}\right)-\frac{\partial}{\partial y}\left(\alpha(x,y)\frac{\partial u}{\partial y}\right)=f(x,y,t) \tag{1.1}$$

governs the temperature distribution in a nonhomogeneous isotropic solid or the pressure distribution in a fluid-containing porous medium. The conductivity $\alpha(x, y)$ is inaccessible to direct measurement and, consequently, its value must be inferred from measurements of u at a finite number of points. A fundamental question arising in such problems is that of *identifiability*, namely, do the measurements provide sufficient information to determine α uniquely.

Relatively little work has been carried out on the identifiability of $\alpha(x)$ in (1.1). Early work by Cannon and coworkers [3–5] is concerned with the steady-state version of (1.1) and identifiability given the temperature u and the heat flux along the boundary. Kitamura and Nakagiri [11] have studied the identifiability of $\alpha(x)$ in the one-dimensional version of (1.1) given measurements of $u(x, t)$ at all x and t. Nakagiri [17] has considered the identifiability of the operator in general first and second order evolution equations in Hilbert spaces given whole domain measurements of the state.

The most relevant measurement configuration is that of one or more point measurements of the state u and we concentrate on that situation here. Specifically, we consider the problem of identifying $\alpha(x)$ in the one-dimensional version of (1.1),

$$\frac{\partial u}{\partial t}-\frac{\partial}{\partial x}\left(\alpha(x)\frac{\partial u}{\partial x}\right)=f(x,t) \tag{1.2}$$

given a measurement of u at a single point x_p, $u(x_p, t)$. The appropriate method of attack to obtain uniqueness and nonuniqueness results is to formulate the problem as an inverse Sturm–Liouville problem.

2. Inverse Sturm–Liouville problems. Their relation to identifiability problems.

The inverse Sturm–Liouville problem was first posed in 1946 by Borg [2] as follows: *Given the eigenvalues λ_n of the Sturm–Liouville problem*

$$y''+[\lambda-q(x)]y=0, \tag{2.1}$$

* Received by the editors June 14, 1983, and in revised form March 15, 1985.

† Department of Chemical Engineering, California Institute of Technology, Pasadena, California 91125. Present address, Department of Chemical Engineering, University of Michigan, Ann Arbor, Michigan 48109.

‡ Department of Chemical Engineering, California Institute of Technology, Pasadena, California 91125.

$$y'(0) - hy(0) = 0, \tag{2.2}$$

$$y'(l) + Hy(l) = 0, \tag{2.3}$$

determine $q(x)$. Borg showed that knowledge of the spectrum alone is not sufficient to determine $q(x)$ uniquely. Since that early work, two not altogether equivalent inverse Sturm-Liouville problems have been considered.

One approach, which has become associated with Gel'fand and Levitan [10], uses the spectral function $\sigma(\lambda)$ as a starting point. If $\phi(x; \lambda)$ denotes the solution of (2.1) satisfying $y(0) = 1$ and $y'(0) = h$ and if we define

$$E_f(\lambda) = \int_0^l f(x)\phi(x; \lambda)\, dx$$

where $f(x)$ is an arbitrary element of $L^2(0, l)$, then by Parseval's theorem,

$$\int_0^l f^2(x)\, dx = \int_{-\infty}^{\infty} E_f^2(\lambda)\, d\sigma(\lambda)$$

where

$$\sigma(\lambda) = \sum_{\lambda_n < \lambda} \frac{1}{\xi_n}$$

and

$$\xi_n = \int_0^l \phi^2(x; \lambda_n)\, dx.$$

Gel'fand and Levitan have shown that knowledge of $\sigma(\lambda)$, or equivalently, of the spectrum $\{\lambda_n\}$ and the normalizing constants $\{\xi_n\}$, determine the potential $q(x)$ uniquely. Furthermore, they provided a method of constructing $q(x)$ from $\sigma(\lambda)$, as well as necessary and sufficient conditions for existence. Note that since $\xi_n = [y_n(0)]^{-2}$, where $y_n(x)$ are the normalized eigenfunctions of the Sturm-Liouville problem (2.1)-(2.3), the results of Gel'fand and Levitan can be interpreted as applicable to the problem of constructing a Sturm-Liouville operator of the form (2.1) given $\{\lambda_n\}$ and $\{|y_n(0)|\}$.

The other approach to the inverse Sturm-Liouville problem consists in using two spectra, such as $\{\lambda_n\}$ associated with (2.1)-(2.3) and $\{\mu_n\}$ associated with (2.1) and a different set of boundary conditions, to determine $q(x)$ [15]. Krein [12], [13] provided a method of constructing $q(x)$ from two spectra as well as necessary and sufficient conditions for existence. The issue of existence was investigated further by Levitan [16] who showed how the normalizing constants ξ_n can be evaluated from $\{\lambda_n\}$ and $\{\mu_n\}$.

It is noteworthy that the prior work on inverse Sturm-Liouville theory is based on the Liouville normal form (2.1)-(2.3). As we shall see shortly, the problem of interest in the present work requires us to consider inverse Sturm-Liouville problems that are not in normal form.

Consider for a moment the parabolic system,

$$\begin{aligned}
&\frac{\partial u}{\partial t} - \frac{\partial}{\partial x}\left(\alpha(x)\frac{\partial u}{\partial x}\right) + q(x)u = f(x, t) \quad \text{in }]0, l[\times]0, T], \\
&u(x, 0) = u_0(x) \quad \text{in }]0, l[, \\
&\frac{\partial u}{\partial x}(0, t) - hu(0, t) = g(t) \quad \text{in }]0, T], \\
&\frac{\partial u}{\partial x}(l, t) + Hu(l, t) = G(t) \quad \text{in }]0, T],
\end{aligned} \tag{2.4}$$

where $f(x, t)$, $u_0(x)$, $g(t)$, $G(t)$, h and H are known. Given the point measurement $z_d(t) = u(x_p, t)$, $t \in]0, T]$, at some $x_p \in [0, l]$, the question is can $\alpha(x)$ and $q(x)$ be uniquely determined.

Note that the solution[1] of (2.4) can be expressed in terms of the eigenvalues λ_n and the eigenfunctions $y_n(x)$ of

$$\begin{aligned} &\frac{d}{dx}\left(\alpha(x)\frac{dy}{dx}\right)+[\lambda - q(x)]y = 0, \\ &y'(0) - hy(0) = 0, \\ &y'(l) + Hy(l) = 0, \end{aligned} \tag{2.5}$$

as follows

$$\begin{aligned} u(x, t) = &\sum_{n=1}^{\infty}\left[\int_0^l u_0(x)y_n(x)\,dx\right]y_n(x)\,e^{-\lambda_n t} \\ &+\int_0^t\int_0^l\left\{\sum_{n=1}^{\infty} y_n(x)y_n(x')\,e^{-\lambda_n(t-\tau)}\right\}f(x', \tau)\,dx'\,d\tau \\ &-\int_0^t\left\{\sum_{n=1}^{\infty}\alpha(0)y_n(0)y_n(x)\,e^{-\lambda_n(t-\tau)}\right\}g(\tau)\,d\tau \\ &+\int_0^t\left\{\sum_{n=1}^{\infty}\alpha(l)y_n(l)y_n(x)\,e^{-\lambda_n(t-\tau)}\right\}G(\tau)\,d\tau \end{aligned} \tag{2.6}$$

and thus the measurement

$$\begin{aligned} z_d(t) = u(x_p, t) = &\sum_{n=1}^{\infty} y_n(x_p)\left[\int_0^l u_0(x)y_n(x)\,dx\right]e^{-\lambda_n t} \\ &+\int_0^t\int_0^l\left\{\sum_{n=1}^{\infty} y_n(x_p)y_n(x')\,e^{-\lambda_n(t-\tau)}\right\}f(x', \tau)\,dx'\,d\tau \\ &-\int_0^t\left\{\sum_{n=1}^{\infty}\alpha(0)y_n(0)y_n(x_p)\,e^{-\lambda_n(t-\tau)}\right\}g(\tau)\,d\tau \\ &+\int_0^t\left\{\sum_{n=1}^{\infty}\alpha(l)y_n(l)y_n(x_p)\,e^{-\lambda_n(t-\tau)}\right\}G(\tau)\,d\tau. \end{aligned} \tag{2.7}$$

Kitamura and Nakagiri [11] (see also [8]) considered (2.4) with $\alpha(x)$ and $q(x)$ both being constant. Using (2.7), they have shown that (under certain assumptions) in the following special cases

(i) $f(x, t) = 0$, $g(t) = G(t) = 0$,

(ii) $f(x, t) = 0$, $u_0(x) = 0$, one of $g(t)$ or $G(t)$ vanishes,

(iii) $u_0(x) = 0$, $g(t) = G(t) = 0$, $f(x, t) = f_1(x)f_2(t)$,

the eigenvalues λ_n can be uniquely determined. Thus the constants α and q can be easily obtained.

When α and q are spatially varying, one can still (under certain assumptions) determine the eigenvalues λ_n as well as some information on the eigenfunctions $y_n(x)$ for the cases (i)-(iii) [18]. Thus the identifiability problem reduces to an inverse Sturm-Liouville problem related to (2.5).

[1] With $\alpha(x)$ strictly positive and $\alpha(x)$, $q(x)$, $f(x, t)$, $u_0(x)$, $g(t)$ and $G(t)$ sufficiently regular, there exists a unique strong solution of (2.4). See [14, pp. 320-321] for appropriate Hölder continuity and compatibility conditions.

Pierce [18] has considered the identifiability of $q(x)$ in (2.4) when $\alpha(x) \equiv 1$. In this case the Sturm-Liouville system (2.5) is in normal form (it is identical to (2.1)-(2.3)). He obtained a number of identifiability results in very special cases, as an immediate consequence of the Gel'fand and Levitan [10] and Levinson [15] theories. Some complementary results to [18] were obtained by Suzuki and Murayama [20] and Suzuki [21]; these results do not rely directly on the Gel'fand-Levitan or the Levinson-Krein theories, but rather on the Povzner representation theorem [19] (this is Lemma 2 in [21, p. 302]).

In the present work we are going to consider the case $q(x) = 0$, i.e.

$$
\begin{aligned}
&\frac{\partial u}{\partial t} - \frac{\partial}{\partial x}\left(\alpha(x)\frac{\partial u}{\partial x}\right) = f(x, t) \quad \text{in }]0, l[\times]0, T], \\
&u(x, 0) = u_0(x) \quad \text{in }]0, l[, \\
&\frac{\partial u}{\partial x}(0, t) - hu(0, t) = g(t) \quad \text{in }]0, T], \\
&\frac{\partial u}{\partial x}(l, t) + Hu(l, t) = G(t) \quad \text{in }]0, T],
\end{aligned} \tag{2.8}
$$

the general problem being to determine $\alpha(x)$ knowing $f(x, t)$, $u_0(x)$, $g(t)$, $G(t)$, h and H and given the point measurement

$$
z_d(t) = u(x_p, t), \qquad t \in]0, T] \tag{2.9}
$$

at some $x_p \in [0, l]$. The Sturm-Liouville system associated wth (2.8) is

$$
\begin{aligned}
&\frac{d}{dx}\left(\alpha(x)\frac{dy}{dx}\right) + \lambda y = 0, \\
&y'(0) - hy(0) = 0, \\
&y'(l) + Hh(l) = 0,
\end{aligned} \tag{2.10}
$$

and the eigenfunction expansion of the solution of (2.8) is still given by (2.6).

In § 3 we define three special cases of (2.8) corresponding to models of physical systems and formulate identifiability problems as inverse Sturm-Liouville problems. In § 4 we state and prove the analogue of Gel'fand and Levitan's result for the Sturm-Liouville problem (2.10). In § 5 we obtain uniqueness and nonuniqueness results for the identification problems of § 3.

3. Problem statement. In the previous section we have stated a general identifiability problem associated with the system (2.8) and the measurement (2.9). With the available tools it does not appear to be possible to attack the problem in its full generality; rather, it is necessary to consider special cases. It will, however, be very important to select cases that are physically relevant and of practical significance. Our selection is based on the following considerations:

(i) In practice one generally has point actuators and thus boundary control and/or point control at some interior point(s) exist as opposed to distributed control.

(ii) Before performing a heat conduction experiment, it is natural to assume that the system is at ambient temperature, i.e. u = constant. Thus, the most important special case for $u_0(x)$ is $u_0(x)$ = constant.

In the present work the analysis will be restricted to the SISO case. In other words, we are going to assume either that only one of $g(t)$, $G(t)$ is nonzero and $f(x, t) = 0$

or $g(t)=G(t)=0$ and $f(x,t)=Q(t)\delta(x-x_p)$. Also, we will restrict ourselves to the special case of $u_0(x)=u_0$ (constant).

We consider

$$
\begin{aligned}
&\frac{\partial u}{\partial t}=\frac{\partial}{\partial x}\left(\alpha(x)\frac{\partial u}{\partial x}\right) \quad \text{in }]0,l[\times]0,T],\\
&u(x,0)=u_0 \quad \text{in }]0,l[,\\
&\alpha(0)\frac{\partial u}{\partial x}(0,t)=Q(t) \quad \text{in }]0,T],\\
&\frac{\partial u}{\partial x}(l,t)=0 \quad \text{in }]0,T],
\end{aligned}
\tag{3.1}
$$

where

$$
\begin{aligned}
&\alpha\in C^1([0,l]) \quad \text{and} \quad \exists\alpha_0>0\colon \alpha(x)\geqq\alpha_0 \quad \forall x\in[0,l],\\
&Q\in H^1(0,T) \quad \text{and} \quad \nexists\varepsilon>0\colon Q(t)=0 \quad \forall t\in]0,\varepsilon[,\\
&u_0\in\mathbb{R}.
\end{aligned}
$$

PROBLEM 1. *To a known input $Q(t)$, a known initial state u_0 and a given measurement $z_d(t)=u(0,t)$, $t\in]0,T]$, does there correspond a unique $\alpha(x)$?*

PROBLEM 2. *To a known input $Q(t)$, a known initial state u_0 and a given measurement $z_d(t)=u(l,t)$, $t\in]0,T]$, does there correspond a unique $\alpha(x)$?*

Now consider

$$
\begin{aligned}
&\frac{\partial u}{\partial t}=\frac{\partial}{\partial x}\left(\alpha(x)\frac{\partial u}{\partial x}\right)+Q(t)\delta(x-x_p) \quad \text{in }]0,l[\times]0,T],\\
&u(x,0)=u_0 \quad \text{in }]0,l[,\\
&\frac{\partial u}{\partial x}(0,t)=\frac{\partial u}{\partial x}(l,t)=0 \quad \text{in }]0,T],
\end{aligned}
\tag{3.2}
$$

where

$$
\begin{aligned}
&\alpha\in C^1([0,l]) \quad \text{and} \quad \exists\alpha_0>0\colon \alpha(x)\geqq\alpha_0 \quad \forall x\in[0,l],\\
&Q\in H^1(0,T) \quad \text{and} \quad \nexists\varepsilon>0\colon Q(t)=0 \quad \forall t\in]0,\varepsilon[,\\
&u_0\in\mathbb{R},\\
&x_p\in]0,l[.
\end{aligned}
$$

PROBLEM 3. *To a known input $Q(t)$, a known initial state u_0 and a given measurement $z_d(t)=u(x_p,t)$, $t\in]0,T]$, does there correspond a unique $\alpha(x)$?*

To be able to formulate Problems 1-3 as inverse Sturm-Liouville problems, we will need the following lemmata:

LEMMA 1. *Let $\{\lambda_n\}$ and $\{\hat{\lambda}_n\}$ be strictly increasing sequences tending to infinity and let $\sum_{n=1}^{\infty}c_ne^{-\lambda_nt}$, $\sum_{n=1}^{\infty}\hat{c}_ne^{-\hat{\lambda}_nt}$ be uniformly convergent on $[\delta,+\infty)$ for every $\delta>0$. Suppose*

$$
\sum_{n=1}^{\infty}c_n\,e^{-\lambda_nt}=\sum_{n=1}^{\infty}\hat{c}_n\,e^{-\hat{\lambda}_nt} \quad \forall t\in]0,T].
\tag{3.3}
$$

If $c_n\neq0$ and $\hat{c}_n\neq0$ $\forall n\in\mathbb{N}$, then $\lambda_n=\hat{\lambda}_n$ and $c_n=\hat{c}_n$ $\forall n\in\mathbb{N}$.

Proof. By analytic continuation we see that (3.3) should hold for all $t>0$. The result follows from the uniqueness of the expansion in Dirichlet series (see for example proof of Lemma 3 in [9]).

LEMMA 2 [22, p. 325]. *Let* Ψ, $Q\in L^1(0, T)$ *and assume*

$$\exists \varepsilon>0\colon Q(t)=0 \quad \text{a.e. in } [0, \varepsilon[.$$

If

$$\int_0^t \Psi(t-\tau)Q(\tau)\,d\tau=0 \quad \text{a.e. in }]0, T[,$$

then

$$\Psi(t)=0 \quad \text{a.e. in }]0, T[.$$

COROLLARY. *Let* $Q\in L^1(0, T)$ *and assume*

$$\exists \varepsilon>0\colon Q(t)=0 \quad \text{a.e. in }]0, \varepsilon[.$$

If the integral equation

$$\int_0^t \Psi(t-\tau)Q(\tau)\,d\tau=R(t)$$

admits a solution $\Psi\in L^1(0, T)$ *then* Ψ *is unique.*

Next consider

$$\frac{d}{dx}\left(\alpha(x)\frac{dy}{dx}\right)+\lambda y=0, \quad y'(0)=0, \quad y'(l)=0, \tag{3.4}$$

where $\alpha\in C^1([0, l])$ and $\exists\alpha_0\colon \alpha(x)\geqq\alpha_0\,\forall x\in[0, l]$ and denote by λ_n the eigenvalues of the above Sturm-Liouville problem and by $y_n(x)$ its normalized eigenfunctions.

PROBLEM 1′. *Referring to* (3.4), *is knowledge of* $\{\lambda_n\}_{n=1}^{\infty}$ *and* $\{|y_n(0)|\}_{n=1}^{\infty}$ *sufficient to determine* $\alpha(x)$ *uniquely*?

Proof of equivalence of Problems 1 *and* 1′. We will show that knowing $Q(t)$ and u_0,

(i) $z_d(t)$ is sufficient to determine $\{\lambda_n\}_{n=1}^{\infty}$ and $\{|y_n(0)|\}_{n=1}^{\infty}$ uniquely and

(ii) $\{\lambda_n\}_{n=1}^{\infty}$ and $\{|y_n(0)|\}_{n=1}^{\infty}$ are sufficient to determine $z_d(t)$ uniquely.

The eigenfunction expansion of the solution of (3.1) is given by

$$u(x, t)=u_0-\int_0^t\left\{\sum_{n=1}^{\infty} y_n(0)y_n(x)\,e^{-\lambda_n(t-\tau)}\right\}Q(\tau)\,d\tau.$$

Hence

$$z_d(t)=u(0, t)=u_0-\int_0^t\left\{\sum_{n=1}^{\infty} |y_n(0)|^2\,e^{-\lambda_n(t-\tau)}\right\}Q(\tau)\,d\tau. \tag{3.5}$$

Now given $Q(t)$, u_0 and $z_d(t)$, it follows from the Corollary of Lemma 2 that the function

$$\Psi(t)=\sum_{n=1}^{\infty} |y_n(0)|^2\,e^{-\lambda_n t}$$

is uniquely determined. But by Lemma 1 $\{\lambda_n\}$ and $\{|y_n(0)|\}$ are uniquely determined by $\Psi(t)$. This proves (i). Part (ii) is an obvious consequence of (3.5).

PROBLEM 2′. *Referring to* (3.4), *is knowledge of* $\{\lambda_n\}_{n=1}^{\infty}$ *and* $\{y_n(0)y_n(l)\}_{n=1}^{\infty}$ *sufficient to determine* $\alpha(x)$ *uniquely*?

Proof of equivalence of Problems 2 *and* 2′. The eigenfunction expansion of the solution of (3.1) is given by

$$u(x, t) = u_0 - \int_0^t \left\{ \sum_{n=1}^{\infty} y_n(0) y_n(x)\, e^{-\lambda_n (t-\tau)} \right\} Q(\tau)\, d\tau.$$

Hence

$$z_d(t) = u(l, t) = u_0 - \int_0^t \left\{ \sum_{n=1}^{\infty} y_n(0) y_n(l)\, e^{-\lambda_n (t-\tau)} \right\} Q(\tau)\, d\tau. \tag{3.6}$$

Using (3.6) and repeating the same argument as with Problems 1 and 1′, we conclude that Problems 2 and 2′ are equivalent.

PROBLEM 3′. *Referring to* (3.4), *is knowledge of* $\{\lambda_n\}_{n=1}^{\infty}$ *and* $\{|y_n(x_p)|\}_{n=1}^{\infty}$ *sufficient to determine* $\alpha(x)$ *uniquely*?

Problem 3 *is equivalent to Problem* 3′ *provided that* $y_n(x_p) \neq 0\ \forall n \in \mathbb{N}$. Indeed, using the eigenfunction expansion of the solution of (3.2),

$$u(x, t) = u_0 + \int_0^t \left\{ \sum_{n=1}^{\infty} y_n(x_p) y_n(x)\, e^{-\lambda_n (t-\tau)} \right\} Q(\tau)\, d\tau$$

we have

$$z_d(t) = u(x_p, t) = u_0 + \int_0^t \left\{ \sum_{n=1}^{\infty} |y_n(x_p)|^2\, e^{-\lambda_n (t-\tau)} \right\} Q(\tau)\, d\tau. \tag{3.7}$$

Thus, repeating the same argument as with Problems 1 and 1′, we conclude that Problems 3 and 3′ are equivalent, provided that $y_n(x_p) \neq 0\ \forall n \in \mathbb{N}$.

4. An associated inverse Sturm–Liouville problem. As mentioned in § 2, Gel'fand and Levitan ([10]) have solved the inverse Sturm-Liouville problem for a Sturm-Liouville operator in *normal* form. Their result (as applied to a finite interval) can be stated as follows:

THEOREM 1 (Gel'fand and Levitan). *Let* $\{\lambda_n\}$ *and* $\{\xi_n\}$ *be two sequences of positive real numbers obeying the asymptotic formulas*

$$\sqrt{\lambda_n} = \frac{\pi}{l} n + \frac{b_1}{n} + \frac{b_3}{n^3} + O\left(\frac{1}{n^4}\right), \qquad \xi_n = \frac{l}{2} + \frac{a_1}{n^2} + O\left(\frac{1}{n^4}\right)$$

where a_1, b_1, b_3 *are constants. Then there exists a unique differential operator, defined by a differential expression of the form*

$$L(y) = y'' - q(x)y, \qquad 0 \leqq x \leqq l$$

with $q \in C([0, l])$ *and by boundary conditions of the form*

$$y'(0) - hy(0) = 0, \qquad y'(l) + Hy(l) = 0$$

which has $\{\lambda_n\}$ *as eigenvalues and* $\{\xi_n\}$ *as normalizing constants. The function* $q(x)$ *and the number* h *can be computed via*

$$q(x) = \frac{1}{2} \frac{\partial K(x, x)}{\partial x}, \qquad h = K(0, 0)$$

where $K(x, t)$ *is the solution of the linear integral equation*

$$F(x, t) + K(x, t) + \int_0^x K(x, s) F(s, t)\, ds = 0$$

and where

$$F(x,t)=\frac{1}{\xi_0}\cos(\sqrt{\lambda_0}x)\cdot\cos(\sqrt{\lambda_0}t)-\frac{1}{l}$$
$$+\sum_{n=1}^{\infty}\left[\frac{\cos(\sqrt{\lambda_n}x)\cos(\sqrt{\lambda_n}t)}{\xi_n}-\frac{2}{l}\cos\left(\frac{n\pi}{l}x\right)\cos\left(\frac{n\pi}{l}t\right)\right].$$

Note that the above theorem gives at the same time existence, uniqueness and method of construction of the differential operator from its eigenvalues and its normalizing constants. For the purpose of studying identifiability problems, one needs only uniqueness. With this in mind, and the fact that $\xi_n=[y_n(0)]^{-2}$, where $y_n(x)$ are the normalized eigenfunctions of the Sturm-Liouville operator, what we wish to retain is

THEOREM 1′ (Gel'fand and Levitan). *Consider*

$$\begin{aligned} &y''+[\lambda-q(x)]y=0,\\ &y'(0)-hy(0)=0,\\ &y'(l)+Hy(l)=0, \end{aligned}\tag{4.1}$$

where $q\in C([0,l])$ and denote by λ_n its eigenvalues and $y_n(x)$ its normalized eigenfunctions. Also, consider

$$\begin{aligned} &y''+[\lambda-r(x)]y=0,\\ &y'(0)-\hat{h}y(0)=0,\\ &y'(l)+\hat{H}y(l)=0, \end{aligned}\tag{4.2}$$

where $r\in C([0,l])$ and denote by μ_n its eigenvalues and $z_n(x)$ its normalized eigenfunctions. If

$$\lambda_n=\mu_n,\qquad |y_n(0)|=|z_n(0)|\quad \forall n\in\mathbb{N},\tag{4.3}$$

then

$$q(x)=r(x),\quad h=\hat{h},\quad H=\hat{H}.$$

The purpose of this section will be to obtain a similar result for the Sturm-Liouville problem (3.4). Note that (3.4) can be reduced to (4.1) via the so-called Liouville transform [1]. Therefore, it is natural to try to "back Liouville transform" the result of Gel'fand and Levitan. In fact, this is possible and it leads to the following result:

Given $\alpha(0)$, λ_n and $|y_n(0)|$ for a differential operator of the form (3.4), *there corresponds a unique $\alpha(x)$.*

Using an entirely different approach than that of Gel'fand and Levitan, we will show that λ_n and $|y_n(0)|$ are sufficient to determine $\alpha(x)$ uniquely. In fact, we will prove the following theorem:

THEOREM 2. *Consider*

$$\begin{aligned} &\frac{d}{dx}\left(\alpha(x)\frac{dy}{dx}\right)+\lambda y=0,\\ &y'(0)-hy(0)=0,\\ &y'(l)+Hy(l)=0, \end{aligned}\tag{4.4}$$

with $\alpha\in C^1([0,l])$, and $\exists\alpha_0>0$: $\alpha(x)\geqq\alpha_0$ $\forall x\in[0,l]$ and denote by λ_n its eigenvalues

and $y_n(x)$ its normalized eigenfunctions. Also, consider

$$\frac{d}{dx}\left(\beta(x)\frac{dy}{dx}\right)+\lambda y=0,$$

$$y'(0)-\hat{h}y(0)=0, \tag{4.5}$$

$$y'(l)+\hat{H}y(l)=0,$$

with $\beta\in C^1([0,l])$, and $\exists\beta_0>0$: $\beta(x)\geqq\beta_0$ $\forall x\in[0,l]$ and denote by μ_n its eigenvalues and $z_n(x)$ its normalized eigenfunctions. If

$$\lambda_n=\mu_n, \qquad |y_n(0)|=|z_n(0)| \quad \forall n\in\mathbb{N}, \tag{4.6}$$

then

$$\alpha(x)=\beta(x), \quad h=\hat{h}, \quad H=\hat{H}.$$

Before proving Theorem 2, we will first prove a number of lemmata. The key to the proof of Theorem 2 will be Lemma 6 which is an analogue of the Povzner representation theorem (see [19] or [21, Lemma 2, p. 302]) for the differential operator $(d/dx)\,(\alpha(x)d/dx)$. An immediate consequence of Lemma 6 will be that we can relate $z_n(x)$ to $y_n(x)$ by

$$\frac{z_n(x)}{z_n(0)}=\frac{U(y_n(x))}{y_n(0)}$$

where U is a bounded linear operator defined by (4.24). Note that U has to relate functions defined on $[0,l]$ to functions defined on $[0,l]$; this is guaranteed by Lemmata 3 and 4. Taking into account that $|z_n(0)|=|y_n(0)|$ we will conclude that U has to be unitary. Lemma 5 will provide the appropriate conditions for U to be unitary. Using these conditions we will finally conclude that $z_n(x)=y_n(x)$.

LEMMA 3. *Consider the class $\mathcal{M}$ of differential operators of the form*

$$A=\frac{d}{dx}\left(\alpha(x)\frac{d}{dx}\right)$$

densely defined in $L^2(0,l)$, where $\alpha\in C^1([0,l])$ bounded below by a positive constant. If $A_1\in\mathcal{M}$ and $A_2\in\mathcal{M}$ have the same spectrum, then

$$\int_0^l\frac{dx}{\sqrt{\alpha_1(x)}}=\int_0^l\frac{dx}{\sqrt{\alpha_2(x)}}.$$

Proof. For every $A\in\mathcal{M}$, the eigenvalues λ_n satisfy the following asymptotic formula (see for example [7])

$$\sqrt{\lambda_n}=\frac{n\pi}{\int_0^l(dx/\sqrt{\alpha(x)})}+O(1).$$

Now if the operators $A_1=(d/dx)\,(\alpha_1(x)d/dx)$ and $A_2=(d/dx)\,(\alpha_2(x)d/dx)$ have

the same eigenvalues λ_n, then

$$\sqrt{\lambda_n}\int_0^l \frac{dx}{\sqrt{\alpha_1(x)}} = n\pi + O(1) = \sqrt{\lambda_n}\int_0^l \frac{dx}{\sqrt{\alpha_2(x)}}$$

$$\Rightarrow \sqrt{\lambda_n}\left(\int_0^l \frac{dx}{\sqrt{\alpha_1(x)}} - \int_0^l \frac{dx}{\sqrt{\alpha_2(x)}}\right) = O(1)$$

$$\Rightarrow \int_0^l \frac{dx}{\sqrt{\alpha_1(x)}} - \int_0^l \frac{dx}{\sqrt{\alpha_2(x)}} = O\left(\frac{1}{\sqrt{\lambda_n}}\right) = O\left(\frac{1}{n}\right) \quad \forall n \in \mathbb{N}.$$

Hence

$$\int_0^l \frac{dx}{\sqrt{\alpha_1(x)}} = \int_0^l \frac{dx}{\sqrt{\alpha_2(x)}}.$$

Q.E.D.

LEMMA 4. *Let $\alpha, \beta \in C^1([0, l])$ bounded below by positive constants and satisfying*

$$\int_0^l \frac{dx}{\sqrt{\alpha(x)}} = \int_0^l \frac{dx}{\sqrt{\beta(x)}}.$$

Denote $\rho \in C^2([0, l])$ the solution of

$$\frac{d\rho}{dx} = [\beta(x)]^{-1/2}[\alpha(\rho)]^{1/2}, \qquad \rho(0) = 0.$$

Then $\rho(x)$ is a bijection of $[0, l]$ onto itself.

Proof. Since $d\rho/dx > 0$, ρ is strictly increasing. Furthermore, from the definition of ρ, we have

$$\frac{d\rho}{[\alpha(\rho)]^{1/2}} = \frac{dx}{[\beta(x)]^{1/2}}$$

which upon integration gives

$$\int_0^{\rho(l)} \frac{dx}{\sqrt{\alpha(x)}} = \int_0^l \frac{dx}{\sqrt{\beta(x)}}.$$

Hence

$$\int_0^{\rho(l)} \frac{dx}{\sqrt{\alpha(x)}} = \int_0^l \frac{dx}{\sqrt{\alpha(x)}} \Leftrightarrow \int_l^{\rho(l)} \frac{dx}{\sqrt{\alpha(x)}} = 0.$$

But since $\alpha(x)$ is strictly positive, $\int_0^x (dx'/\sqrt{\alpha(x')})$ is a strictly increasing function. Hence $\rho(l) = l$. So ρ is a strictly increasing continuous mapping of $[0, l]$ onto $[0, l]$. Hence ρ is a bijection.

LEMMA 5. *Let $\alpha(x)$, $\beta(x)$, $\rho(x)$ as in Lemma 4 and let $U: L^2(0, l) \to L^2(0, l)$ be defined by*

$$U(f) = \kappa\left[\frac{\alpha(\rho(x))}{\beta(x)}\right]^{1/4} f(\rho(x)) + \int_0^x K(x, t) f(\rho(t))\, dt$$

where $\kappa \in \mathbb{R}^+$ and $K \in C([0, l] \times [0, l])$. U will be unitary iff $\kappa = 1$ and $K = 0$.

Proof. Let

$$T(f)=\left[\frac{\alpha(\rho(x))}{\beta(x)}\right]^{1/4} f(\rho(x)), \qquad V(f)=\int_0^x K(x,t)f(\rho(t))\,dt$$

so that

$$U=\kappa T+V.$$

To prove the "if" part of the lemma, we need to show that T is unitary. A straightforward calculation gives the adjoint of T

$$T^*(f)=\left[\frac{\beta(\rho^{-1}(x))}{\alpha(x)}\right]^{1/4} f(\rho^{-1}(x))$$

and thus $TT^*=T^*T=I$.

To prove the "only if" part of the lemma, we will first show that the spectral radius of V is zero. To see this, observe that

$$|(V^n f)(x)|^2 \leqq \frac{\gamma\Gamma^{(n-1)^2}M^n}{2^{n-1}(n-1)!}\|f\|_{L^2}^2 x^{2n-1} \quad \forall x\in[0,l], \tag{4.7}$$

where

$$\gamma=\sup_{0\leqq x\leqq l}\left[\frac{\beta(x)}{\alpha(\rho(x))}\right]^{1/2},$$

$$\Gamma=\sup_{0\leqq x\leqq l}\left[\frac{\alpha(\rho(x))}{\beta(x)}\right]^{1/2},$$

$$M=\sup_{0\leqq t\leqq x\leqq l}[K(x,t)]^2.$$

(See Appendix A for a proof). Hence

$$\|V^n\|_{L^2}\leqq\frac{\gamma^{1/2}}{\sqrt{n!}}\Gamma^{(n-1)^2/2}\left(\frac{Ml^2}{2}\right)^{n/2}, \qquad r(V)=\lim_{n\to\infty}\|V^n\|^{1/n}=0.$$

Now

$$V=U-\kappa T.$$

Since U and T are unitary, it is easy to show that $(U-\kappa T)$ and $(U-\kappa T)^*$ commute, i.e. $V=U-\kappa T$ is normal. But for every normal operator, the spectral radius equals the norm of the operator. Hence $\|V\|_{L^2}=0\Rightarrow K(x,t)=0$. So

$$U=\kappa T$$

and since they are both unitary and $\kappa>0$, it follows that $\kappa=1$. This completes the proof of the lemma.

LEMMA 6. *Let $\alpha(x)$ and $\beta(x)$ be C^1-functions bounded below by positive constants. Furthermore, let $\phi(x;\lambda)$ be the solution of*

$$\frac{d}{dx}\left(\alpha(x)\frac{dw}{dx}\right)+\lambda w=0, \qquad x\geqq 0,$$

$$w(0)=1, \tag{4.8}$$

$$w'(0)=h,$$

$\psi(x;\lambda)$ be the solution of

$$\frac{d}{dx}\left(\beta(x)\frac{dw}{dx}\right)+\lambda w=0, \qquad x\geqq 0,$$

$$w(0)=1, \tag{4.9}$$

$$w'(0)=\hat{h}.$$

Then there exists a continuous kernel $K(x,t)$ such that

$$\psi(x;\lambda)=\left[\frac{\beta(0)}{\alpha(0)}\right]^{1/4}\left[\frac{\alpha(\rho(x))}{\beta(x)}\right]^{1/4}\phi(\rho(x);\lambda)+\int_0^x K(x,t)\phi(\rho(t);\lambda)\,dt \quad \forall x\geqq 0, \tag{4.10}$$

where $\rho(x)$ is the solution of

$$\frac{d\rho}{dx}=[\beta(x)]^{-1/2}[\alpha(\rho)]^{1/2}, \qquad \rho(0)=0. \tag{4.11}$$

Proof. Let $u(x,y;\lambda)=\phi(x;\lambda)\psi(y;\lambda)$.

It can be easily seen that u satisfies the hyperbolic P.D.E.

$$\frac{\partial}{\partial x}\left(\alpha(x)\frac{\partial u}{\partial x}\right)-\frac{\partial}{\partial y}\left(\beta(y)\frac{\partial u}{\partial y}\right)=0 \tag{4.12}$$

and the initial conditions

$$u|_{y=0}=\phi(x;\lambda) \tag{4.13}$$

$$\left.\frac{\partial u}{\partial y}\right|_{y=0}=\hat{h}\phi(x;\lambda). \tag{4.14}$$

The Cauchy problem (4.12)-(4.14) admits a unique solution which can be computed by Riemann's method (See Appendix B for details). First, we make the change of variable

$$X=\int_0^x \frac{dx'}{\sqrt{\alpha(x')}}=\mathscr{A}(x), \qquad Y=\int_0^y \frac{dy'}{\sqrt{\beta(y')}}=\mathscr{B}(y).$$

Then, applying (B.9) we find

$$\begin{aligned}\tilde{u}(X,Y;\lambda)=\frac{1}{2}\left[\frac{\beta(0)}{\tilde{\beta}(Y)\tilde{\alpha}(X)}\right]^{1/4}&\{[\tilde{\alpha}(X-Y)]^{1/4}\tilde{\phi}(X-Y;\lambda)\\ &+[\tilde{\alpha}(X+Y)]^{1/4}\tilde{\phi}(X+Y;\lambda)\}\\ &+\frac{1}{2}\int_{X-Y}^{X+Y}\tilde{W}(X,Y,t)\tilde{\phi}(t;\lambda)\,dt\end{aligned} \tag{4.15}$$

where $\tilde{W}(X,Y,t)=W_1(X,Y,t)+\hat{h}W_2(X,Y,t)$. Applying (4.15) at $X=0$ we obtain

$$\begin{aligned}\tilde{\psi}(Y;\lambda)=\tilde{u}(0,Y;\lambda)=\frac{1}{2}\left[\frac{\beta(0)}{\alpha(0)}\right]^{1/4}&\frac{[\tilde{\alpha}(-Y)]^{1/4}\tilde{\phi}(-Y;\lambda)+[\tilde{\alpha}(Y)]^{1/4}\tilde{\phi}(Y;\lambda)}{[\tilde{\beta}(Y)]^{1/4}}\\ &+\frac{1}{2}\int_{-Y}^{Y}\tilde{W}(0,Y,t)\tilde{\phi}(t;\lambda)\,dt.\end{aligned} \tag{4.16}$$

Now, if the function $\alpha(x)$ is continued so as to be even, it is easy to see that

$\phi(x;\lambda)=\phi(-x;\lambda)$. Thus we easily deduce that

$$\tilde{\alpha}(Y)=\tilde{\alpha}(-Y), \qquad \tilde{\phi}(Y;\lambda)=\tilde{\phi}(-Y;\lambda). \tag{4.17}$$

Furthermore

$$\begin{aligned}\int_{-Y}^{Y}\tilde{W}(0,Y,t)\tilde{\phi}(t;\lambda)\,dt &= \int_{0}^{Y}\tilde{W}(0,Y,t)\tilde{\phi}(t;\lambda)\,dt+\int_{0}^{Y}\tilde{W}(0,Y,-t)\tilde{\phi}(t;\lambda)\,dt \\ &= \int_{0}^{Y}\{\tilde{W}(0,Y,t)+\tilde{W}(0,Y,-t)\}\tilde{\phi}(t;\lambda)\,dt. \end{aligned} \tag{4.18}$$

Thus setting

$$\tilde{K}(Y,t)=\frac{1}{2}\{\tilde{W}(0,Y,t)+\tilde{W}(0,Y,-t)\}$$

it follows from (4.16), (4.17) and (4.18) that

$$\tilde{\psi}(Y;\lambda)=\left[\frac{\beta(0)}{\alpha(0)}\right]^{1/4}\left[\frac{\tilde{\alpha}(Y)}{\tilde{\beta}(Y)}\right]^{1/4}\tilde{\phi}(Y;\lambda)+\int_{0}^{Y}\tilde{K}(Y;t)\tilde{\phi}(t,\lambda)\,dt. \tag{4.19}$$

It remains to back-transform to the original variables. It is easy to see that

$$\rho=\mathscr{A}^{-1}\circ\mathscr{B}.$$

Also recall (from Appendix B) the notation $\tilde{\alpha}=\alpha\circ\mathscr{A}^{-1}$, $\tilde{\beta}=\beta\circ\mathscr{B}^{-1}$, $\tilde{\phi}=\phi\circ\mathscr{A}^{-1}$ etc. Thus (4.19) becomes

$$\psi(y,\lambda)=\left[\frac{\beta(0)}{\alpha(0)}\right]^{1/4}\left[\frac{\alpha(\rho(y))}{\beta(y)}\right]^{1/4}\phi(\rho(y);\lambda)+\int_{0}^{y}K(y,t)\phi(\rho(t);\lambda)\,dt$$

where

$$K(y,t)=\frac{\tilde{K}(\mathscr{B}(y),\mathscr{B}(t))}{\sqrt{\beta(t)}}. \qquad \text{Q.E.D.}$$

Proof of Theorem 2. Let $\phi(x;\lambda)$ be the solution of

$$\frac{d}{dx}\left(\alpha(x)\frac{dy}{dx}\right)+\lambda y=0, \quad y(0)=1, \quad y'(0)=h, \tag{4.20}$$

and $\psi(x;\lambda)$ the solution of

$$\frac{d}{dx}\left(\beta(x)\frac{dy}{dx}\right)+\lambda y=0, \quad y(0)=1, \quad y'(0)=\hat{h}. \tag{4.21}$$

Clearly,

$$y_n(x)=y_n(0)\phi(x;\lambda_n), \quad z_n(x)=z_n(0)\psi(x;\lambda_n) \quad \forall n\in\mathbb{N}.$$

From Lemma 6 we have

$$\psi(x;\lambda)=\left[\frac{\beta(0)}{\alpha(0)}\right]^{1/4}\left[\frac{\alpha(\rho(x))}{\beta(x)}\right]^{1/4}\phi(\rho(x);\lambda)+\int_{0}^{x}K(x,t)\phi(\rho(t);\lambda)\,dt. \tag{4.22}$$

Hence,

$$\frac{z_n(x)}{z_n(0)}=\frac{[\beta(0)/\alpha(0)]^{1/4}[\alpha(\rho(x))/\beta(x)]^{1/4}y_n(\rho(x))+\int_{0}^{x}K(x,t)y_n(\rho(t))\,dt}{y_n(0)}. \tag{4.23}$$

Now define the operator $U: L^2(0, l) \to L^2(0, l)$ by

$$U(f) = \left[\frac{\beta(0)}{\alpha(0)}\right]^{1/4}\left[\frac{\alpha(\rho(x))}{\beta(x)}\right]^{1/4} f(\rho(x)) + \int_0^x K(x, t) f(\rho(t))\, dt. \tag{4.24}$$

Since every $f \in L^2(0, l)$ can be expanded as

$$f(x) = \sum_{n=1}^{\infty} f_n y_n(x)$$

where

$$f_n = \int_0^l f(x) y_n(x)\, dx,$$

we have

$$\begin{aligned}(Uf)(x) &= \left[\frac{\beta(0)}{\alpha(0)}\right]^{1/4}\left[\frac{\alpha(\rho(x))}{\beta(x)}\right]^{1/4} f(\rho(x)) + \int_0^x K(x, t) f(\rho(t))\, dt \\ &= \left[\frac{\beta(0)}{\alpha(0)}\right]^{1/4}\left[\frac{\alpha(\rho(x))}{\beta(x)}\right]^{1/4} \sum_{n=1}^{\infty} f_n y_n(\rho(x)) + \int_0^x K(x, t) \sum_{n=1}^{\infty} f_n y_n(\rho(t))\, dt \\ &= \sum_{n=1}^{\infty} f_n \left\{\left[\frac{\beta(0)}{\alpha(0)}\right]^{1/4}\left[\frac{\alpha(\rho(x))}{\beta(x)}\right]^{1/4} y_n(\rho(x)) + \int_0^x K(x, t) y_n(\rho(t))\, dt\right\}\end{aligned}$$

and taking into account (4.23),

$$(Uf)(x) = \sum_{n=1}^{\infty} f_n \frac{y_n(0)}{z_n(0)} z_n(x).$$

Thus, from Parseval's theorem we obtain

$$\|Uf\|_{L^2}^2 = \sum_{n=1}^{\infty} f_n^2 \left|\frac{y_n(0)}{z_n(0)}\right|^2$$

and since $|y_n(0)| = |z_n(0)|\ \forall n \in \mathbb{N}$ (by (4.6)),

$$\|Uf\|_{L^2}^2 = \sum_{n=1}^{\infty} f_n^2 = \|f\|_{L^2}^2,$$

which means that U is unitary. But from Lemma 5, this implies $[\beta(0)/\alpha(0)]^{1/4} = 1$ and $K(x, t) = 0$. So

$$\alpha(0) = \beta(0), \tag{4.25}$$

$$\psi(x; \lambda) = \left[\frac{\alpha(\rho(x))}{\beta(x)}\right]^{1/4} \phi(\rho(x); \lambda). \tag{4.26}$$

Now, due to (4.26), we have

$$\begin{aligned}\frac{d}{dx}\left(\beta(x)\frac{d\psi}{dx}\right) + \lambda\psi = [\alpha(\rho(x))]^{1/4}[\beta(x)]^{3/4}\frac{d^2\phi(\rho(x); \lambda)}{dx^2} \\ + \frac{1}{2}\left\{\left[\frac{\beta(x)}{\alpha(\rho(x))}\right]^{3/4}\frac{d}{dx}[\alpha(\rho(x))]\right. \\ \left. + \left[\frac{\alpha(\rho(x))}{\beta(x)}\right]^{1/4}\beta'(x)\right\}\frac{d\phi(\rho(x); \lambda)}{dx}\end{aligned}$$

$$+\left\{\frac{d}{dx}\left(\beta(x)\frac{d}{dx}\left(\left[\frac{\alpha(\rho(x))}{\beta(x)}\right]^{1/4}\right)\right)\right\}\phi(\rho(x);\lambda)$$

$$+\lambda\left[\frac{\alpha(\rho(x))}{\beta(x)}\right]^{1/4}\phi(\rho(x);\lambda)$$

$$=\left[\frac{\alpha(\rho(x))}{\beta(x)}\right]^{1/4}\left\{\frac{d}{d\rho(x)}\left(\alpha(\rho(x))\frac{d\phi(\rho(x);\lambda)}{d\rho(x)}\right)+\lambda\phi(\rho(x);\lambda)\right\}$$

$$+\left\{\frac{d}{dx}\left(\beta(x)\frac{d}{dx}\left(\left[\frac{\alpha(\rho(x))}{\beta(x)}\right]^{1/4}\right)\right)\right\}\phi(\rho(x);\lambda).$$

Since $\phi(x;\lambda)$ and $\psi(x;\lambda)$ are solutions of (4.20) and (4.21) respectively, it follows that

$$\frac{d}{dx}\left(\beta(x)\frac{d}{dx}\left(\left[\frac{\alpha(\rho(x))}{\beta(x)}\right]^{1/4}\right)\right)=0.$$

Integrating and taking into account (4.25) we find

$$\left[\frac{\alpha(\rho(x))}{\beta(x)}\right]^{1/4}=1+c\int_0^x\frac{d\xi}{\beta(\xi)};$$

hence

$$\rho(x)=\int_0^x\left(1+c\int_0^\xi\frac{d\xi'}{\beta(\xi')}\right)^2 d\xi.$$

Finally, since $\rho(l)=l$ and $\int_0^x(d\xi/\beta(\xi))$ is a strictly increasing positive function, it easily follows that $c=0$. So $\rho(x)=x$, hence $\alpha(x)=\beta(x)$, hence $\psi(x;\lambda)=\phi(x;\lambda)$, hence $h=\hat{h}$ and $H=\hat{H}$. This completes the proof.

5. Identifiability and nonidentifiability results. Using Theorem 2 we can now solve completely Problems 1 and 2, which were posed in § 3. Also we can solve a special case of Problem 3, namely that for which $x_p=l/2$.

RESULT 1. *Consider Problem* 1. *To a known input* $Q(t)$, *a known initial state* u_0 *and a given measurement,* $z_d(t)=u(0,t)$, $t\in]0,T]$ *there corresponds a unique* $\alpha(x)$.

Proof. Immediate consequence of Theorem 2.

The next result will establish the fact that Problem 2 has in general a nonunique solution. We first establish the following lemma:

LEMMA 7. *Consider* (3.4) *with eigenvalues* λ_n *and normalized eigenfunctions* $y_n(x)$. *Also consider*

$$\frac{d}{dx}\left(\alpha(l-x)\frac{dy}{dx}\right)+\lambda y=0,\quad y'(0)=0,\quad y'(l)=0, \tag{5.1}$$

with eigenvalues $\bar{\lambda}_n$ *and normalized eigenfunctions* $\bar{y}_n(x)$. *Then*

$$\bar{\lambda}_n=\lambda_n,\qquad \bar{y}_n(x)=y_n(l-x)\quad \forall n\in\mathbb{N}. \tag{5.2}$$

Proof. Under the affine transformation

$$\bar{x}=l-x$$

(5.1) reduces to (3.4). Equation (5.2) follows immediately.

RESULT 2. *If* $u(x, t)$ *is the solution of* (3.1) *and* $\bar{u}(x, t)$ *is the solution of*

$$
\begin{aligned}
&\frac{\partial u}{\partial t}=\frac{\partial}{\partial x}\left(\alpha(l-x)\frac{\partial u}{\partial x}\right) \quad \text{in }]0, l[\times]0, T],\\
&u(x, 0)=u_0 \quad \text{in }]0, l[,\\
&\alpha(0)\frac{\partial u}{\partial x}(0, t)=Q(t) \quad \text{in }]0, T],\\
&\frac{\partial u}{\partial x}(l, t)=0 \quad \text{in }]0, T],
\end{aligned}
\tag{5.3}
$$

then

$$
\bar{u}(l, t)=u(l, t). \tag{5.4}
$$

Proof. We have

$$
u(l, t)=u_0-\int_0^t\left\{\sum_{n=1}^{\infty} y_n(0)y_n(l)\, e^{-\lambda_n(t-\tau)}\right\}Q(\tau)\, d\tau,
$$

$$
\bar{u}(l, t)=u_0-\int_0^t\left\{\sum_{n=1}^{\infty} \bar{y}_n(0)\bar{y}_n(l)\, e^{-\bar{\lambda}_n(t-\tau)}\right\}Q(\tau)\, d\tau.
$$

Using Lemma 7, we immediately conclude that

$$
\lambda_n=\bar{\lambda}_n, \quad y_n(0)y_n(l)=\bar{y}_n(0)\bar{y}_n(l) \quad \forall n\in\mathbb{N}.
$$

Hence the result.

LEMMA 8. *Consider again* (3.4). *If* $\alpha(x)$ *is symmetric, i.e.* $\alpha(x)=\alpha(l-x)$, *then*

$$
y_n(x)=y_n(l-x). \tag{5.5}
$$

Proof. Immediate consequence of Lemma 7.

RESULT 3. *Consider again* (3.1) *and assume that* $\alpha(x)$ *is symmetric. Then to a known* $Q(t)$, *a known* u_0 *and a given measurement*

$$
z_d(t)=u(l, t)
$$

there corresponds a unique $\alpha(x)$.

Proof. We have

$$
u(l, t)=u_0-\int_0^t\left\{\sum_{n=1}^{\infty} y_n(0)y_n(l)\, e^{-\lambda_n(t-\tau)}\right\}Q(\tau)\, d\tau.
$$

From (5.5) we have $y_n(0)=y_n(l)$, hence

$$
z_d(t)=u(l, t)=u_0-\int_0^t\left\{\sum_{n=1}^{\infty} |y_n(0)|^2\, e^{-\lambda_n(t-\tau)}\right\}Q(\tau)\, d\tau.
$$

Thus the problem of identifying $\alpha(x)$ from $Q(t)$, u_0 and $z_d(t)$ reduces to the one of identifying $\alpha(x)$ from $\{\lambda_n\}$ and $\{|y_n(0)|\}$. Hence by Theorem 2 $\alpha(x)$ is unique.

The next two results solve the special case $x_p=l/2$ for Problem 3.

RESULT 4. *If $u(x, t)$ is the solution of* (3.2) *for $x_p = l/2$ and $\bar{u}(x, t)$ is the solution of*

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(\alpha(l-x)\frac{\partial u}{\partial x}\right) + Q(t)\delta\left(x - \frac{l}{2}\right) \quad \text{in }]0, l[\times]0, T],$$

$$u(x, 0) = u_0 \quad \text{in }]0, l[, \tag{5.6}$$

$$\frac{\partial u}{\partial x}(0, t) = \frac{\partial u}{\partial x}(l, t) = 0 \quad \text{in }]0, T],$$

then

$$\bar{u}\left(\frac{l}{2}, t\right) = u\left(\frac{l}{2}, t\right).$$

Proof. We have

$$u\left(\frac{l}{2}, t\right) = u_0 + \int_0^t \left\{\sum_{n=1}^{\infty}\left[y_n\left(\frac{l}{2}\right)\right]^2 e^{-\lambda_n(t-\tau)}\right\} Q(\tau)\, d\tau,$$

$$\bar{u}\left(\frac{l}{2}, t\right) = u_0 + \int_0^t \left\{\sum_{n=1}^{\infty}\left[\bar{y}_n\left(\frac{l}{2}\right)\right]^2 e^{-\lambda_n(t-\tau)}\right\} Q(\tau)\, d\tau.$$

Using Lemma 7, we immediately conclude that

$$y_n\left(\frac{l}{2}\right) = \bar{y}_n\left(\frac{l}{2}\right) \quad \forall n \in \mathbb{N}.$$

Hence the result.

RESULT 5. *Consider again* (3.2) *with $x_p = l/2$ and assume that $\alpha(x)$ is symmetric. Then to a known $Q(t)$, a known u_0 and a given measurement*

$$z_d(t) = u\left(\frac{l}{2}, t\right)$$

there corresponds a unique $\alpha(x)$.

Proof. Since $\alpha(x)$ is symmetric, $y_n(x)$ are symmetric $\forall n \in \mathbb{N}$ by Lemma 8. Hence,

$$y_n'\left(\frac{l}{2}\right) = 0 \quad \forall n \in \mathbb{N}.$$

Now consider the Sturm–Liouville problem

$$\frac{d}{dx}\left(\alpha(x)\frac{dy}{dx}\right) + \lambda y = 0, \quad y'(0) = 0, \quad y'\left(\frac{l}{2}\right) = 0, \tag{5.7}$$

and observe that (3.4) and (5.7) have the same eigenvalues λ_n and the same normalized eigenfunctions $y_n(x)$, $x \in [0, l/2]$. Thus the problem of identifying $\alpha(x)$ from $Q(t)$, u_0 and $z_d(t)$ reduces to the one of identifying $\alpha(x)$ in (5.7) from $\{\lambda_n\}$ and $\{|y_n(l/2)|\}$. Hence by Theorem 2 $\alpha(x)$ is unique.

6. Conclusions and significance. In this paper the problem of identifiability of spatially varying conductivity from point measurement of temperature in the linear, one-dimensional heat equation is addressed. Uniqueness and nonuniqueness results are derived referring to special cases of the above general problem. More specifically, in Problem 1 (system (3.1)) we have shown that in a rod, which is insulated at one

end and heated at the other end (with known heat flux), measurement of the temperature as a function of time at the heated end determines uniquely the conductivity as a function of position. Uniqueness is not obtained, however, when in the above physical system the measurement is placed at the insulated end (Problem 2). In Problem 3 (system (3.2)) we have addressed the case of a rod, which is insulated at both ends, with a known heat source at the point $x = x_p$ and a temperature measurement as a function of time at x_p. We have shown that in the special case where both the heat source and the sensor are placed in the middle of the rod ($x_p = l/2$), there corresponds in general a nonunique conductivity. Only in the highly exceptional case where the conductivity is a symmetric function (with respect to the middle of the rod) Problems 2 and 3 can have a unique solution.

There are still important questions that remain unanswered, such as: (i) Is the system described by (3.1) identifiable if $0 < x_p < l$? In other words, is uniqueness the "rule" or the "exception"? (ii) In the system described by (3.2), is the point $x_p = l/2$ an "exceptional" or a "typical" point? What happens for other x_p's?

The above questions reduce to inverse Sturm-Liouville problems for which, at the moment, results are not available. However, we intuitively expect that uniqueness will be the case for all x_p's, except for a set of measure zero.

The motivation of undertaking this work is to establish identifiability conditions for (1.1). Since, for example, (1.1) governs the pressure distribution in petroleum reservoirs and subsurface aquifers, and since the identification of $\alpha(x, y)$ is a key problem in describing these systems, elucidation of the fundamental question of identifiability will have a significant impact on the estimation of such parameters. It should be pointed out, however, that we are far from being able to make the extension from one to two spatial dimensions. Although the work of Nakagiri [17] has established that the identification of eigenvalues from point observation is possible for n-dimensional parabolic systems, the existing inverse Sturm-Liouville theories are still one-dimensional.

Appendix A. Proof of (4.7). We will first prove (4.7) for $n = 1$. We have

$$|(Vf)(x)|^2 = \left| \int_0^x K(x, t) f(\rho(t))\, dt \right|^2$$

$$= \left| \int_0^x \left[\frac{\beta(t)}{\alpha(\rho(t))} \right]^{1/4} K(x, t) \left[\frac{\alpha(\rho(t))}{\beta(t)} \right]^{1/4} f(\rho(t))\, dt \right|^2$$

$$\leqq \left(\int_0^x \left[\frac{\beta(t)}{\alpha(\rho(t))} \right]^{1/2} [K(x, t)]^2\, dt \right) \left(\int_0^x \left[\frac{\alpha(\rho(t))}{\beta(t)} \right]^{1/2} [f(\rho(t))]^2\, dt \right)$$

$$\text{First integral} \leqq \int_0^x \gamma M\, dt = \gamma M x$$

$$\text{Second integral} = \int_0^{\rho(x)} f^2(\zeta)\, d\zeta \leqq \int_0^l f^2(\zeta)\, d\zeta = \|f\|_{L^2}^2$$

and thus $|(Vf)(x)|^2 \leqq \gamma M \|f\|_{L^2}^2 x$.

Suppose now that (4.7) is true for $n = k$, i.e.

$$|(V^k f)(x)|^2 \leqq \frac{\gamma \Gamma^{(k-1)^2} M^k}{2^{k-1}(k-1)!} \|f\|_{L^2}^2 x^{2k-1}.$$

We have

$$|(V^{k+1}f)(x)|^2 = \left|\int_0^x K(x,t)(V^k f)(\rho(t))\right|^2 \leqq \left(\int_0^x [K(x,t)]^2\,dt\right)\left(\int_0^x [(V^k f)(\rho(t))]^2\,dt\right)$$

$$\text{First integral} \leqq \int_0^x M\,dt = Mx$$

$$\text{Second integral} \leqq \frac{\gamma\Gamma^{(k-1)^2}M^k}{2^{k-1}(k-1)!}\|f\|_{L^2}^2\int_0^x [\rho(t)]^{2k-1}\,dt$$

$$\leqq \frac{\gamma\Gamma^{(k-1)^2}M^k}{2^{k-1}(k-1)!}\|f\|_{L^2}^2\Gamma^{2k-1}\int_0^x t^{2k-1}\,dt$$

$$= \frac{\gamma\Gamma^{k^2}M^k}{2^{k-1}(k-1)!}\|f\|_{L^2}^2\frac{x^{2k}}{2k} = \frac{\gamma\Gamma^{k^2}M^k}{2^k k!}\|f\|_{L^2}^2 x^{2k}.$$

Hence

$$|(V^{k+1}f)(x)|^2 \leqq \frac{\gamma\Gamma^{k^2}M^{k+1}}{2^k k!}\|f\|_{L^2}^2 x^{2k+1}.$$

So (4.7) is also true for $n = k+1$. This completes the proof.

Appendix B. Solution of the Cauchy problem for $(\alpha u_x)_x - (\beta u_y)_y = 0$ by Riemann's method. Consider

$$\frac{\partial}{\partial x}\left(\alpha(x)\frac{\partial u}{\partial x}\right) - \frac{\partial}{\partial y}\left(\beta(y)\frac{\partial u}{\partial y}\right) = 0,$$

(B.1)
$$u|_{y=0} = f(x),$$

$$\left.\frac{\partial u}{\partial y}\right|_{y=0} = g(x).$$

By making the change of variable

(B.2)
$$X = \int_0^x \frac{dx'}{\sqrt{\alpha(x')}} = \mathscr{A}(x), \qquad Y = \int_0^y \frac{dy'}{\sqrt{\beta(y')}} = \mathscr{B}(y),$$

(B.1) becomes

$$\frac{\partial^2\tilde{u}}{\partial X^2} - \frac{\partial^2\tilde{u}}{\partial Y^2} + 2\tilde{a}(X)\frac{\partial\tilde{u}}{\partial X} + 2\tilde{b}(Y)\frac{\partial\tilde{u}}{\partial Y} = 0,$$

(B.3)
$$\tilde{u}|_{Y=0} = \tilde{f}(X),$$

$$\left.\frac{\partial\tilde{u}}{\partial Y}\right|_{Y=0} = \sqrt{\beta(0)}\tilde{g}(X),$$

where

(B.4)
$$a(x) = \frac{d\alpha(x)/dx}{4\sqrt{\alpha(x)}}, \qquad b(y) = -\frac{d\beta(y)/dy}{4\sqrt{\beta(y)}}$$

and where we have used the symbols $\tilde{u}(X, Y)$, $\tilde{a}(X)$, $\tilde{b}(Y)$, $\tilde{f}(X)$, $\tilde{g}(X)$ in place of $u(\mathscr{A}^{-1}(X), \mathscr{B}^{-1}(Y))$, $a(\mathscr{A}^{-1}(X))$, $b(\mathscr{B}^{-1}(Y))$, $f(\mathscr{A}^{-1}(X))$, $g(\mathscr{A}^{-1}(X))$ respectively.

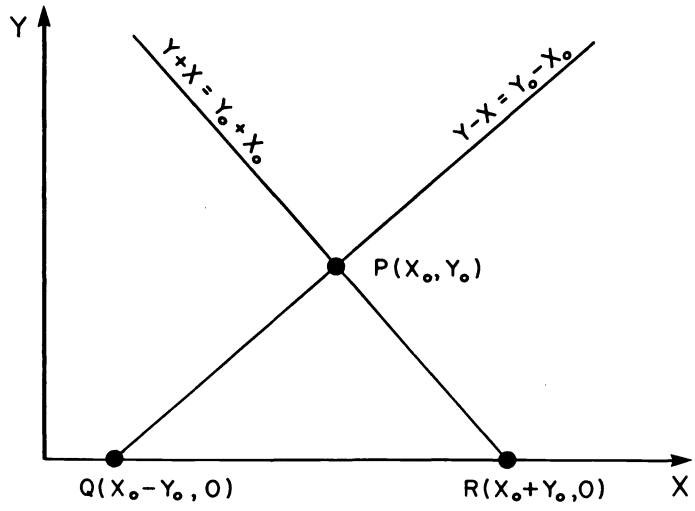


FIG. 1

For the solution of (B.3) we make use of Riemann's formula [6, p. 81] (see Fig. 1),

$$\tilde{u}(P)=\frac{1}{2}[\tilde{u}(Q)\tilde{v}(Q)+\tilde{u}(R)\tilde{v}(R)]+\frac{1}{2}\int_{QR}\left(\tilde{v}\frac{\partial\tilde{u}}{\partial Y}-\tilde{u}\frac{\partial\tilde{v}}{\partial Y}-2\tilde{b}\tilde{u}\tilde{v}\right)dX$$

(B.5)

$$+\left(\tilde{v}\frac{\partial\tilde{u}}{\partial X}-\tilde{u}\frac{\partial\tilde{v}}{\partial X}+2\tilde{a}\tilde{u}\tilde{v}\right)dY$$

where $\tilde{v}(X, Y; X_0, Y_0)$ is the solution of

$$\frac{\partial^2\tilde{v}}{\partial X^2}-\frac{\partial^2\tilde{v}}{\partial Y^2}-2\tilde{a}(X)\frac{\partial\tilde{u}}{\partial X}-2\tilde{b}(Y)\frac{\partial\tilde{v}}{\partial Y}-\left(\frac{d\tilde{a}}{dX}+\frac{d\tilde{b}}{dY}\right)\tilde{v}=0,$$

(B.6)

$$\frac{\partial\tilde{v}}{\partial X}-\frac{\partial\tilde{v}}{\partial Y}=(\tilde{a}+\tilde{b})\tilde{v} \quad \text{on } Y+X=Y_0+X_0,$$

$$\frac{\partial\tilde{v}}{\partial X}+\frac{\partial\tilde{v}}{\partial Y}=(\tilde{a}-\tilde{b})\tilde{v} \quad \text{on } Y-X=Y_0-X_0,$$

$$\tilde{v}(X_0, Y_0)=1.$$

It can be easily seen that the boundary conditions in (B.6) are equivalent to

$$\text{(B.7)}\quad \tilde{v}=\exp\left[\int_{X_0}^{X}\tilde{a}(X')\,dX'-\int_{Y_0}^{Y}\tilde{b}(Y')\,dY'\right] \quad \text{on the lines}\begin{cases} Y+X=Y_0+X_0 \\ Y-X=Y_0-X_0 \end{cases}$$

Thus (B.5) gives

$$\tilde{u}(X_0, Y_0)=\frac{1}{2}\exp\left[\int_0^{Y_0}\tilde{b}(Y')\,dY'\right]$$

$$\cdot\left\{\tilde{f}(X_0-Y_0)\exp\left[\int_{X_0}^{X_0-Y_0}\tilde{a}(X')\,dX'\right]\right.$$

(B.8)

$$\left.+\tilde{f}(X_0+Y_0)\exp\left[\int_{X_0}^{X_0+Y_0}\tilde{a}(X')\,dX'\right]\right\}$$

$$+\frac{1}{2}\int_{X_0-Y_0}^{X_0+Y_0}W_1(X_0, Y_0, t)\tilde{f}(t)\,dt+\frac{1}{2}\int_{X_0-Y_0}^{X_0+Y_0}W_2(X_0, Y_0, t)\tilde{g}(t)\,dt$$

where

$$W_1(X_0, Y_0, t) = -\frac{\partial \tilde{\nu}}{\partial Y}(t, 0; X_0, Y_0) - 2b(0)\tilde{\nu}(t, 0; X_0, Y_0),$$

$$W_2(X_0, Y_0, t) = \sqrt{\beta(0)}\,\tilde{\nu}(t, 0; X_0, Y_0).$$

Finally, taking into account (B.4), the integrals of the first term of (B.8) can be easily evaluated. Thus (B.8) (dropping the subscript 0) gives

$$\tilde{u}(X, Y) = \frac{1}{2}\left[\frac{\beta(0)}{\tilde{\beta}(Y)\tilde{\alpha}(X)}\right]^{1/4}\{[\tilde{\alpha}(X-Y)]^{1/4}\tilde{f}(X-Y) + [\tilde{\alpha}(X+Y)]^{1/4}\tilde{f}(X+Y)\}$$
$$+\frac{1}{2}\int_{X-Y}^{X+Y} W_1(X, Y, t)\tilde{f}(t)\,dt + \frac{1}{2}\int_{X-Y}^{X+Y} W_2(X, Y, t)\tilde{g}(t)\,dt \tag{B.9}$$

where $\tilde{\alpha}(\cdot) = \alpha(\mathscr{A}^{-1}(\cdot))$ and $\tilde{\beta}(\cdot) \equiv \beta(\mathscr{B}^{-1}(\cdot))$.

REFERENCES

[1] G. BIRKHOFF AND G. C. ROTA, *Ordinary Differential Equations*, 3rd. ed., John Wiley, New York, 1968.

[2] G. BORG, *Eine Umkehrung der Sturm-Liouvilleschen Eigenwertaufgabe*, Acta Math., 78 (1946), pp. 1-96.

[3] J. R. CANNON, J. DOUGLAS AND B. F. JONES, *Determination of the diffusivity of an isotropic medium*, Int. J. Engng. Sci., 1 (1963), pp. 453-455.

[4] J. R. CANNON AND B. F. JONES, *Determination of the diffusivity of an anisotropic medium*, Int. J. Engng. Sci., 1 (1963), pp. 457-460.

[5] J. R. CANNON AND P. DUCHATEAU, *Determination of the conductivity of an isotropic medium*, J. Math. Anal. Appl., 48 (1974), pp. 699-704.

[6] E. T. COPSON, *Partial Differential Equations*, Cambridge Univ. Press, Cambridge, 1975.

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

[8] M. COURDESSES, *Comments on identifiability of spatially varying and constant parameters in distributed systems of parabolic type*, this Journal, 21 (1983), pp. 410-412.

[9] M. COURDESSES, M. P. POLIS AND M. AMOUROUX, *On identifiability of parameters in a class of parabolic distributed systems*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 474-477.

[10] I. M. GEL'FAND AND B. M. LEVITAN, *On the determination of a differential equation from its spectral function*, Izv. Akad. Nauk. SSSR Ser. Mat., 15 (1951), pp. 309-360; Amer. Math. Soc. Transl. (Ser. 2), 1 (1955), pp. 253-304.

[11] 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.

[12] M. G. KREIN, *Solution of the inverse Sturm-Liouville problem*, Dokl. Akad. Nauk SSSR (N.S.), 76 (1951), pp. 21-24.

[13] ———, *Determination of the density of a non-homogeneous symmetric cord by its frequency spectrum*, Dokl. Akad. Nauk SSSR (N.S.), 76 (1951), pp. 345-348.

[14] O. A. LADYŽENSKAJA, V. A. SOLONNIKOV AND N. N. URAL'CEVA, *Linear and Quasi-linear Equations of Parabolic Type*, American Mathematical Society, Providence, RI, 1968.

[15] N. LEVINSON, *The inverse Sturm-Liouville problem*, Mat. Tidsskr. B, (1949), pp. 25-30.

[16] B. M. LEVITAN, *On the determination of a differential equation by two of its spectra*, Izv. Akad. Nauk SSSR, 28 (1964), pp. 63-78; Amer. Math. Soc. Transl. (Ser. 2), 68 (1968), pp. 1-20.

[17] S.-I. NAKAGIRI, *Identifiability of Linear Systems in Hilbert Spaces*, this Journal, 21 (1983), pp. 501-530.

[18] A. PIERCE, *Unique identification of eigenvalues and coefficients in a parabolic problem*, this Journal, 17 (1979), pp. 494-499.

[19] A. YA. POVZNER, *On differential equations of Sturm-Liouville type on a half-axis*, Mat. Sb. (N.S.), 23(65) (1948), pp. 3-52; Amer. Math. Soc. Transl. (Ser 1), 4 (1962), pp. 24-101.

[20] T. SUZUKI AND R. MURAYAMA, *A uniqueness theorem in an identification problem for coefficients of parabolic equations*, Proc. Japan Acad. Ser. A. Math. Sci., 56 (1980), pp. 259-263.

[21] T. SUZUKI, *Uniqueness and nonuniqueness in an inverse problem for the parabolic equation*, J. Differential Equations, 47 (1983), pp. 296-316.

[22] E. C. TITCHMARSH, *Introduction to the Theory of Fourier Integrals*, 2nd ed., Clarendon Press, Oxford, 1948.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 3, May 1986

013

HYPERBOLIC STATE SPACE DECOMPOSITION FOR A LINEAR STOCHASTIC DELAY EQUATION*

S. MOHAMMED†, M. SCHEUTZOW‡ AND H. v. WEIZSÄCKER‡

Abstract. We consider the equation $dX(t)=(\int_{-r}^{0}X(s+t)\,d\mu(s))\,dt+G\,dW(t)$ where $r>0$, μ is a matrix-valued signed measure on $[-r,0]$, W is an n-dimensional Wiener process and G is a fixed $n\times n$ matrix. Using the known decomposition $C=S\oplus U$ of $C=C[-r,0]$ into a stable closed subspace S and a finite dimensional unstable subspace U for the analogous deterministic equation, we prove that the projection of the solution X onto S converges exponentially fast to a stationary Gaussian process. Also the speed of explosion of the projection onto U is exhibited.

Key words. asymptotic stationarity, infinite dimensional diffusion, stochastic functional DE, hyperbolic decomposition, stochastic variation of constants

AMS(MOS) subject classifications. Primary 60H10; secondary 34K20, 60G10, 93E03

1. Introduction. Let F be an $\mathbb{R}^n$-valued continuous linear map on the Banach space $C=C([-r,0],\mathbb{R}^n)$ of continuous functions from $[-r,0]$ to $\mathbb{R}^n$ $(r>0)$. For any $t\geqq 0$ and any continuous function $x:[-r,+\infty)\to\mathbb{R}^n$ we define $x_t\in C$ by $x_t(u)=x(t+u)$, $u\in[-r,0]$. Let $(f(t))_{t\geqq 0}$ be an $\mathbb{R}^n$-valued, locally (in t) integrable function.

Deterministic functional differential equations of the form

$$\begin{aligned} &x'(t)=F(x_t)+f(t), \qquad t\geqq 0,\\ &x_0=\eta \end{aligned} \tag{1}$$

with a "hyperbolic" linear functional $F:C\to\mathbb{R}^n$ have frequently been studied. It is known [3, p. 187] that there exists a decomposition $C=U\oplus S$ where U is a finite dimensional linear subspace, such that the projection x_t^U of x_t on U behaves like an unstable finite system of ordinary differential equations and x_t^S converges to 0 as $t\to\infty$ for any initial condition. This decomposition is uniquely determined by the functional F.

In this paper we establish an analogous result if f is replaced by "white noise", i.e., we consider the equation

$$\begin{aligned} &dX(t)=F(X_t)\,dt+G\,dW(t),\\ &X_0=\eta \end{aligned} \tag{2}$$

where $(W(t))_{t\geqq 0}$ is an n-dimensional Wiener process and G is a fixed element of the space $L(\mathbb{R}^n)$ of all $n\times n$ matrices.

Although in general no solution of (2) converges to a stationary process, we show that the projection X_t^S of the solution X_t of (2) on the space S associated with (1) does so. The complementary projection X_t^U is a solution of a purely unstable finite system of stochastic differential equations without delay; in particular on every non-degenerated subspace V of U the variance of the induced probability distribution converges to infinity exponentially.

Our aim was to show how in a specific infinite dimensional situation a stochastic hyperbolic decomposition makes sense. The result could be extended to more general

* Received by the editors May 29, 1984, and in revised form February 15, 1985.

† Department of Mathematics, Southern Illinois University at Carbondale, Carbondale, Illinois 62901.

‡ Fachbereich Mathematik, Universität Kaiserslautern, Postfach 3049, D-6750 Kaiserslautern, West Germany.

situations like time dependent G with appropriate regularity conditions. A more interesting challenge is the formulation of some general principles about infinite dimensional linear stochastic equations from which our results could be derived.

2. The variation of constants. First we want to write down a stochastic variation-of-constants formula of the solution of (2) which looks the same as for equation (1) (cf. [3, pp. 80–87] and Theorems 1 and 2a below). In this section $F: C \to \mathbb{R}^n$ is any continuous linear functional. The assumption of hyperbolicity will appear only in Theorem 4.

LEMMA 1. *Let μ be a finite signed measure on $[a, b]$. Let $\varphi: [a, b] \times [c, d] \to \mathbb{R}$ be in $L_2([a, b] \times [c, d], \lambda \otimes |\mu|)$. Then for every 1-dimensional Brownian motion W*

$$\int_c^d \int_a^b \varphi(s, u)\, dW(s)\, d\mu(u) = \int_a^b \int_c^d \varphi(s, u)\, d\mu(u)\, dW(s) \quad a.s.$$

Proof. If $\varphi = 1_{]\alpha, \beta] \times]\gamma, \delta]}$ then this is the obvious equality

$$\int_\gamma^\delta W(\beta) - W(\alpha)\, d\mu = (W(\beta) - W(\alpha))(\mu]\gamma, \delta]).$$

Both sides of the equation to be proved are linear and are easily seen to be continuous on $L^2([a, b] \times [c, d], \lambda \otimes |\mu|)$. The step functions with rectangular steps being dense in this space, the lemma follows. □

Let $\tilde{C}$ (resp. $\bar{C}$) be the Banach space of all bounded Borel measurable $\mathbb{R}^n$ (resp. $\mathbb{R}^{n \times n}$-) valued functions on $[-r, 0]$ given the sup norm.

For $D \in \bar{C}$ and a linear operator R on $\tilde{C}$ define $RD := (Rd_1, \cdots, Rd_n)$ where d_j is the jth column of D. Define $\bar{\tilde{C}}$ as the set of all functions $B: [a, b] \to \bar{C}$ such that $B(\cdot)(\cdot)$ is a product measurable $\mathbb{R}^{n \times n}$-valued bounded function. For $B \in \bar{\tilde{C}}$ and an n-dimensional Brownian motion we define Wiener integrals of the form $\int_a^b B(s)\, dW(s)$ by

$$\left(\int_a^b B(s)\, dW(s)\right)(u) = \int_a^b B(s)(u)\, dW(s) = \sum_{j=1}^n \int_a^b (B(s)(u))_j\, dW_j(s)$$

$$= \sum_{j=1}^n \int_a^b (b_j(s))(u)\, dW_j(s), \qquad u \in [-r, 0].$$

LEMMA 2. *Let $L: C \to \mathbb{R}^n$ be continuous, linear and $B \in \bar{\tilde{C}}$. Then*

$$\tilde{L}\left(\int_a^b B(s)\, dW(s)\right) = \int_a^b \tilde{L}(B(s))\, dW(s) = \sum_{j=1}^n \int_a^b \tilde{L}(b_j(s))\, dW_j(s) \quad a.e.$$

where $\tilde{L}$ is the canonical extension of L to $\tilde{C}$ using the Riesz representation theorem.

Proof. Follows from Lemma 1. □

For every initial condition $\eta \in \tilde{C}$ there is a unique solution of

$$x'(t) = \tilde{F}(x_t), \qquad x_0 = \eta \tag{3}$$

where $\tilde{F}$ is the canonical extension of F to the space $\tilde{C}$. This can be seen by a straightforward extension of the usual contraction argument. The resulting semigroup $(T(t), t \geqq 0)$ of linear operators on $\tilde{C}$ where $T(t)\eta = x_t$ as in (3), however, is not strongly continuous in contrast to the analogous semigroup on C.

Let $\Delta: [-r, 0] \to L(\mathbb{R}^n)$ be defined by

$$\Delta(u) = 0, \qquad u \in [-r, 0[,$$

$$\Delta(0) = I \qquad (I = \text{identity matrix on } \mathbb{R}^n).$$

THEOREM 1. (variation-of-constant formula). *The solution of* (2) *can be written in the form* $X_t = T(t)\eta + \int_0^t T(t-s)\Delta G\, dW(s)$.

Proof. Define $Y_t = T(t)\eta + \int_0^t T(t-s)\Delta G\, dW(s)$, $t \geqq 0$. For $u \geqq -t$, $u \in [-r, 0]$ we get

$$Y_t(u) = (T(t)\eta)(u) + \int_0^t (T(t-s)\Delta G)(u)\, dW(s)$$

$$= (T(t+u)\eta)(0) + \int_0^{t+u} (T(t+u-s)\Delta G)(0)\, dW(s) = Y_{t+u}(0)$$

as $T(t-s)\Delta G(u) = (T(t+u-s)\Delta G)(0)$ for $0 \leqq s \leqq t+u$ and $(T(t-s)\Delta G)(u) = 0$ for $t+u < s \leqq t$. Also $Y_t(u) = \eta(t+u) = Y_0(t+u)$ if $-r \leqq t+u < 0$. Therefore Y_t corresponds to a stochastic process on $[-r, \infty[$,

$$Y(t) = Y_t(0) = (T(t)\eta)(0) + \int_0^t (T(t-s)\Delta G)(0)\, dW(s)$$

$$= (T(t)\eta)(0) + \sum_{j=1}^n \int_0^t (T(t-s)(\Delta G)_j)(0)\, dW_j(s).$$

Lemma 1 implies the general identity

$$\int_0^t f(t-s)\, dW(s) - \int_0^t f(0)\, dW(s) = \int_0^t \int_s^t \frac{\partial}{\partial u} f(u-s)\, du\, dW(s)$$

$$= \int_0^t \int_0^u \frac{\partial}{\partial u} f(u-s)\, dW(s)\, du.$$

Hence

$$dY(t) = F(T(t)\eta)\, dt + \sum_{j=1}^n \left((\Delta G)_j(0)\, dW_j(t) + \int_0^t \frac{\partial T(t-s)(\Delta G)_j(0)}{\partial t}\, dW_j(s) \right) dt$$

$$= F(T(t)\eta)\, dt + G\, dW(t) + \sum_{j=1}^n \left(\int_0^t F(T(t-s)(\Delta G)_j)\, dW_j(s) \right) dt$$

$$= F\left(T(t)\eta + \int_0^t T(t-s)\Delta G\, dW(s) \right) dt + G\, dW(t) = F(Y_t)\, dt + G\, dW(t)$$

i.e., $Y(t)$ satisfies (2) which proves the assertion as (2) has a unique solution. □

Let A be the unbounded operator on $\tilde{C}$ defined by $D(A) = \{\phi : \phi \in C^1[-r, 0]$ and $\phi'(0) = F(\phi)\}$ and

$$(A\phi)(u) = \begin{cases} \dfrac{d\phi}{du}(u), & -r \leqq u < 0, \\ F(\phi) = \displaystyle\int_{-r}^0 \phi(s)\, d\mu(s), & u = 0, \end{cases}$$

where we have represented F by a matrix-valued signed measure (or function of bounded variation in each component) μ on $[-r, 0]$ (see [3, p. 167]).

We now define U to be the sum of the generalized eigenspaces of the eigenvalues of A with nonnegative real part. The set of eigenvalues of A coincides with the solutions of the characteristic equation $\det(\lambda I - \int_{-r}^0 e^{\lambda u}\, d\mu(u)) = 0$ [3, p. 168]. One can choose a closed complementary subspace S such that $T(t)(S) \subseteq S$ for all $t \geqq 0$. U is finite-dimensional with $T(t)U \subseteq U$ for all t [3, p. 171].

The projection π^U of C onto U corresponding to the decomposition $C = U \oplus S$ is a continuous linear map since S is closed. Thus, U being finite dimensional, it has a representation by a $L(\mathbb{R}^n, U)$-valued, signed measure (or function of bounded variation) ρ^U. This induces a canonical extension of π^U to a continuous linear functional π^U on $\tilde{C}$ by the formula $\pi^U(\phi) = \int_{-r}^{0} \phi \, d\rho^U$, $\phi \in \tilde{C}$. Then $\tilde{C} = U \oplus \tilde{S}$ where $\tilde{S} = \{\phi : \pi^U(\phi) = 0\}$. The associated projection onto $\tilde{S}$ is given by the formula $\pi^{\tilde{S}}(\phi) = \phi - \pi^U(\phi)$, or shortly: $\pi^{\tilde{S}} = 1 - \pi^U$ where 1 is the identity on $\tilde{C}$. We write η^U and $\eta^{\tilde{S}}$ instead of $\pi^U \eta$ and $\pi^{\tilde{S}} \eta$ respectively.

LEMMA 3. *$(T(t)\eta)^U = T(t)(\eta^U)$ and $(T(t)\eta)^{\tilde{S}} = T(t)(\eta^{\tilde{S}})$ holds for every $\eta \in \tilde{C}$.*

Proof. For $\eta \in C$ this follows immediately from the invariance of U and S under $T(t)$; e.g.,

$$(T(t)\eta)^U = (T(t)\eta^U + T(t)\eta^S)^U = (T(t)\eta^U)^U + (T(t)\eta^S)^U = T(t)\eta^U + 0.$$

Let us consider the following continuity property of linear operators $R : \tilde{C} \to \tilde{C}$:

(**) If η^k is a uniformly bounded sequence in $\tilde{C}$ such that $\eta^k(u) \to 0$ for all $u \in [-r, 0]$, then $R\eta^k(u) \to 0$ for all $u \in [-r, 0]$.

From the uniqueness part in the Riesz representation theorem for every continuous linear $R : C \to U$ there is exactly one extension $R : \tilde{C} \to U$ with (**). Hence for the first part of the lemma it is sufficient to show that $\pi^U \circ T(t)$ and $T(t) \circ \pi^U$ both have (**) (we have seen already that they coincide on C). From the definition of $\pi^U : \tilde{C} \to U$, π^U has (**). Thus it remains to prove (**) for $T(t)$. This follows from Lemma 4.

The statement concerning S now follows algebraically:

$$(T(t))\eta^{\tilde{S}} = T(t)\eta + (T(t)\eta)^U = T(t)\eta - T(t)\eta^U = T(t)(\eta - \eta^U) = T(t)\eta^{\tilde{S}}. \qquad \square$$

LEMMA 4. *$T(t)$ has the continuity property (**) for every $t > 0$.*

Proof. Let $|\mu|$ be the variation measure associated with the matrix valued measure μ on $[-r, 0]$ defined by the functional F. Fix η. Define ${}^{\eta}g$ by ${}^{\eta}g(t) := \|T(t)\eta\|_{1,|\mu|} + |(T(t)\eta)(0)|$. We have

$$(T(t)\eta)(u) = \begin{cases} \eta(0) + \int_0^{t+u} F(T(s)\eta)\, ds & \text{for } t \geqq -u, \\ \eta(t+u) & \text{for } t < -u. \end{cases}$$

Thus

$$\begin{aligned}
{}^{\eta}g(t) &= \int_{-r}^{0} |(T(t)\eta)(u)| d|\mu|(u) + |(T(t)\eta)(0)| \\
&= \int_{-r}^{\max(-r,-t)} |\eta(t+u)| d\mu|(u) \\
&\quad + \int_{\max(-r,-t)}^{0} \left| \eta(0) + \int_0^{t+u} F(T(s)\eta)\, ds \right| d|\mu|(u) \\
&\quad + \left| \eta(0) + \int_0^{t} F(T(s)\eta)\, ds \right| \\
&\leqq \int_{-r}^{\max(-t,-r)} |\eta(t+u)| d|\mu|(u) + (\|\mu\|_1 + 1)\left(|\eta(0)| + \int_0^t \|T(s)\eta\|_{1,|\mu|}\, ds \right) \\
&\leqq {}^{\eta}h(t) + C \int_0^t {}^{\eta}g(s)\, ds
\end{aligned}$$

where

$$ {}^{\eta}h(t) = \int_{-r}^{\max(-t,-r)} |\eta(t+u)|d|\mu|(u) + (|||\mu|||_1 + 1)|\eta(0)|, $$

$$ C = |||\mu|||_1 + 1 \quad \text{and} \quad |||\mu|||_1 = |\mu|[-r, 0]. $$

Then Gronwall's lemma [2, p. 393] implies

$$ {}^{\eta}g(t) \leqq {}^{\eta}h(t) + C \int_0^t {}^{\eta}h(s)\, e^{C(t-s)}\, ds. $$

Now suppose (η^k) is a uniformly bounded sequence which converges to 0 pointwise. Then the sequence $({}^{\eta^k}h)$ is also uniformly bounded on bounded intervals and converges to 0 for every t by dominated convergence. The Gronwall estimate then implies again by dominated convergence ${}^{\eta^k}g(t) \to 0$ for all t. In particular $|(T(t)\eta^k)(0)| \to 0$ for all t. Because of

$$ (T(t)\eta^k)(u) = \begin{cases} \eta^k(t+u), & -r \leqq t+u \leqq 0, \\ (T(t+u)\eta^k)(0), & t+u > 0, \end{cases} $$

we get finally that $(T(t)\eta_k)(u)$ converges for every u to 0. □

As in [3, p. 173] define $C^* = C([0, r], \mathbb{R}^{n^*})$, where $\mathbb{R}^{n^*}$ is the n-dimensional space of row vectors. For α in C^* and ϕ in C define

$$ (\alpha, \phi) = \alpha(0)\phi(0) - \int_{-r}^{0} \int_0^{\theta} \alpha(\xi - \theta)\phi(\xi)\, d\xi\, d\mu(\theta) \tag{4} $$

where μ is the matrix-valued measure representing the functional F. Let A^* be the (formal) adjoint operator of A defined by

$$ A^*\alpha(s) = \begin{cases} -\dfrac{d\alpha(s)}{ds}, & 0 < s \leqq r, \\ \displaystyle\int_{-r}^{0} \alpha(-\theta)\, d\mu(\theta), & s = 0, \end{cases} $$

where $D(A^*)$ consists of all functions α in C^* which have a continuous derivative such that

$$ -\frac{d\alpha(0)}{ds} = \int_{-r}^{0} \alpha(-\theta)\, d\mu(\theta) \qquad ([3, p. 174]). $$

According to [3, p. 175] the spectra of A and A^* consist only of eigenvalues and coincide (including multiplicity).

Let U^* be the sum of the generalized eigenspaces of A^* corresponding to the eigenvalues with nonnegative real part. Suppose dim $U = d$ (=dim U^*). Take a basis $\Phi = (\Phi_1, \cdots, \Phi_d)$ for U and a basis

$$ \Psi = \begin{pmatrix} \Psi_1 \\ \vdots \\ \Psi_d \end{pmatrix} $$

for U^* satisfying $(\Phi_i, \Psi_j) = \delta_{ij}$ $(i, j = 1, \cdots, d)$ and let $B \in L(\mathbb{R}^d)$ satisfy $A\Phi = \Phi B$ and $A^*\Psi = B\Psi$ where $A\Phi = (A\Phi_1, \cdots, A\Phi_d)$, $A^*\Psi = (A\Psi_1, \cdots, A\Psi_d)^T$ and ΦB and $B\Psi$ are defined like matrix multiplication. According to [3, p. 186] such a B exists and is unique and its eigenvalues coincide with those of A (or A^*) having nonnegative real

part (including multiplicity). For any $\phi \in C$ the projection of ϕ on U can be written explicitly as

$$\phi^U = \Phi(\Psi, \phi) \tag{5}$$

([3, p. 186]). Now the right-hand side of the last equation also make sense for $\phi \in \tilde{C}$ by defining the bilinear form (4) as above. Both sides are continuous with respect to pointwise convergence of uniformly bounded sequences. For the left side this was stated before Lemma 3, and for the right side it follows from dominated convergence. As $\tilde{C}$ is the smallest class of functions containing C which is closed under pointwise limits of uniformly bounded sequences, the identity also holds for $\phi \in \tilde{C}$.

THEOREM 2. (a) *The projection of X_t on U and $\tilde{S}$ can be written as*

$$X_t^U = T(t)\eta^U + \int_0^t T(t-s)\Delta^U G\, dW(s),$$

$$X_t^{\tilde{S}} = T(t)\eta^{\tilde{S}} + \int_0^t T(t-s)\Delta^{\tilde{S}}\, G\, dW(s).$$

Furthermore

$$\Delta^U = \Phi(\Psi(0)), \qquad \Delta^{\tilde{S}} = \Delta - \Delta^U$$

and

$$X_t^U = \Phi(Y(t)), \qquad t \geqq 0$$

where $Y(t) = (\Psi, X_t)$ is a continuous process on $\mathbb{R}^d$ satisfying the equation

$$dY(t) = BY(t)\, dt + \Psi(0) G\, dW(t), \qquad Y(0) = (\Psi, \eta).$$

(b) *There exist numbers $\alpha < 0$ and $M > 0$ such that*

$$\| T(t)\eta^{\tilde{S}} \| \leqq M e^{\alpha t} \| \eta^{\tilde{S}} \|$$

and

$$\| T(t)\Delta^{\tilde{S}} \| \leqq M e^{\alpha t}, \qquad t \geqq 0$$

where

$$\| T(t)\Delta^{\tilde{S}} \| := \left(\sum_{j=1}^{n} \| T(t)\delta_j^{\tilde{S}} \|^2 \right)^{1/2}$$

and $\delta_j^{\tilde{S}}$ is the jth column of $\Delta^{\tilde{S}}$.

Proof. (a) The representations of X_t^U and $X_t^{\tilde{S}}$ follow from Lemma 2, Theorem 1 and Lemma 3. The formula $\Delta^U = \Phi(\Psi(0))$ is an application of (5). The representation $X_t^U = \Phi Y(t)$ follows exactly like the proof in the deterministic case in [3, p. 186] with $f(s)\, ds$ replaced by $G\, dW(s)$.

(b) [3, p. 187]. □

3. Asymptotics for X^U and X^S.

THEOREM 3. *Let $W(t)$ be an n-dimensional Wiener process defined for all $t \in \mathbb{R}$, normalized by $W(0) = 0$. Then for every distribution of $X_0 = \eta$ with $E\|\eta\|^2 < \infty$, the process $X_t^{\tilde{S}}$ converges to the stationary continuous Gaussian process $Z_t = \int_{-\infty}^t T(t-s)\Delta^{\tilde{S}}\, G\, dW(s)$ in L^2 with exponential speed. In particular the distribution of $X_t^{\tilde{S}}$ converges weakly to the Gaussian distribution of the stationary process $(Z_t)_{t \geqq 0}$.*

Proof. As in the proof of Theorem 1 one can easily show that $(Z_t)_{t\geqq 0}$ corresponds to a process on $[-r,\infty[$ having values in $\mathbb{R}^n$ where

$$Z(t)=Z_t(0)=\int_{-\infty}^{t}(T(t-s)\Delta^{\tilde{S}}G)(0)\,dW(s),\qquad t\geqq 0,$$
$$Z(t)=Z_0(t),\qquad -r\leqq t\leqq 0.$$

$Z(t)$ is well defined as

$$\begin{aligned}E\|Z(t)\|_2^2&=E\left\|\int_{-\infty}^{t}(T(t-s)\Delta^{\tilde{S}}G)(0)\,dW(s)\right\|_2^2\\&=E\sum_{i=1}^{n}\left(\sum_{j=1}^{n}\int_{-\infty}^{t}(((T(t-s)\Delta^{\tilde{S}}G)(0))_j)_i\,dW_j(s)\right)^2\\&=\sum_{i=1}^{n}\sum_{j=1}^{n}\int_{-\infty}^{t}((((T(t-s)\Delta^{\tilde{S}}G)(0))_j)_i)^2\,ds\\&\leqq m_1\int_{-\infty}^{t}\|(T(t-s)\Delta^{\tilde{S}}G)(0)\|^2\,ds\\&\leqq m_2\|G\|^2\int_{-\infty}^{t}\|T(t-s)\Delta^{\tilde{S}}\|^2\,ds\\&\leqq m_2\|G\|^2M^2\int_{-\infty}^{t}e^{2\alpha(t-s)}\,ds=\frac{-m_2\|G\|^2M^2}{2\alpha}<\infty.\end{aligned}$$

Obviously $Z(t)$, $t\geqq -r$ is Gaussian and has continuous sample paths (almost surely). $Z(t)$ is stationary since

$$Z(t)=\int_{-\infty}^{t}(T(t-s)\Delta^{\tilde{S}}G)(0)\,dW(s)=\int_{0}^{\infty}(T(s)\Delta^{\tilde{S}}G)(0)\,dW(t-s).$$

Now

$$E\|X_t^{\tilde{S}}-Z_t\|_C^2\leqq 2E\|T(t)\eta^{\tilde{S}}\|_C^2+2E\left\|\int_{-\infty}^{0}T(t-s)\Delta^{\tilde{S}}G\,dW(s)\right\|_C^2$$

and

$$E\|T(t)\eta^{\tilde{S}}\|_C^2\leqq E(Me^{\alpha t}\|\eta^{\tilde{S}}\|_C)^2=M^2e^{2\alpha t}E\|\eta^{\tilde{S}}\|_C^2\to 0\quad\text{as } t\to\infty.$$

Also, for $t>2r$, we have

$$\begin{aligned}&E\left\|\int_{-\infty}^{0}T(t-s)\Delta^{\tilde{S}}G\,dW(s)\right\|_C^2\\&\quad=E\sup_{-r\leqq u\leqq 0}\lim_{v\to-\infty}\left\|\int_{v}^{0}\sum_{j=1}^{n}T(t-s)(\Delta^{\tilde{S}}G)_j(u)\,dW_j(s)\right\|^2\\&\quad=E\sup_{-r\leqq u\leqq 0}\lim_{v\to-\infty}\sum_{i=1}^{n}\left(\sum_{j=1}^{n}\int_{v}^{0}(T(t-s)(\Delta^{\tilde{S}}G)_j(u))_i\,dW_j(s)\right)^2\\&\quad=E\sup_{-r\leqq u\leqq 0}\lim_{v\to-\infty}\sum_{i=1}^{n}\left(\sum_{j=1}^{n}(T(t-s)(\Delta^{\tilde{S}}G)_j(u))_iW_j(s)\Big|_{v}^{0}\right.\\&\qquad\left.-\int_{v}^{0}\frac{\partial(T(t-s)(\Delta^{\tilde{S}}G)_j(u))_i}{\partial s}W_j(s)\,ds\right)^2\end{aligned}$$

$$\leqq E \lim_{v\to-\infty} 2n \sum_{k=1}^{n} (W_k(v))^2 \sup_{-r\leqq u\leqq 0} \sum_{j=1}^{n} \sum_{i=1}^{n} (T(t-v)(\Delta^{\tilde{S}}G)_j(u))_i^2$$

$$+E \lim_{v\to-\infty} 2n \sup_{-r\leqq u\leqq 0} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\int_v^0 F_i(T(t-s+u)(\Delta^{\tilde{S}}G)_j) |W_j(s)|\, ds \right)^2.$$

Now

$$E \lim_{v\to-\infty} 2n \sum_{k=1}^{n} (W_k(v))^2 \sup_{-r\leqq u\leqq 0} \sum_{j=1}^{n} \sum_{i=1}^{n} (T(t-v)(\Delta^{\tilde{S}}G)_j(u))_i^2$$

$$\leqq E \lim_{v\to-\infty} m_3 \sum_{k=1}^{n} (W_k(v))^2 \sup_{-r\leqq u\leqq 0} \|T(t-v)\Delta^{\tilde{S}}G(u)\|^2$$

$$\leqq E \lim_{v\to-\infty} m_3 \sum_{k=1}^{n} (W_k(v))^2 \sup_{-r\leqq u\leqq 0} \|T(t-v)\Delta^{\tilde{S}}(u)\|^2 \|G\|^2$$

$$\leqq E \lim_{v\to-\infty} m_4 \sum_{k=1}^{n} (W_k(v))^2 \|G\|^2 \|T(t-v)\Delta^{\tilde{S}}\|_C^2$$

$$\leqq E \lim_{v\to-\infty} m_4 \sum_{k=1}^{n} (W_k(v))^2 \|G\|^2 M^2 e^{2\alpha(t-v)}$$

where m_3 and m_4 are suitable constants.

The limit is 0 almost surely as

$$\overline{\lim_{v\to-\infty}} \frac{W_k(v)}{\sqrt{2|v|\log\log|v|}} = 1 \quad \text{a.s.}$$

Using the fact that $T(t-s+u)(\Delta^{\tilde{S}}G)_j \in C$ for $t > 2r$, the second limit is estimated as follows:

$$E \lim_{v\to-\infty} 2n \sup_{-r\leqq u\leqq 0} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\int_v^0 |F_i(T(t-s+u)(\Delta^{\tilde{S}}G)_j)\| W_j(s)|\, ds \right)^2$$

$$\leqq E \lim_{v\to-\infty} (2n)^2 \sup_{-r\leqq u\leqq 0} \sum_{j=1}^{n} \left(\int_v^0 \|F(T(t-s+u)(\Delta^{\tilde{S}}G)_j)\|\, |W_j(s)|\, ds \right)^2$$

$$\leqq E \lim_{v\to-\infty} (2n)^2 \|F\|^2 \sup_{-r\leqq u\leqq 0} \sum_{j=1}^{n} \left(\int_v^0 \|T(t-s+u)(\Delta^{\tilde{S}}G)_j\|\, |W_j(s)|\, ds \right)^2$$

$$\leqq E(2n)^2 \|F\|^2 \|G\|^2 m_5 \lim_{v\to-\infty} \sup_{-r\leqq u\leqq 0} \sum_{j=1}^{n} \left(\int_v^0 \|T(t-s+u)\Delta^{\tilde{S}}\|\, |W_j(s)|\, ds \right)^2$$

$$\leqq E(2n)^2 \|F\|^2 \|G\|^2 m_5 e^{2\alpha t} e^{-2\alpha r} \lim_{v\to-\infty} \sum_{j=1}^{n} \left(\int_v^0 M e^{-\alpha s} |W_j(s)|\, ds \right)^2.$$

The limit is finite almost surely because of the law of the iterated logarithm and its expectation is finite:

$$E\left(\int_{-\infty}^0 e^{-\alpha s} |W_j(s)|\, ds \right)^2 = \int_{-\infty}^0 \int_{-\infty}^0 e^{-\alpha s} e^{-\alpha t} E|W_j(s) W_j(t)|\, ds\, dt$$

$$\leqq -\frac{1}{\alpha} \int_{-\infty}^0 e^{-\alpha s}\, ds < \infty. \qquad \square$$

According to Theorem 2, X_t^U is completely characterized by the finite-dimensional process $Y(t)$. In the next theorem we will describe the growth of $Y(t)$ by looking at its projections on one-dimensional subspaces. Let d be the dimension of the state space of $Y(t)$. Define $C := \psi(0) \cdot G$. Applying the well-known formulas for the mean and covariance of the solution of a linear stochastic differential equation (see e.g. [1, p. 143]),

$$EY(t) = e^{Bt} EY(0) \tag{6}$$

and

$$K(t) = e^{Bt}\left(K(0) + \int_0^t (e^{-Bu}C)(e^{-Bu}C)^T\,du\right)(e^{Bt})^T \tag{7}$$

where $K(t)$ is the covariance matrix of the components of Y, we get

THEOREM 4. *Assume all eigenvalues of B have strictly positive real parts.*

a) *If $EY(0) \neq 0$ then $EY(t)$ grows with exponential speed i.e. there exists some $a > 0$ such that for any $\varepsilon > 0$ there exist positive constants D_1 and D_2 such that*

$$D_1 e^{at} \leqq |EY(t)| \leqq D_2 e^{(a+\varepsilon)t} \quad \text{for all } t \geqq 0.$$

b) *If the "variance of $Y(t)$ in direction z", given by $z^T K(t) z / \|z\|^2$ is strictly positive for some $t_0 > 0$, then it is also strictly positive for all $t \geqq t_0$ and it grows with exponential speed.*

Proof. The proofs follow from (6) and (7) and from the fact that $K(0) + \int_0^t (e^{-Bu}C)(e^{-Bu}C)^T\,du$ is nondecreasing in t in the sense of positive definite matrices. □

Remark. If either $K(0)$ is positive definite or (B, C) is completely controllable i.e. rank $(C, BC, B^2C, \cdots, B^{d-1}C) = d$, then $(z^T K(t) z)/\|z\|^2 > 0$ for all $t > 0$ and for all $z \in \mathbb{R}^d \setminus \{0\}$. The second condition is equivalent to the positive definiteness of the integral in (7) for some (and then for all) $t > 0$ (see [4], p. 39).

The assumption on B means that the operator A has no purely imaginary eigenvalues, i.e., that F is "hyperbola".

Final remark. After submission the essential results of this paper have been incorporated into [6]. Related, but somewhat different questions have been treated in [5] and [7].

REFERENCES

[1] L. ARNOLD, *Stochastische Differentialgleichungen*, Oldenbourg Verlag, München, 1973.

[2] I. I. GIKHMAN AND A. V. SKOROKHOD, *Stochastische Differentialgleichungen*, Springer-Verlag, Berlin, 1971 (translation from Russian).

[3] J. HALE, *Theory of Functional Differential Equations*, Springer-Verlag, New York, 1977.

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

[5] V. J. MIZEL AND V. TRUTZER, *Stochastic hereditary equations: existence and asymptotic stability*, J. Integral Equations, 7 (1984), pp. 1-72.

[6] S. MOHAMMED, *Stochastic Functional Differential Equations*, Research Notes in Mathematics 99, Pitman, London, 1984.

[7] M. SCHEUTZOW, *Qualitative Behaviour of stochastic delay equations with a bounded memory*, Stochastics, 12 (1984), pp. 41-80.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 3, May 1986

014

OPTIMAL CONTROL WITH STATE-SPACE CONSTRAINT I*

HALİL METE SONER†

Abstract. We investigate the optimal value of a deterministic control problem with state space constraint. We show that the optimal value function is the only viscosity subsolution, on the open domain, and the viscosity supersolution, on the closed domain, of the corresponding Bellman equation. Finally, the uniform continuity of the optimal value function is obtained under an assumption on the vector field.

Key words. optimal control, viscosity solutions, Hamilton-Jacobi-Bellman equations, state-space constraint

AMS(MOS) subject classifications. 93E20, 35J65, 35K60, 60J60

Introduction. This paper is concerned with the optimal control of deterministic trajectories given a state-space constraint. The dynamics of the controlled process are (0.1) below. More precisely, let the control u be a Borel measurable map from $[0, \infty)$ into a compact, separable, metric space U and $y(x, \cdot, u)$ be the controlled process. The trajectories $y(x, t, u)$ are the solutions of

$$\frac{d}{dt}y(x, t, u) = b(y(x, t, u), u(t)) \tag{0.1}$$

with initial data $y(x, 0, u) = x$. Let θ be an open subset of R^n and $\mathcal{A}_x$ be the set of strategies under which $y(x, t, u)$ lies in $\bar{\theta}$ (bar denotes the closure). In this paper we refer to $\mathcal{A}_x$ as the set of admissible controls. The structure of $\mathcal{A}_x$ constitutes a state-space constraint. We now associate a discounted cost to every admissible control u and x in $\bar{\theta}$. Given these the optimal value function is

$$v(x) = \inf_{u \in \mathcal{A}_x} \int_0^\infty e^{-t} f(y(x, t, u), u(t))\, dt. \tag{0.2}$$

Note that v is not necessarily continuous. This is caused by the complicated structure of the set valued function $x \to \mathcal{A}_x$. However, as will be shown in § 3 the optimal value function is uniformly continuous on $\bar{\theta}$ given that at every point x on $\partial\theta$ (boundary of θ) there is an $\alpha(x)$ in U such that $b(x, \alpha(x)) \cdot \nu(x) \leqq -\beta < 0$. Here $\nu(x)$ is the exterior normal vector.

If v is uniformly continuous one can make use of the notion of weak (or so-called viscosity) solution of Hamilton-Jacobi equations introduced by M. G. Crandall and P.-L. Lions [2]. In [2] they proved the uniqueness of the viscosity solutions of Hamilton-Jacobi equations in a wide-class of cases. In [1], M. G. Crandall, L. C. Evans, P.-L. Lions provide a simpler introduction to the subject. The book by P.-L. Lions [5] and the review paper by M. G. Crandall and P. E. Souganidis [3] provide a view of the scope of the theory and the references to much of the recent literature. Finally, P.-L. Lions in [6] states results related to constrained problems and viscosity solutions. He proves that under an assumption, stronger than the one above, the optimal value function is locally Lipschitz. Then by using the "everywhere characterization"

* Received by the editors November 5, 1984, and in revised form March 19, 1985. This research was supported by the National Science Foundation under grant MCS-8121940.

† Lefschetz Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, Rhode Island, 02912.

of Lipschitz viscosity solutions [7], he obtains an existence and uniqueness result for the Bellman equation $v+H(x, Dv)=0$.

Let $H \in C(\bar{\theta} \times R^n; R)$ be given by

$$H(x, p)=\sup_{\alpha \in U}\{-b(x, \alpha) \cdot p-f(x, \alpha)\}. \tag{0.3}$$

In § 2, the optimal value function is characterized as the only viscosity solution of $v(x)+H(x, Dv(x))=0$ on θ given the appropriate boundary conditions. Note that v is not a priori defined on $\partial\theta$. The only information at the boundary is given by the state-space constraint. To motivate the boundary condition assume that there is a continuous optimal feedback strategy $\alpha^*(x)$ and v is continuously differentiable on $\bar{\theta}$. The constraint imposes the inequality $b(x, \alpha^*(x)) \cdot \nu(x) \leqq 0$ at the boundary of θ. Also, the optimality of α^* yields $H(x, \nabla v(x))=-b(x, \alpha^*(x)) \cdot \nabla v(x)-f(x, \alpha^*(x))$. Given these, one can show

$$H(x, \nabla v(x)) \leqq H(x, \nabla v(x)+\beta \nu(x)) \quad \text{for all } \beta \geqq 0 \text{ and } x \in \partial\theta. \tag{0.4}$$

Moreover if ψ is differentiable and $v-\psi$ has a minimum on $\bar{\theta}$ at $x \in \partial\theta$, then $\nabla\psi(x)=\nabla v(x)+\beta\nu(x)$ for some positive β. In view of (0.4),

$$v(x)+H(x, \nabla v(x)) \leqq v(x)+H(x, \nabla\psi(x)). \tag{0.5}$$

Since v is smooth, it is easy to show that $v(x)+H(x, \nabla v(x))=0$ on $\bar{\theta}$. Hence $v(x)+H(x, \nabla\psi(x)) \geqq 0$ whenever ψ is smooth and $v-\psi$ has a minimum, relative to $\bar{\theta}$, at $x \in \partial\theta$. In fact, it is proved that v is the only solution of the Bellman equation with this property (Theorem 2.2).

One can view the inequality (0.4) as a constraint on the normal derivative of v at the boundary. Suppose the Hamiltonian $H(x, p)$ is differentiable with respect to p. Then (0.4) reads as $H_p(x, \nabla v(x)) \cdot \nu(x) \geqq 0$. This implicitly imposes a constraint on $\nabla v(x)$ at the boundary. We give the following simple example to clarify this point.

Example. Let $\theta=(0, 1)$, $U=[-1, 1]$, $b(x, u)=u$, $f(x, u)=-u$ if $u \in [0, 1]$ and $f(x, u)=0$ otherwise. The corresponding Hamiltonian $H(x, p)$ is given by

$$H(x, p)=\begin{cases} 1-p & \text{if } p \leqq \frac{1}{2}, \\ p & \text{if } p > \frac{1}{2}. \end{cases}$$

At $x=1$ the condition (0.4) implies that $v_x(1) \geqq \frac{1}{2}$ and at $x=0$ we have $v_x(0) \leqq \frac{1}{2}$. For this example $v(x)=\frac{1}{2}e^{x-1}-1$ is the only solution of $v(x)+H(x, v_x(x))=0$ on $x \in (0, 1)$ satisfying the inequalities $v_x(0) \leqq \frac{1}{2}$ and $v_x(1) \geqq \frac{1}{2}$.

1. Statement of the problem. Let θ be an open subset of $\mathbb{R}^n$ with a connected boundary satisfying:

(A1) There are positive constants h, r and an R^n-value bounded, uniformly continuous map η of $\bar{\theta}$ satisfying

$$B(x+t\eta(x), rt) \subset \theta \quad \text{for all } x \in \bar{\theta} \quad \text{and } t \in (0, h].$$

Here $B(x, r)$ denotes the ball with center x and radius r.

Remark. If θ is bounded and $\partial\theta$ is C^1, then it satisfies (A1). Also boundaries with isolated corners may satisfy (A1), for example, $\theta=\{(x, y) \in R^2: x>0, y>0\}$.

We assume the following throughout the paper:

(1.0) The controls take values in a compact metric space U.

For all $x, y \in R^n$, $u \in U$ the functions

$$b: R^n \times U \to R^n,$$

$$f: R^n \times U \to R$$

satisfy

$$\sup_{\alpha \in U} |b(x, \alpha) - b(v, \alpha)| \leqq L(b)|x - y| \quad \text{for all } x, y, \tag{1.1}$$

$$\sup_{a \in U} |b(x, \alpha)| \leqq K(b) \quad \text{for all } x, \tag{1.2}$$

$$\sup_{a \in U} |f(x, \alpha) - f(y, \alpha)| \leqq \omega_f(|x - y|) \quad \text{for all } x, y, \tag{1.3}$$

$$\sup_{\alpha \in U} |f(x, \alpha)| \leqq K(f) \quad \text{for all } x \tag{1.4}$$

where ω_f is a nondecreasing continuous function with $\omega_f(0) = 0$.

Consider $\mathcal{A}$, the set of all measurable maps of $[0, \infty)$ into U. For any $u \in \mathcal{A}$ and $x \in \bar{\theta}$ let $y(x, \cdot, u)$ be the solution of (0.1) with initial data $y(x, 0, u) = x$. The associated discounted cost $J(x, u)$ is

$$J(x, u) = \int_0^\infty e^{-t} f(y(x, t, u), u(t))\, dt. \tag{1.5}$$

We allow only the controls which leave $y(x, \cdot, u)$ in $\bar{\theta}$. To have a feasible problem, we assume that the set of admissible controls is nonempty, i.e.

$$\mathcal{A}_x = \{u \in \mathcal{A}: y(x, t, u) \in \bar{\theta} \text{ for all } t \geqq 0\} \neq \varnothing \quad \text{for all } x \in \bar{\theta}. \tag{A2}$$

Under these assumptions the optimal value function

$$v(x) = \inf_{u \in \mathcal{A}_x} J(x, u), \qquad x \in \bar{\theta} \tag{1.6}$$

is bounded.

2. Hamilton–Jacobi–Bellman equation. We begin by recalling the notion of viscosity solutions [1], [2]. Let K be a subset of R^n. We will use the notations $C^1(K)$ and $BUC(\bar{\theta})$ to mean the set of continuously differentiable functions in a neighborhood of K and the set of bounded uniformly continuous functions on $\bar{\theta}$, respectively.

DEFINITIONS 1.1. Let K be a subset of R^n and $v \in BUC(\bar{K})$.

(i) We say v is a *viscosity subsolution* of $v(x) + H(x, Dv(x)) = 0$ on K if

$$v(x_0) + H(x_0, \nabla\psi(x_0)) \leqq 0$$

whenever $\psi \in C^1(\bar{K})$ and $v - \psi$ has a maximum, relative to K, at $x_0 \in K$.

(ii) We say v is a *viscosity supersolution* of $v(x) + H(x, Dv(x)) = 0$ on K if

$$v(x_0) + H(x_0, \nabla\psi(x_0)) \geqq 0$$

whenever $\psi \in C^1(\bar{K})$ and $v - \psi$ has a minimum, relative to K, at $x_0 \in K$.

If v is both subsolution and supersolution, then v is called a viscosity solution.

Remark. A viscosity solution v satisfies the equation at every point where v is differentiable.

In order to consider the state space constraint problem, we extend the definition as follows.

DEFINITION 2.1. $v \in BUC(\bar{\theta})$ is said to be a *constrained viscosity solution* of $v(x) + H(x, Dv(x)) = 0$ on $\bar{\theta}$ if it is a subsolution on θ and a supersolution on $\bar{\theta}$.

Remark. The fact that v is a supersolution on the closed domain imposes a boundary condition. To demonstrate this, suppose $v \in C^1(\bar{\theta})$ is a constrained viscosity solution; then $v(x)+H(x, \nabla v(x))=0$ for all $x \in \bar{\theta}$. But also $v(x)+H(x, \nabla v(x)+\alpha\nu(x)) \geqq 0$, for all $x \in \partial\theta$ and α positive, because if $v-\psi$ has a minimum at $x_0 \in \partial\theta$, then $\nabla\psi(x_0)=\nabla v(x_0)+\alpha\nu(x_0)$ for some $\alpha \geqq 0$. Hence v satisfies (0.5).

THEOREM 2.1. *Suppose that* (A1), (A2), (1.0)-(1.4) *hold and that the optimal value function v is in $BUC(\bar{\theta})$. Then v is the only constrained viscosity solution of $v(x)+H(x, \nabla(x))=0$ on $\bar{\theta}$.*

Proof. First recall that the optimal value satisfies the dynamic programming principle, i.e.: for any positive T

$$v(x)=\inf_{u \in \mathscr{A}_x}\left\{\int_0^T e^{-t} f(y(x, t, u), u(t))\, dt + e^{-T} v(y(x, T, u))\right\}. \tag{2.1}$$

Let $\psi \in C^1(\bar{\theta})$, $x_0 \in \theta$ and $(v-\psi)(x_0)=\max[(v-\psi)(x): x \in \bar{\theta}]=0$. Then, for any $u \in \mathscr{A}_{x_0}$ and t positive, the dynamic programming relation yields:

$$\psi(x_0) \leqq \int_0^t e^{-s} f(y(x_0, s, u), u(s))\, ds + e^{-t}\psi(y(x_0, t, u))$$

which implies

$$\begin{aligned}\frac{1}{t}\int_0^t [\psi(y(x_0, s, u))-b(y(x_0, s, u), u(s)) \\ \cdot \nabla\psi(y(x_0, s, u))-f(y(x_0, s, u), u(s))]\, e^{-s}\, ds \leqq 0.\end{aligned} \tag{2.2}$$

Use (1.1), (1.3) and the fact $|y(x_0, s, u)-x_0| \leqq K(b)s$ to obtain:

$$\psi(x_0)-\frac{1}{t}\int_0^t b(x_0, u(s))\, ds \cdot \nabla\psi(x_0)-\frac{1}{t}\int_0^t f(x_0, u(s))\, ds \leqq h(t). \tag{2.3}$$

Here $h(t)$ denotes a continuous function of $[0, \infty)$ into R with value zero at the origin. Put $t_0=\operatorname{dist}(x_0, \partial\theta)/K(b)$ and for any $\alpha \in U$ define u as follows:

$$u(t)=\alpha\chi_{[0, t_0)}(t)+\tilde{u}(t-t_0)\chi_{[t_0, \infty)}(t) \tag{2.4}$$

where $\tilde{u}$ is any control in $\mathscr{A}_{y(x_0, t_0, u)}$. Then u is in $\mathscr{A}_{x_0}$. Use u in (2.3) to get:

$$\psi(x_0)-b(x_0, \alpha)\cdot\nabla\psi(x_0)-f(x_0, \alpha) \leqq h(t) \quad \text{for all } \alpha \in U \text{ and } t \leqq t_0. \tag{2.5}$$

Send t to zero to prove v is a subsolution on θ. Now let $\psi \in C^1(\bar{\theta})$ and $(v-\psi)(x_0)=\min_{x \in \bar{\theta}}[(v-\psi)(x)]=0$ for some $x_0 \in \bar{\theta}$. Then we have

$$\psi(x_0)=\inf_{u \in \mathscr{A}_{x_0}}\left[\int_0^T f(y(x_0, t, u), u(t))\, e^{-t}\, dt + e^{-T} v(y(x_0, T, u))\right] \quad \text{for } T \geqq 0. \tag{2.6}$$

Thus there is a sequence $\{u^m\}_{m=1}^{\infty} \subset \mathscr{A}_{x_0}$ such that

$$\psi(x_0)+\frac{1}{m^2} \geqq \int_0^{1/m} e^{-t} f(y(x_0, t, u^m), u^m(t))\, dt + e^{-1/m}\psi\left(y\left(x_0, \frac{1}{m}, u^m\right)\right). \tag{2.7}$$

Use (1.1) and (1.3) and proceed as in (2.3) and (2.4) to obtain

$$\psi(x_0)-m\int_0^{1/m} b(x_0, u^m(t))\, dt \cdot \nabla\psi(x_0)-m\int_0^{1/m} f(x_0, u^m(t))\, dt \geqq -K(m) \tag{2.8}$$

where $K(m)$ denotes a sequence of numbers which converges to zero as m tends to

infinity. Observe that $(b^m, f^m) = (m\int_0^{1/m} b(x_0, u^m(t))\,dt,\ m\int_0^{1/m} f(x_0, u^m(t))\,dt)$ lies in the closed convex hull of $BF(x_0) = \{(b(x_0, \alpha), f(x_0, \alpha)),\ \alpha \in U\}$ which is compact. Thus there is a subsequence denoted by m again and $(b, f) \in \overline{co}\,BF(x_0)$ such that (b^m, f^m) converges to (b, f). Send m to infinity in (2.8) to get

$$\psi(x_0) - b \cdot \nabla\psi(x_0) - f \geqq 0: \tag{2.9}$$

hence

$$\psi(x_0) + \sup\{-b \cdot \nabla\psi(x_0) - f: (b, f) \in \overline{co}\,BF(x_0)\} \geqq 0. \tag{2.10}$$

But $H(x_0, \nabla\psi(x_0)) = \sup\{-b \cdot \nabla\psi(x_0) - f: (b, f) \in \overline{co}\ BF(x_0)\}$. This proves that the optimal value v is a constrained viscosity solution. The uniqueness is an immediate consequence of the following theorem. □

Consider two running costs $\{f_i;\ i = 1, 2\}$ and the corresponding Hamiltonians H_i defined as in (0.3).

THEOREM 2.2. *Suppose v_1 is a viscosity subsolution of $v(x) + H_1(x, Dv(x)) = 0$ on θ and v_2 is a viscosity supersolution of $v(x) + H_2(x, Dv(x)) = 0$ on $\bar{\theta}$. Let* (A1), (1.1), (1.2) *hold and f_i satisfy* (1.3) *and* (1.4) *for $i = 1, 2$. Then*

$$\sup_{x \in \bar{\theta}} [v_1(x) - v_2(x)] \leqq \sup_{\substack{x \in \bar{\theta} \\ \alpha \in U}} [f_1(x, \alpha) - f_2(x, \alpha)]. \tag{2.11}$$

Before we give the proof, we briefly sketch the technique introduced by Crandall, Evans and Lions [1] and point out the modification we need. Let ξ be a smooth bump function ξ, for example, $\xi(r) = 1 - r^2$. Let $m = \max\{\|v_1\|_\infty, \|v_2\|_\infty\}$ and define

$$\Phi(x, y) = v_1(x) - v_2(y) + 3m\xi\left(\frac{x - y}{\varepsilon}\right). \tag{2.12}$$

Suppose Φ attains its maximum at $(x_0, y_0) \in \bar{\theta} \times \bar{\theta}$. It follows that $|x_0 - y_0| \leqq \varepsilon$. Now consider the map $x \to v_1(x) - v_2(y_0) + 3m\xi(x - y_0/\varepsilon)$. It has a maximum at x_0. If $x_0 \in \theta$, the viscosity property yields

$$v_1(x_0) + H_1(x_0, p_\varepsilon) \leqq 0, \tag{2.13}$$

$$p_\varepsilon = \frac{3m}{\varepsilon}\nabla\xi\left(\frac{x_0 - y_0}{\varepsilon}\right). \tag{2.14}$$

Similarly, consider the map $y \to v_1(x_0) - v_2(y) + 3m\xi(x_0 - y/\varepsilon)$. At $y_0 \in \bar{\theta}$ it has a maximum. The viscosity property implies

$$v_2(y_0) + H_2(y_0, p_\varepsilon) \geqq 0. \tag{2.15}$$

Subtract (2.15) from (2.13) and use the fact $|x_0 - y_0| \leqq \varepsilon$ to obtain

$$\begin{aligned} v_1(x_0) - v_2(y_0) &\leqq H_2(y_0, p_\varepsilon) - H_1(x_0, p_\varepsilon) \\ &\leqq 0(\varepsilon) + \sup_{x \in \theta \alpha \in U} [f_1(x, \alpha) - f_2(x, \alpha)]. \end{aligned} \tag{2.16}$$

This will give (2.11) since one can estimate $\sup_{x \in \theta}[v_1(x) - v_2(x)]$ by $v_1(x_0) - v_2(y_0)$. In general, however, x_0 may lie on $\partial\Omega$. To complete the proof of the theorem, we have to modify Φ so that x_0 lies in θ.

Proof of Theorem 2.1. Let η, r be as in (A1), pick $z_\delta \in \bar{\theta}$ and ρ positive such that

$$|\eta(x) - \eta(y)| \leqq \frac{r}{2} \quad \text{for all } x, y \in \bar{\theta} \text{ and } |x - y| < \rho, \tag{2.17}$$

$$v_1(z_\delta) - v_2(z_\delta) \geqq \sup_{x \in \bar{\theta}} [v_1(x) - v_2(x)] - \frac{\delta}{2}. \tag{2.18}$$

Define $\Phi^\varepsilon : \bar{\theta} \times \bar{\theta} \to R$ as follows:

$$\Phi^\varepsilon(x, y) = v_1(x) - v_2(y) - \left| \frac{x-y}{\varepsilon} - \frac{2}{r}\eta(z_\delta) \right|^2 - \left| \frac{y - z_\delta}{\rho} \right|^2. \tag{2.19}$$

Note that $z_\delta + (\varepsilon 2/r)\eta(z_\delta)$ is in θ for small ε. Use these to obtain

$$\begin{aligned} \Phi^\varepsilon\left(z_\delta + \frac{2\varepsilon}{r}\eta(z_\delta), z_\delta \right) &\geqq v_1(z_\delta) - v_2(z_\delta) - \omega_1(c\varepsilon) \\ &\geqq \sup_{x \in \theta} [v_1(x) - v_2(x)] - \frac{\delta}{2} - \omega_1(c\varepsilon) \end{aligned} \tag{2.20}$$

where $\omega_1(r)$ is the modulus of continuity and c is a positive constant. We also have

$$\Phi^\varepsilon(x, y) \leqq v_1(x) - v_2(x) + \omega_1(|x-y|) - \left| \frac{x-y}{\varepsilon} - \frac{2}{r}\eta(z_\delta) \right|^2 - \left| \frac{y - z_\delta}{\rho} \right|^2. \tag{2.21}$$

Suppose $\Phi^\varepsilon(x, y) \geqq \Phi^\varepsilon(z_\delta + (2\varepsilon/r)\eta(z_\delta), z_\delta)$. Use (2.20) and (2.21) to obtain

$$\left| \frac{y - z_\delta}{\rho} \right|^2 + \left| \frac{x-y}{\varepsilon} - \frac{2}{r}\eta(z_\delta) \right|^2 \leqq \omega_1(c\varepsilon) + \frac{\delta}{2} + \omega_1(|x-y|). \tag{2.22}$$

Since ω_1 is bounded it follows that $\Phi^\varepsilon(x, y) \leqq \Phi^\varepsilon(z_\delta + (2\varepsilon/r)\eta(z_\delta), z_\delta)$ for $x, y \notin B(z_\delta, K)$ for sufficiently large K. Hence Φ^ε achieves its maximum, say at x_0, y_0. Also (2.22) yields that there is m positive such that $|x_0 - y_0| \leqq m\varepsilon$. We use this in (2.22) to obtain

$$\left| \frac{y_0 - z_\delta}{\rho} \right|^2 + \left| \frac{x_0 - y_0}{\varepsilon} - \frac{2}{r}\eta(z_\delta) \right|^2 \leqq \omega_1(c\varepsilon) + \frac{\delta}{2} + \omega_1(m\varepsilon). \tag{2.23}$$

Pick ε and δ so that the right-hand side of (2.23) is less than one. Hence $|y_0 - z_\delta| \leqq \rho$, (2.17) implies there in e is the unit ball such that $\eta(y_0) = \eta(z_\delta) + (r/2)e$. Also, there is e' again in the unit ball such that $x_0 = y_0 + (\varepsilon 2/r)\eta(z_\delta) + \varepsilon e'$. Combining these yields

$$x_0 = y_0 + \frac{2\varepsilon}{r}\left(\eta(y_0) + r\left[-\frac{e}{2} + \frac{e'}{2} \right] \right) \in B(y_0 + t\eta(y_0), tr) \tag{2.24}$$

with $t = 2\varepsilon/r$. Thus, (A1) implies $x_0 \in \theta$ if ε is small. Now consider the maps

$$\bar{\psi}(x) = v_2(y_0) + \left| \frac{x - y_0}{\varepsilon} - \frac{2}{r}\eta(z_\delta) \right|^2 + \left| \frac{y_0 - z_\delta}{\rho} \right|^2, \tag{2.25}$$

$$\psi(y) = v_1(x_0) - \left| \frac{x_0 - y}{\varepsilon} - \frac{2}{r}\eta(z_\delta) \right|^2 - \left| \frac{y - z_\delta}{\rho} \right|^2. \tag{2.26}$$

Then $v_1 - \bar{\psi}$ has a maximum at $x_0 \in \theta$ and $v_2 - \psi$ has a minimum at $y_0 \in \bar{\theta}$. The viscosity property yields

$$v_1(x_0) + H_1(x_0, p_\varepsilon) \leqq 0, \tag{2.27}$$

$$v_2(y_0) + H_2(y_0, p_\varepsilon + q_\varepsilon) \geqq 0 \tag{2.28}$$

where

$$p_\varepsilon = 2\left(\frac{x_0 - y_0}{\varepsilon} - \frac{2}{r}\eta(z_\delta)\right)\frac{1}{\varepsilon}, \qquad q_\varepsilon = -2\left(\frac{y_0 - z_\delta}{\rho^2}\right).$$

Subtract (2.28) from (2.27),

$$\begin{aligned} v_1(x_0) - v_2(y_0) &\leqq [H_2(y_0, p_\varepsilon + q_\varepsilon) - H_2(x_0, p_\varepsilon)] \\ &\quad + [H_2(x_0, p_\varepsilon) - H_1(x_0, p_\varepsilon)] := I(\varepsilon) + J(\varepsilon), \end{aligned} \tag{2.29}$$

$$J(\varepsilon) \leqq \sup_{\alpha \in U} [f_1(x_0, \alpha) - f_2(x_0, \alpha)]. \tag{2.30}$$

Using (1.2) and (1.3) yields

$$I(\varepsilon) \leqq \omega_{f_2}(|x_0 - y_0|) + |p_\varepsilon| L(b)|x_0 - y_0| + K(b)|q_\varepsilon|. \tag{2.31}$$

(2.23) yields $|q_\varepsilon| \leqq h(\varepsilon) + \delta/2$, $\varepsilon|p_\varepsilon| \leqq h(\varepsilon) + \delta/2$, and $|x_0 - y_0|\varepsilon^{-1} \leqq C$ for some C independent of ε. Here $h(\varepsilon)$ denotes a continuous function of ε which has value zero at the origin. Thus, we have

$$I(\varepsilon) \leqq \omega_{f_2}(C\varepsilon) + L(b)h(\varepsilon)C + K(b)h(\varepsilon) \leqq h(\varepsilon) + C\delta. \tag{2.32}$$

Substitute (2.30), (2.32) into (2.29) to get

$$v_1(x_0) - v_2(y_0) \leqq h(\varepsilon) + \sup_{\alpha \in U} [f_1(x_0, \alpha) - f_2(x_0, \alpha)] + C\delta. \tag{2.33}$$

Also we have,

$$\begin{aligned} \max_{x \in \bar{\theta}} \{v_1(x) - v_2(x)\} &\leqq v_1\left(z_\delta + \frac{2\varepsilon}{r}\eta(z_\delta)\right) - v_2(z_\delta) + \delta + \omega_1(c\varepsilon) \\ &\leqq \max\{\phi^\varepsilon(x, y) : x, y \in \bar{\theta}\} + \delta + \omega_1(c\varepsilon). \end{aligned}$$

Using (2.23) and (2.33), one can show that

$$\max\{\phi^\varepsilon(x, y) : x, y \in \bar{\theta}\} \leqq h(\varepsilon) + c\delta + \sup_{x \in \bar{\theta}, \alpha \in U} \{f_1(x, \alpha) - f_2(y, \alpha)\}. \tag{2.34}$$

Now send first ε then δ to zero. □

In fact, using the fact that x_0 is close to z_δ in (2.33), one can improve the result as follows:

COROLLARY 2.3. *Let $z_\delta \in \bar{\theta}$ be as in* (2.18); *then under the hypothesis of Theorem* 2 *we have*

$$v_1(z_\delta) - v_2(z_\delta) \leqq \sup_{\alpha \in U} [f_1(z_\delta, \alpha) - f_2(z_\delta, \alpha)] + C\delta + \omega_{f_1}(C\delta) + \omega_{f_2}(C\delta)$$

where C is a positive constant depending on $K(b)$, $L(b)$.

3. Uniform continuity of the value function. In this section we prove the continuity of the value function under the following assumptions.

(A3) There is a positive constant β such that for any $x \in \partial\theta$ there is $\alpha(x) \in U$ satisfying $b(x, \alpha(x)) \cdot \nu(x) \leqq -\beta < 0$, where ν is the exterior normal vector.

(A4) The boundary $\partial\theta$ is of class C^2.

(A5) If $\partial\theta$ is not compact there are positive constants ρ and l such that, for any $x \in \partial\theta$ there is a $T \in C^2(B(x, \rho)$ with inverse $T^{-1} \in C^1(B(x, \rho))$ satisfying

$$\begin{aligned} &\text{(i)} \quad T(B(x,\rho) \cap \theta) \subset \{y \in R^n, y_n > 0\}, \\ (3.1) \quad &\text{(ii)} \quad T(B(x,\rho) \cap \partial\theta) \subset \{y \in R^n, y_n = 0\}, \\ &\text{(iii)} \quad \|T\|_{C^2(B(x,\rho))} + \|T^{-1}\|_{C^1(B(x,\rho))} \leqq l. \end{aligned}$$

The subscript n denotes the nth component.

Remark. The condition (3.1) is satisfied locally if (A4) holds. By using a technique similar to the one indicated below, one can prove the continuity of the optimal value function under (A3) and (A4). Instead of this we use (A5) together with (A3) and (A4) to obtain a uniform modulus of continuity for the optimal value function.

LEMMA 3.1. *Suppose* (A3)-(A5) *hold. Then, for all* $x \in \partial\theta$, $\bar{b}(x, \alpha(x))_n \geqq l\beta$ *where*

$$(3.2) \qquad \bar{b}(y, \alpha) = \nabla T(T^{-1}(y)) \cdot b(T^{-1}(y), \alpha), \; y \in B(x, \rho) \text{ and } \alpha \in U.$$

Proof. Given $x_0 \in \partial\theta$ pick T as in (A5). Observe $\partial\theta \cap B(x_0, \rho) \subset \{x: T_n(x) = 0\}$. Hence $\nu(x_0) = -\nabla T_n(x_0)/|\nabla T_n(x_0)|$.

$$\begin{aligned} \bar{b}(T(x_0), \alpha(x_0))_n &= \nabla T_n(x_0) \cdot b(x_0, \alpha(x_0)) \\ &= -|\nabla T_n(x_0)| \nu(x_0) \cdot b(x_0, \alpha(x_0)) \geqq l\beta. \end{aligned} \tag{3.3}$$

□

Remark. The vector field $\bar{b}$ is the image of b under the transformation T.

Let u be an admissible control for $x_0 \in \bar{\theta}$. Then u is not necessarily admissible at any point x, regardless how close that point is to x_0. The following lemma provides a way to project it into $\mathcal{A}_x$ by changing the cost proportionally to $|x - x_0|$.

LEMMA 3.2. *Assume that* (A3)-(A5), (1.1)-(1.4) *hold. Then there exist* $t^* > 0$ *and* $L > 0$ *such that for any* $x \in \bar{\theta}$ *and* $u \in \mathcal{A}$ *there is* $\bar{u}$ *in* $\mathcal{A}_x$ *satisfying*

$$(3.4) \qquad |J_{t^*}(x, \bar{u}) - J_{t^*}(x, u)| \leqq L \sup_{t \in [0, t^*]} [\operatorname{dist}(y(x, t, u), \bar{\theta})]$$

where

$$(3.5) \qquad J_{t^*}(x, u) = \int_0^{t^*} e^{-t} f(y(x, t, u), u(t))\, dt.$$

Proof. In the proof we shall determine t^*, sufficiently small. Let t_0 be the first entrance to $\partial\theta$, i.e.,

$$\begin{aligned} t_0 &= \inf\{0 < t \leqq t^*, y(x, t, u) \in \partial\theta\} \\ &\quad \text{or } t^* \text{ if } y(x, t, u) \in \theta \text{ for all } t \leqq t^*. \end{aligned} \tag{3.6}$$

Let $\varepsilon = \sup\{\text{dist. } (y(x, t, u), \bar{\theta}); t \in [0, t^*]\}$. Define $\bar{u}$ as follows

$$(3.7) \qquad \bar{u}(t) = u(t)\chi_{[0, t_0) \cup (t_0 + k\varepsilon, \infty)}(t) + \alpha(y(x, t_0, u))\chi_{[t_0, t_0 + k\varepsilon]}(t)$$

where k is to be chosen and $\alpha(x)$ is as in (A3). We claim that $y(x, t, \bar{u}) \in \bar{\theta}$ for $t \leqq t^*$. At $y(x, t_0, u) \in \partial\theta$ there is a map T satisfying (3.1). Set $z(x, t, u) = T(y(x, t, u))$ for any u in $\mathcal{A}$, then z obeys the differential equation

$$(3.8) \qquad \frac{d}{dt} z(x, t, u) = \bar{b}(z(x, t, u), u(t))$$

where $\bar{b}$ is as in (3.2). The vector field $\bar{b}$ is Lipschitz continuous on $N = T(B(y(x, t_0, u), \rho))$. Moreover, on N it is bounded by $K(\bar{b}) = lK(b)$ and its Lipschitz

constant $L(\bar{b})$ is no more than $l^2K(b)+l^2L(b)$. (Here l is as in (A5).) If we choose t^* less than $\rho(K(b))^{-1}$, then $y(x, t, \bar{u})$ lies in $B(y(x, t_0, u), \rho)$ for all $t \leqq t^*$. Hence, $z(x, t, \bar{u}) \in N$ for all $t \leqq t^*$ and without loss of generality we assume $\bar{b}$ is Lipschitz continuous with Lipschitz constant $l^2K(b)+l^2L(b)$. To prove the claim, it suffices to show $(z(x, t, \bar{u}))_n \geqq 0$ on $t \in [0, t^*]$. Consider

$$\psi(t) = (z(x, t+k\varepsilon, \bar{u}) - z(x, t, u))_n \quad \text{for } t \geqq t_0. \tag{3.9}$$

Then

$$\begin{aligned}
\psi(t) &= \psi(t_0) + \int_{t_0}^{t} [\bar{b}(z(x, s+k\varepsilon, \bar{u}), u(t) - \bar{b}(z(x, s, u), u(t))]_n \, ds \\
&\geqq \psi(t_0) - L(\bar{b})|z(x, t_0+k\varepsilon, \bar{u}) - z(x, t_0, u)| \int_{t_0}^{t} e^{L(\bar{b})(s-t_0)} \, ds \qquad (3.10) \\
&\geqq \psi(t_0) - K(\bar{b})k\varepsilon(e^{L(\bar{b})(t-t_0)} - 1) \\
&\geqq \psi(t_0) - K(\bar{b})k\varepsilon(e^{L(\bar{b})(t^*-t_0)} - 1) \quad \text{for } t \in [t_0, t^*].
\end{aligned}$$

Now choose t^* less than $(L(\bar{b}))^{-1} \ln(1+\beta l/4K(\bar{b}))$, where β is as in (A3). We have

$$\psi(t) \geqq \psi(t_0) - k\varepsilon\beta l_4^1 \quad \text{for } t \in [t_0, t^*]. \tag{3.11}$$

We need an estimate for $\psi(t_0)$. To simplify the notation, let $\bar{b}_0 = \bar{b}(z(x, t_0, u))$, $\alpha(y(x, t_0, u))$ and $\omega(t) = z(x, t_0, u) + (t-t_0)\bar{b}_0$. Then, Lemma 3.1 yields that $(\bar{b}_0)_n \geqq \beta l$ and hence,

$$\omega(t)_n \leqq \beta l(t-t_0) \quad \text{for } t \geqq t_0. \tag{3.12}$$

Using standard O.D.E. estimates, one can obtain

$$|z(x, t, \bar{u}) - \omega(t)| \leqq \tfrac{1}{2}L(\bar{b})K(\bar{b})(t-t_0)^2 \quad \text{for } t \geqq t_0. \tag{3.13}$$

Thus

$$z(x, t_0+k\varepsilon, \bar{u})_n \geqq \beta lk\varepsilon - \tfrac{1}{2}L(\bar{b})K(\bar{b})(k\varepsilon)^2. \tag{3.14}$$

Since $y(x, t_0, u) \in \partial\theta$, we have $\psi(t_0) = z(x, t_0+k\varepsilon, \bar{u})_n$. Substitute (3.14) into (3.11) to get

$$\psi(t) \geqq k\varepsilon(\beta l_4^3 - \tfrac{1}{2}L(\bar{b})K(\bar{b})k\varepsilon). \tag{3.15}$$

Choose k to be the minimum of $\beta l/2\varepsilon L(\bar{b})K(\bar{b})$ and $2/\beta$. Then for $t \leqq t^*$

$$\psi(t) \geqq \varepsilon l = \sup\{[-z(x, t, u)_n], t \in [0, t^*]\}. \tag{3.16}$$

Hence, $z(x, t+k\varepsilon, \bar{u})_n = \psi(t) + z(x, t, u)_n \geqq 0$ for all $t \in [t_0, t^*]$. One can prove (3.4) by using the standard estimates. □

THEOREM 3.3. *Suppose* (A3)-(A5), (1.1), (1.2) *and* (1.4) *hold. Then the value function v is in $BUC(\bar{\theta})$.*

Proof. Without loss of generality one can assume f is Lipschitz in x uniformly with respect to α. If not, we take a sequence f^n of Lipschitz continuous functions converging to f uniformly. Let $x, y \in \bar{\theta}$ and $|x-y| < r$. For any positive δ, pick a δ-optimal control u in $\mathcal{A}_y$, i.e.

$$J_{t^*}(y, u) + e^{-t^*}v(y(y, t^*, u)) \leqq v(y) + \delta \tag{3.17}$$

where t^* is as in Lemma 3.2. Construct $\bar{u} \in \mathcal{A}_x$ as in Lemma 3.2, and set $\varepsilon = \sup[\text{dist}(y(x, t, u), \bar{\theta}), t \in [0, t^*]]$. Using standard estimates, one can get $\varepsilon \leqq Cr$ for

some C positive. Thus, we have

$$|J_{t^*}(x, u) - J_{t^*}(x, \bar{u})| \leqq LCr. \tag{3.18}$$

Also the construction of $\bar{u}$ yields

$$|y(x, t^*, u) - y(x, t^*, \bar{u})| \leqq \bar{C}r \quad \text{for some } \bar{C} > 0. \tag{3.19}$$

Also

$$|y(x, t^*, \bar{u}) - y(y, t^*, u)| \leqq \bar{C}r + |y(x, t^*, u) - y(y, t^*, u)| \leqq \tilde{C}r \quad \text{for some } \tilde{C} > 1. \tag{3.20}$$

And

$$|J_{t^*}(x, \bar{u}) - J_{t^*}(y, u)| \leqq LCr + |J_{t^*}(x, u) - J_{t^*}(y, u)| \leqq Cr \quad \text{for some } C > 0. \tag{3.21}$$

Let $\omega(r) = \sup\{|v(x) - v(y)|, x, y \in \bar{\theta}, |x - y| < r\}$ for $r > 0$. At the origin $\omega(0) = \lim_{r \downarrow 0} \omega(r)$. Combine (3.17), (3.20) and (3.21) and use the dynamic programming principle to obtain

$$\begin{aligned} v(x) - v(y) &\leqq J_{t^*}(x, \bar{u}) + e^{-t^*} v(y(x, t^*, \bar{u})) \\ &\quad - J_{t^*}(y, u) - e^{-t^*} v(y(y, t^*, u)) + \delta \\ &\leqq Cr + e^{-t^*} \omega(\tilde{C}r) + \delta \quad \text{for all } \delta > 0. \end{aligned} \tag{3.22}$$

Hence we have

$$\omega(r) \leqq Cr + e^{-t^*} \omega(\tilde{C}r) \quad \text{and} \quad \tilde{C} > 1. \tag{3.23}$$

Assume $\tilde{C} e^{-t^*} \neq 1$ and iterate (3.23) to obtain

$$\begin{aligned} \omega(0) &= \lim_{n \to \infty} \omega(\tilde{C}^{-n}) \leqq \lim_{n \to \infty} \left[C\tilde{C}^{-n} \sum_{l=0}^{n-1} [\tilde{C} e^{-t^*}]^l + e^{-nt^*} \omega(1) \right] \\ &= C \lim_{n \to \infty} \left[\frac{e^{-nt^*} - \tilde{C}^{-n}}{\tilde{C} e^{-t^*} - 1} \right] = 0. \end{aligned}$$

□

Acknowledgments. Finally, I would like to express thanks to Professor P. E. Souganidis for introducing me to the notion of viscosity solutions and their applications to control problems and to Professor W. H. Fleming for suggesting the problem, helpful converstations and good advice. This paper comprises part of the author's dissertation written under the direction of Professor W. H. Fleming at Brown University.

REFERENCES

[1] 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.

[2] M. G. CRANDALL AND P.-L. LIONS, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1-42.

[3] M. G. CRANDALL AND P. E. SOUGANIDIS, *Developments in the theory of nonlinear first-order partial differential equations*, in Proc. International Symposium on Differential Equations, Birmingham, Alabama, Knowles and Lewis, eds., North-Holland, Amsterdam, 1983.

[4] W. H. FLEMING AND R. RISHEL, *Deterministic and Stochastic Optimal Control*, Springer, Berlin, 1975.

[5] P.-L. LIONS, *Generalized Solutions of Hamilton-Jacobi Equations*, Research Notes in Mathematics 69, Pitman, London, 1982.

[6] ———, *Optimal control and viscosity solutions*, Proc. Rome meeting, 1984, to appear in Springer Lecture Notes in Mathematics.

[7] P. E. SOUGANIDIS AND P.-L. LIONS, *Differential games, optimal control and directional derivatives to viscosity solutions of Bellman's and Isaacs' equations*, this Journal, 23 (1985), pp. 566-573.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 3, May 1986

015

DECOUPLAGE DES SYSTEMES NON LINEAIRES, SERIES GENERATRICES NON COMMUTATIVES ET ALGEBRES DE LIE*

DANIEL CLAUDE†

Résumé. L'utilisation conjointe des séries génératrices et des algèbres de Lie rend possible une approche algébrique du découplage des systèmes non linéaires. On présente une méthode de découplage cohérente grâce à la notion supplémentaire d'immersion d'un système dans un autre, concept qui formalise la notion de même comportement entrée-sortie. Ainsi, un choix judicieux des lois de bouclage assurant le découplage permet d'immerger un système dans un système défini sur $\mathbb{R}^\eta$, avec η dépendant des nombres caractéristiques du système.

Abstract. The joint use of noncommutative generating power series and Lie algebras gives an algebraic approach for decoupling problem of nonlinear systems. We produce a coherent decoupling method using the additional notion of the immersion of a system into another, a concept which makes clear the fact that two systems have a same input-output behaviour. Thus, a judicious choice of feedback which gives decoupling is able to immerse a system into a system defined on $\mathbb{R}^\eta$, with η depending on the characteristic numbers of the system.

Key words. nonlinear systems, decoupling, noninteracting control, disturbance rejection, noncommutative generating series, Lie algebras

On considère des fonctionnelles causales, en temps continu, associant une entrée vectorielle $\mathbf{e}$ à une sortie vectorielle y par l'intermédiaire d'une dynamique donnée par un système différentiel. Ces systèmes dynamiques multivariables sont de la forme:

$$\begin{aligned} &\frac{dq}{dt} = \dot{q}(t) = f(q) + \mathbf{e}(t)g(q), \\ &y(t) = h(q) \end{aligned} \tag{1}$$

où l'état q est de dimension N.

La commande d'un système de type (1) pose immédiatement à tout utilisateur le problème suivant: comment éviter qu'une entrée perturbe une sortie ou bien, à cause des couplages inhérents à ces systèmes, affecte plusieurs sorties simultanément? Par exemple, dans le pilotage d'un avion, que faire pour qu'une action sur le gouvernail d'altitude ne modifie pas en même temps le vecteur instantané de rotation et les angles d'attaque, de roulis, de virage et de tangage (cf. Singh et Schy [33])? La réponse à ces interrogations est donnée par la théorie du découplage qui consiste à rendre les sorties d'un système indépendantes de certaines entrées. Pour y parvenir, un système ayant rarement les propriétés souhaitées, on utilisera ici des bouclages ou "feedbacks", statiques par retour d'état du type:

$$\mathbf{e}(t) = \alpha(q) + \mathbf{v}(t)\beta(q) \tag{2}$$

où $\mathbf{v}$ indique une nouvelle entrée vectorielle permettant la commande du système.

Le système (1) bouclé devient:

$$\begin{aligned} &\dot{q} = [f(q) + \alpha(q)g(q)] + \mathbf{v}(t)\beta(q)g(q), \\ &y = h(q) = (h_1(q), \cdots, h_r(q)). \end{aligned} \tag{3}$$

* Received by the editors September 14, 1983, and in final revised form January 13, 1985.

† Laboratoire des Signaux et Systèmes, C.N.R.S.-E.S.E., Plateau du Moulon, 91190 Gif-sur-Yvette, France.

On cherche alors les bouclages donnant le découplage désiré entre l'entrée $\mathbf{v}$ et la sortie y.

La solution du problème pour les systèmes linéaires[1]—systèmes où l'espace d'état Q est un espace vectoriel de dimension finie N, où f et g sont respectivement des champs de vecteurs linéaire et constant, et où la fonction de sortie $h: Q \to \mathbb{R}^r$ est linéaire—a été donnée simultanément autour des années soixante-dix, grâce à l'approche géométrique, par Basile et Marro [1] et par Wonham et Morse [39].

Pour les systèmes non linéaires, ce problème, important par ses applications pratiques, restait ouvert.

Les progrès récents dus à l'usage de la géométrie différentielle, introduite en automatique non linéaire par Hermann [18] et Lobry [24][2], ont suggéré à Hirschorn [20] et Isidori, Krener, Gori-Giorgi et Monaco [22] l'utilisation d'un certain type de distributions, dites (f, g)-invariantes, pour attaquer le découplage en non linéaire. Cette approche géométrique, dans laquelle la notion de sous-espace vectoriel, utilisée en linéaire, est remplacée par celle de distribution, se révèle fructueuse et apparaît dans l'étude de systèmes très généraux (cf. Nijmeijer et van der Schaft [28]).

Cependant, seules les distributions de rang constant présentent un réel intérêt géométrique par les changements de coordonnées qu'elles permettent (cf. Isidori [21], Respondek [32], Nijmeijer [27], Bournonville [2]) et l'existence de singularités conduit, contrairement au cas linéaire, à des difficultés redoutables (cf. Byrnes et Krener [3]). Comme le montre Isidori [21], le découplage peut être assuré par l'existence d'une carte locale dans laquelle la dynamique du système se décompose en deux parties dont une est inobservable. Mais, pour avoir une valeur globale, cela impose des conditions restrictives incompatibles avec la présence des singularités les plus communes. Tout ceci nous amène à retenir, pour le non linéaire, deux types de découplage d'un système, le découplage structurel lié à une décomposition de la dynamique (cf. Isidori [21]) et le découplage fonctionnel, déterminé uniquement par le comportement externe (entrée-sortie), comportement lui-même caractérisé par la série génératrice non commutative du système (cf. Fliess [11]). Ce dernier découplage a l'avantage de permettre des solutions sans les conditions de régularité nécessaires au découplage structurel qu'il redonne en plus facilement. Mais les difficultés évoquées plus haut demeurent et la recherche de bouclages donnant un découplage fonctionnel conduit en général à travailler sur un ouvert dense de la variété d'état (cf. Hirschorn [20]). Dans beaucoup d'exemples concrets définis sur $\mathbb{R}^N$, cet ouvert est le complémentaire d'un hyperplan (cf. Singh et Schy [33]) dont la simple approche peut entraîner des instabilités numériques rédhibitoires.

L'utilisation des séries génératrices non commutatives, inséparablement liées à la géométrie différentielle (cf. Fliess [13]), permet de traiter sous un aspect plus algébrique le découplage des systèmes non linéaires. L'emploi d'algèbres de Lie, sur l'anneau des fonctions analytiques sur la variété d'état, fournit alors un équivalent algébrique aux distributions analytiques. Cela conduit de façon naturelle aux lois de bouclage cherchées et complète la méthode géométrique tout en généralisant une méthode bien connue des ingénieurs (cf. Porter [30]). Il est alors possible de tenir compte de nouvelles données comme par exemple les contraintes physiques (cf. Gauthier, Bornard, Bacha et Idir [16]). De plus, la notion de découplage fini qui implique la nullité d'un nombre fini de premiers termes de la série génératrice du système au lieu d'une infinité pour

[1] On pourra consulter le livre de Wonham [38] pour une étude complète du découplage des systèmes linéaires.

[2] On pourra consulter un cours avancé de Sussmann [35] pour un "survey" récent.

le découplage fonctionnel, permet des procédés algorithmiques nouveaux (cf. Claude et Dufresne [7] et Kasinski et Lévine [23]). Le découplage fonctionnel, complété par le découplage fini, rend possible la présentation d'une méthode de découplage cohérente grâce à la notion dernière d'immersion d'un système dans un autre. Ce concept d'immersion (cf. Fliess [12] et Fliess et Kupka [14]) qui formalise la notion de même comportement entrée-sortie, offre une solution intéressante. Ainsi, un choix judicieux de lois de bouclage assurant le découplage permet d'immerger le système dans un système défini sur $\mathbb{R}^\eta$, avec η dépendant des nombres caractéristiques du système.

Il est à noter que nos démonstrations restent dans un cadre strictement algébrique.

La plupart des résultats de cet article ont déjà été publiés sous une forme souvent succincte en [4], [5], [6].

Table des Matières

1. Séries génératrices et découplage. Le découplage d'un système consistant à rendre les sorties d'un système indépendantes de certaines entrées, on est conduit dans le système (3) à associer à chaque sortie une partition de l'entrée $\mathbf{v}$ en deux sous-ensembles: les commandes admissibles et les perturbations. Ainsi, tout découplage se ramène à une association de rejets de perturbations, et nous ne traitons que ce dernier problème.

Nous considérons des systèmes à coefficients analytiques où les entrées apparaissent au premier degré, de la forme:

$$\text{(I)} \qquad \begin{aligned} \dot{q}(t) &= A^0(q) + \sum_{l=1}^{d} \mathbf{u}_l(t) A^l(q) + \sum_{j=1}^{k} \mathbf{p}_j(t) P^j(q), \\ y &= h(q). \end{aligned}$$

L'état q appartient à une variété analytique réelle Q connexe et de dimension N; les champs de vecteurs $A^0, A^1, \cdots, A^d, P^1, \cdots, P^k: Q \to TQ$ (fibré tangent) et la fonction vectorielle $h = (h_1, \cdots, h_r)$ de Q dans $\mathbb{R}^r$, sont analytiques; les entrées $\mathbf{u}_1, \cdots, \mathbf{u}_d, \mathbf{p}_1, \cdots, \mathbf{p}_k$ sont des fonctions réelles continues par morceaux. On désire que la sortie vectorielle $y = (y_1, \cdots, y_r)$ ne dépende pas de l'entrée $\mathbf{p} = (\mathbf{p}_1, \cdots, \mathbf{p}_k)$, au moyen de bouclages du genre:

$$(*) \qquad \begin{aligned} \mathbf{u}_l(t) &= \alpha_l(q) + \sum_{i=1}^{n} \beta_l^i(q) \mathbf{a}_i(t) \quad \text{avec } l = 1, \cdots, d, \\ \mathbf{p}_j(t) &= \mathbf{p}_j(t) \quad \text{avec } j = 1, \cdots, k. \end{aligned}$$

Les fonctions α_l et β_l^i sont analytiques et $\mathbf{a}_1, \cdots, \mathbf{a}_n$ désignent de nouvelles entrées permettant la commande du système bouclé.

En prenant $\alpha_l = 0$ et $\beta_l^i = \delta_l^i$ (symbole de Kronecker) avec $l = 1, \cdots, d$ et $i = 1, \cdots, n$, on retrouve le système (I).

Le système bouclé peut s'écrire:

$$\text{(II)} \qquad \begin{aligned} \dot{q} &= \hat{A}^0 + \sum_{i=1}^{n} \mathbf{a}_i \hat{A}^i + \sum_{j=1}^{k} \mathbf{p}_j P^j, \\ y &= h(q) \end{aligned}$$

avec

$$\hat{A}^0 = A^0 + \sum_{l=1}^{d} \alpha_l A^l \quad \text{et} \quad \hat{A}^i = \sum_{l=1}^{d} \beta_l^i A^l \qquad (i = 1, \cdots, n).$$

Comme l'a montré Fliess [11], la sortie $y(t)$ du système différentiel (II), initialisé en $q(0)$, est une fonctionnelle causale analytique des entrées $\mathbf{a}_1, \cdots, \mathbf{a}_n, \mathbf{p}_1, \cdots, \mathbf{p}_k$ définie par une série génératrice. Cette série génératrice $\mathfrak{h}|q(0) = (\mathfrak{h}_1|q(0), \cdots, \mathfrak{h}_r|q(0))$ et la sortie vectorielle $y = (y_1, \cdots, y_r)$ sont données par:

$$\text{(4)} \qquad \mathfrak{h}_s|q(0) = h_s|q(0) \cdot 1 + \sum_{\nu \geqq 0} \sum_{j_0, \cdots, j_\nu = 0}^{n+k} C^{j_0} \cdots C^{j_\nu} \cdot h_s|q(0) \cdot \mathbf{c}_{j_\nu} \cdots \mathbf{c}_{j_0} \qquad (s = 1, \cdots, r),$$

$$\text{(5)} \qquad y_s(t) = h_s|q(0) + \sum_{\nu \geqq 0} \sum_{j_0, \cdots, j_\nu = 0}^{n+k} C^{j_0} \cdots C^{j_\nu} \cdot h_s|q(0) \cdot \int_0^t d\xi_{j_\nu} \cdots d\xi_{j_0} \qquad (s = 1, \cdots, r).$$

Cette dernière série n'étant convergente en général que pour des temps courts et de petites entrées.

Chaque $\mathbf{c}_j$ appartient à l'alphabet

$$\begin{aligned} \mathbf{c} &= \{\mathbf{a}_0, \mathbf{a}_1, \cdots, \mathbf{a}_n, \mathbf{p}_1, \cdots, \mathbf{p}_k\} \\ &= \{\mathbf{c}_0, \mathbf{c}_1, \cdots, \mathbf{c}_n, \mathbf{c}_{n+1}, \cdots, \mathbf{c}_{n+k}\} \\ &= \mathbf{a} \cup \mathbf{p} \quad \text{avec } \mathbf{a} = \{\mathbf{a}_0, \cdots, \mathbf{a}_n\} \quad \text{et} \quad \mathbf{p} = \{\mathbf{p}_1, \cdots, \mathbf{p}_k\}. \end{aligned}$$

Chaque C^j appartient à l'ensemble des champs de vecteurs du système

$$\begin{aligned} \mathscr{C} &= \{\hat{A}^0, \cdots, \hat{A}^n, P^1, \cdots, P^k\} = \{C^0, C^1, \cdots, C^n, C^{n+1}, \cdots, C^{n+k}\} \\ &= \mathscr{A} \cup \mathscr{P} \quad \text{avec } \mathscr{A} = \{\hat{A}^0, \cdots, \hat{A}^n\} \quad \text{et} \quad \mathscr{P} = \{P^1, \cdots, P^k\}. \end{aligned}$$

La barre $|q(0)$ désigne l'évaluation en $q(0)$.

L'intégrale itérée $\int_0^t d\xi_{j_\nu} \cdots d\xi_{j_0}$ est définie par récurrence:

$$\xi_0(t) = t, \qquad \xi_j(t) = \int_0^t \mathbf{a}_i(\tau)\, d\tau \qquad (i = 1, \cdots, n),$$

$$\xi_l(t) = \int_0^t \mathbf{p}_{l-n}(\tau)\, d\tau \qquad (l = n+1, \cdots, n+k)$$

avec

$$\int_0^t d\xi_j = \xi_j(t) \qquad (j = 0, \cdots, n+k),$$

$$\int_0^t d\xi_{j_\nu} \cdots d\xi_{j_0} = \int_0^t d\xi_{j_\nu}(\tau) \int_0^\tau d\xi_{j_{\nu-1}} \cdots d\xi_{j_0},$$

la dernière intégrale étant prise au sens de Stieltjes.

Nous avons montré en [6] que pour les systèmes initialisés, le découplage peut seulement s'exprimer à l'aide des séries génératrices. L'initialisation ne permet pas de formuler des conditions équivalentes de découplage liées aux champs de vecteurs du système, ni de calculer aisément les lois de bouclages nécessaires au découplage. Aussi, considérons-nous des systèmes de type (II) sans initialisation préalable.

Pour un système de type (II) non initialisé, on peut encore associer une série génératrice vectorielle $\mathfrak{h} = (\mathfrak{h}_1, \cdots, \mathfrak{h}_r)$ dans laquelle les coefficients de $\mathfrak{h}_s$ $(s = 1, \cdots, r)$ sont des éléments de $C^\omega(Q)$, l'anneau des fonctions analytiques sur la variété Q. $\mathfrak{h}$ est définie par:

$$\mathfrak{h}_s = h_s \cdot 1 + \sum_{\nu \geqq 0} \sum_{j_0, \cdots, j_\nu = 0}^{n+k} C^{j_0} \cdots C^{j_\nu} \cdot h_s \cdot \mathbf{c}_{j_\nu} \cdots \mathbf{c}_{j_0} \qquad (s = 1, \cdots, r).$$

On peut maintenant préciser la notion de découplage que nous utilisons.

Découplage total (fonctionnel). La sortie y d'un système de type (II) est découplée totalement par rapport à l'entrée $\mathbf{p}$, si et seulement si chaque sortie $y_1, \cdots, y_r$ ne dépend pas des entrées $\mathbf{p}_1, \cdots, \mathbf{p}_k$, quelle que soit l'initialisation $q(0)$ du système.

L'utilisation des séries génératrices éclaire le concept de découplage d'un système en le reliant à la nullité de certains coefficients.

PROPOSITION 1. *La sortie y d'un système de type* (II) *est totalement découplée par rapport à* $\mathbf{p}$ *si et seulement si tous les coefficients* $C^{j_0} \cdots C^{j_\nu} \cdot h_s$ $(s = 1, \cdots, r)$ *sont nuls dès qu'il existe au moins un* C^j *appartenant à* $\mathcal{P} = \{P^1, \cdots, P^k\}$, *l'ensemble des champs de vecteurs liés aux perturbations* $\mathbf{p}_1, \cdots, \mathbf{p}_k$.

Preuve. L'utilisation de la formule (5) montre que la condition est évidemment suffisante. Réciproquement, l'unicité de la série génératrice, montrée par Fliess [11] dans le cas de coefficients réels, s'étend bien évidemment au cas de coefficients analytiques et permet de conclure. En effet, si la sortie y du système (II) est totalement découplée par rapport à $\mathbf{p}$, elle représente une fonctionnelle causale analytique des entrées $\mathbf{a}_1, \cdots, \mathbf{a}_n$. Par conséquent, cette fonctionnelle est définie par une série génératrice $\mathfrak{h} = (\mathfrak{h}_1, \cdots, \mathfrak{h}_r)$ où $\mathfrak{h}_s$ $(s = 1, \cdots, r)$ est élément de $C^\omega(Q)\langle\langle \mathbf{a} \rangle\rangle$, la $C^\omega(Q)$-algèbre des séries formelles, à coefficients dans $C^\omega(Q)$, en les variables associatives $\mathbf{a}_0, \cdots, \mathbf{a}_n$. Comme chaque série $\mathfrak{h}_s$ $(s = 1, \cdots, r)$ est aussi par définition élément de $C^\omega(Q)\langle\langle \mathbf{c} \rangle\rangle$, la $C^\omega(Q)$-algèbre des séries formelles à coefficients dans $C^\omega(Q)$, en les indéterminées associatives $\mathbf{a}_0, \cdots, \mathbf{a}_n, \mathbf{p}_1, \cdots, \mathbf{p}_k$, on a bien la nullité des coefficients $C^{j_0} \cdots C^{j_\nu} \cdot h_s$ $(s = 1, \cdots, r)$ pour lesquels il existe au moins un C^j appartenant à $\mathcal{P}$. □

De la même manière, la valeur d'une fonctionnelle analytique dépendant d'intégrales itérées, nous voyons (cf. Fliess [11], formules de majoration) que la contribution des termes apparaissant dans la série génératrice décroît avec la longueur et est négligeable, en pratique, à partir d'un certain rang. Nous allons appliquer cette propriété au découplage. On note $\mathbf{c}^*$ le monoïde libre engendré par l'alphabet $\mathbf{c}$. $\mathbf{c}^*$ est l'ensemble des mots formés avec les lettres de $\mathbf{c}$ et est muni de la multiplication non commutative définie par la concaténation de deux mots, l'élément neutre étant le mot vide noté 1. On a ainsi, pour tout mot $\mathbf{w}$ de $\mathbf{c}^*$:

$$1 \cdot \mathbf{w} = \mathbf{w} \cdot 1 = \mathbf{w}$$

$$\text{et si, } \mathbf{w}_1 = \mathbf{c}_{j_\nu} \cdots \mathbf{c}_{j_0};\ \mathbf{w}_2 = \mathbf{c}_{j'_{\nu'}} \cdots \mathbf{c}_{j'_0} \text{ alors } \mathbf{w}_1 \cdot \mathbf{w}_2 = \mathbf{c}_{j_\nu} \cdots \mathbf{c}_{j_0} \mathbf{c}_{j'_{\nu'}} \cdots \mathbf{c}_{j'_0}.$$

$\mathbf{c}^*\mathbf{p}\mathbf{c}^*$ désigne l'ensemble des mots de $\mathbf{c}^*$ contenant au moins une occurrence de $\mathbf{p}_j$ $(j = 1, \cdots, k)$. En notant W l'expression obtenue en remplaçant $\mathbf{a}_0$ par $\hat{A}^0$, $\mathbf{a}_i$ par $\hat{A}^i$ $(i = 1, \cdots, n)$ et $\mathbf{p}_j$ par P^j $(j = 1, \cdots, k)$ dans chaque mot $\mathbf{w}$ de $\mathbf{c}^*$, on peut énoncer:

Découplage fini. Un système de type (II) a sa sortie y_s $(s=1,\cdots,r)$ découplée par rapport à l'entrée **p**, à l'ordre μ si et seulement si l'une des conditions suivantes est satisfaite:

(i) $\mathscr{P}\cdot h_s \not\equiv 0$;

(ii) $W\cdot h_s = C^{j_0}\cdots C^{j_\nu}\cdot h_s \equiv 0$

pour tout $\mathbf{w}\in(\mathbf{c^*pc^*})_\mu$, l'ensemble des mots de **c*pc*** de longueur inférieure ou égale à μ avec $\mu \geqq 1$.

L'ordre μ est nul dans le premier cas alors qu'il est supérieur ou égal à un dans le second.

Si une sortie y_s du système (II) n'est pas découplée totalement par rapport à **p**, on peut définir *l'ordre de découplage fini maximal* auquel est découplée y_s par rapport à **p**, c'est le plus grand des ordres de découplage fini de y_s par rapport à **p**.

2. Algèbres de découplage. Les systèmes (II) étant pris non initialisés, on va pouvoir donner, à l'aide des propositions qui suivent, des conditions de découplage liées aux champs de vecteurs du système et aux opérateurs différentiels qu'ils définissent.

Soient $\mathscr{D}_{\mathscr{A}}$ et $\mathfrak{L}_{\mathscr{A}}$, respectivement, la $\mathbb{R}$-algèbre des opérateurs différentiels et la $C^\omega(Q)$-algèbre de Lie engendrées par $\mathscr{A}$. On peut énoncer:

PROPOSITION 2. *Les conditions suivantes sont équivalentes*:

1. *Tous les coefficients* $C^{j_0}\cdots C^{j_\nu}\cdot h_s$ $(s=1,\cdots,r)$ *contenant au moins un élément de* $\mathscr{P}$ *sont identiquement nuls.*

2. $\mathscr{P}\cdot f\equiv 0$ *pour les fonctions* f *du* $\mathscr{D}_{\mathscr{A}}$*-module à gauche engendré par les fonctions* h_s $(s=1,\cdots,r)$.

3. *Les opérateurs différentiels* D *du* $\mathscr{D}_{\mathscr{A}}$*-module bilatère engendré par* $\mathscr{P}$ *vérifient* $D\cdot h_s\equiv 0$ $(s=1,\cdots,r)$.

4. *Les opérateurs différentiels* D *du* $\mathscr{D}_{\mathscr{A}}$*-module à droite engendré par* $\mathscr{P}$ *satisfont à* $D\cdot h_s\equiv 0$ $(s=1,\cdots,r)$.

5. *La plus petite* $C^\omega(Q)$*-algèbre de Lie de champs de vecteurs contenant* $\mathscr{P}$ *et normalisée par* $\mathfrak{L}_{\mathscr{A}}$*, annihile les fonctions* h_s $(s=1,\cdots,r)$. *On note cette algèbre* $\mathfrak{T}_{\mathbf{p}}$.

Preuve. On a immédiatement les équivalences $1\Leftrightarrow 2$, $2\Leftrightarrow 3$ et $3\Leftrightarrow 4$. D'autre part, $\mathfrak{T}_{\mathbf{p}}$ est engendrée par $\mathscr{T}$, l'ensemble des champs de vecteurs de la forme

$$\operatorname{ad}\hat{A}^{i_1}\circ\cdots\circ\operatorname{ad}\hat{A}^{i_\nu}(P^j)\quad \text{avec } j\in\{1,\cdots,k\}$$

et $i_1,\cdots,i_\nu\in\{0,\cdots,n\}$. On va montrer que $5\Rightarrow 4$ par récurrence.

Prenons un élément $C\cdot\hat{A}^{i_1}\cdots\hat{A}^{i_\nu}\cdot h_s$ avec $C\in\mathscr{T}$, $\nu\geqq 0$ et $s=1,\cdots,r$. On a:

$$C\cdot\hat{A}^{i_1}\cdots\hat{A}^{i_\nu}\cdot h_s=[C,\hat{A}^{i_1}]\cdot\hat{A}^{i_2}\cdots\hat{A}^{i_\nu}\cdot h_s+\hat{A}^{i_1}\cdot C\cdot\hat{A}^{i_2}\cdots\hat{A}^{i_\nu}\cdot h_s.$$

Mais, $-[C,\hat{A}^{i_1}]=\operatorname{ad}\hat{A}^{i_1}(C)$ appartient à $\mathscr{T}$ et d'après l'hypothèse de récurrence,

$$[C,\hat{A}^{i_1}]\cdot\hat{A}^{i_2}\cdots\hat{A}^{i_\nu}\cdot h_s\equiv 0\quad\text{et}\quad C\cdot\hat{A}^{i_2}\cdots\hat{A}^{i_\nu}\cdot h_s\equiv 0;$$

donc:

$$\hat{A}^{i_1}\cdot C\cdot\hat{A}^{i_2}\cdots\hat{A}^{i_\nu}\cdot hs\equiv 0\quad\text{et ainsi}\quad C\cdot\hat{A}^{i_1}\cdots\hat{A}^{i_\nu}\cdot h_s\equiv 0.$$

La propriété étant vérifiée pour $\nu=0$, l'implication annoncée est bien démontrée si on prend $C\in\mathscr{P}$. Il est clair de plus que $4\Rightarrow 5$. □

Nous allons maintenant énoncer le pendant des résultats bien connus en automatique linéaire (cf. Wonham [38]).

On désigne par $\mathfrak{H}$, la $C^\omega(Q)$-algèbre de Lie des champs de vecteurs associés à h par:

$$\mathfrak{H}=\bigcap_{s=1}^{r}\mathfrak{H}_s\quad\text{avec } C\in\mathfrak{H}_s \text{ ssi } C\cdot h_s\equiv 0.$$

THÉORÈME 1. *La sortie y d'un système noninitialisé de type* (II) *est totalement découplée par rapport à l'entrée* **p**, *si et seulement si, il existe une sous-algèbre* $\mathfrak{D}$ *de la* $C^\omega(Q)$*-algèbre de Lie* $\mathfrak{Q}$ *des champs de vecteurs sur* Q, *telle que*:

$$[\mathscr{A}, \mathfrak{D}] \subset \mathfrak{D} \quad \text{et} \quad \mathscr{P} \subset \mathfrak{D} \subset \mathfrak{H}. \tag{6}$$

Preuve. D'après la définition de l'algèbre $\mathfrak{T}_{\mathbf{p}}$ on a

$$\mathscr{P} \subset \mathfrak{T}_{\mathbf{p}} \quad \text{et} \quad [\mathscr{A}, \mathfrak{T}_{\mathbf{p}}] \subset \mathfrak{T}_{\mathbf{p}}.$$

Des propositions 1 et 2, si le système (II) a sa sortie y totalement découplée par rapport à **p**, on a aussi:

$$\mathfrak{T}_{\mathbf{p}} \subset \mathfrak{H}.$$

Réciproquement, s'il existe une sous-algèbre $\mathfrak{D}$ de l'algèbre de Lie $\mathfrak{Q}$ telle que:

$$[\mathscr{A}, \mathfrak{D}] \subset \mathfrak{D} \quad \text{et} \quad \mathscr{P} \subset \mathfrak{D} \subset \mathfrak{H}$$

alors, on a bien $\mathscr{T} \subset \mathfrak{H}$ et $\mathfrak{T}_{\mathbf{p}} \subset \mathfrak{H}$. Les propositions 1 et 2 donnent la conclusion. □

Nous appelons *algèbre de découplage* du système (II) toute sous-algèbre $\mathfrak{D}$ de l'algèbre de Lie $\mathfrak{Q}$ vérifiant (6).

On note $\mathfrak{D}_h^{\mathbf{a}}$, la $C^\omega(Q)$-algèbre de Lie des champs de vecteur sur Q qui annihilent les fonctions du $\mathscr{D}_{\mathscr{A}}$-module à gauche engendré par les fonctions h_s ($s = 1, \cdots, r$). On a alors:

COROLLAIRE. *Une condition nécessaire et suffisante pour qu'un système non initialisé de type* (II) *ait sa sortie y totalement découplée par rapport à l'entrée* **p** *est que l'une des conditions équivalentes suivantes soit vérifiée*:

(i) $\mathfrak{T}_{\mathbf{p}} \subset \mathfrak{H}$;

(ii) $\mathscr{P} \subset \mathfrak{D}_h^{\mathbf{a}}$.

Pour un bouclage de type (*) donné, $\mathfrak{T}_{\mathbf{p}}$ est *la plus petite algèbre de découplage* de la sortie y du système (II) par rapport à **p** et $\mathfrak{D}_h^{\mathbf{a}}$ *la plus grande.*

Preuve. Pour un bouclage de type (*) tel que le système non initialisé de type (II) ait sa sortie y totalement découplée par rapport à l'entrée **p**, nous avons vu au cours de la démonstration du théorème 1 que $\mathfrak{T}_{\mathbf{p}}$ est la plus petite algèbre de découplage.

D'après les définitions, on a:

$$\mathfrak{D}_h^{\mathbf{a}} \subset \mathfrak{H} \quad \text{et} \quad [\mathscr{A}, \mathfrak{D}_h^{\mathbf{a}}] \subset \mathfrak{D}_h^{\mathbf{a}}$$

et la condition $\mathscr{P} \subset \mathfrak{D}_h^{\mathbf{a}}$ est ainsi une condition nécessaire et suffisante de découplage. De plus, si $\mathfrak{M}$ est un $C^\omega(Q)$-module de champs de vecteurs sur Q tel que $\mathfrak{M} \subset \mathfrak{H}$ et $[\mathscr{A}, \mathfrak{M}] \subset \mathfrak{M}$ alors $\mathfrak{M} \subset \mathfrak{D}_h^{\mathbf{a}}$.

On fait une démonstration par récurrence. Prenons un élément $C \cdot \hat{A}^{i_1} \cdots \hat{A}^{i_\nu} \cdot h_s$ où $C \in \mathfrak{M}$, $\nu \geqq 0$ et $s = 1, \cdots, r$.

On a, comme dans la démonstration de la proposition 1:

$$C \cdot \hat{A}^{i_1} \cdots \hat{A}^{i_\nu} \cdot h_s = [C, \hat{A}^{i_1}] \cdot \hat{A}^{i_2} \cdots \hat{A}^{i_\nu} \cdot h_s + \hat{A}^{i_1} \cdot C \cdot \hat{A}^{i_2} \cdots \hat{A}^{i_\nu} \cdot h_s$$

mais, $[C, \hat{A}^{i_1}] \in \mathfrak{M}$ et d'après l'hypothèse de récurrence,

$$[C, \hat{A}^{i_1}] \cdot \hat{A}^{i_2} \cdots \hat{A}^{i_\nu} \cdot h_s \equiv 0 \quad \text{et} \quad C \cdot \hat{A}^{i_2} \cdots \hat{A}^{i_\nu} \cdot h_s \equiv 0,$$

d'où la nullité de $C \cdot \hat{A}^{i_1} \cdots \hat{A}^{i_\nu} \cdot h_s$ et l'implication désirée puisque la propriété est vérifiée pour $\nu = 0$. □

Remarque 1. Le corollaire explique pourquoi dans le théorème 1 nous avons préféré la structure de $C^\omega(Q)$-algèbre de Lie à celle de $C^\omega(Q)$-module qui aurait suffi.

Remarque 2. En prenant les distributions associées aux algèbres de découplage on retrouve les résultats de Hirschorn [20] et de Isidori, Krener, Gori-Giorgi et Monaco [22]. On peut se reporter aux articles d'Isidori [21], d'Hirschorn [20] ou à [6] pour comprendre qu'il s'agit bien d'une extension du linéaire au non linéaire.

3. Nombres caractéristiques. Le découplage d'un système linéaire peut être caractérisé par des considérations sur l'ordre de ses zéros infinis (cf. Descusse et Dion [10]). De même, l'étude du découplage du système non linéaire de type (I) nécessite tout d'abord le calcul de ce que nous nommons ses *nombres caractéristiques* φ_s $(s = 1, \cdots, r)$.

On note $\mathfrak{L}_f$ la $C^\omega(Q)$-algèbre de Lie des champs de vecteurs associés à la fonction f, élément de $C^\omega(Q)$. Elle est définie par:

$$C \in \mathfrak{L}_f \quad \text{ssi} \quad C \cdot f \equiv 0 \quad \text{avec} f \in C^\omega(Q).$$

Si $L^m_{A^0} h_s$ représente l'itérée m-ième de la dérivée de Lie L_{A^0}, avec $L^0_{A^0} = Id$, φ_s est le nombre entier tel que:

$$\forall l \, A^l \in \bigcap_{0 \leqq m < \varphi_s} \mathfrak{L}_{L^m_{A^0} h_s} \quad \text{et} \quad \exists l \, A^l \cdot L^{\varphi_s}_{A^0} h_s \not\equiv 0$$

où $l = 1, \cdots, d$; $s = 1, \cdots, r$.

Exemple. On prend le système de type (I), défini sur $\mathbb{R}^N$ par:

$$\dot{q} = A^0 + \mathbf{u} A^1, \qquad y = q_1 \tag{7}$$

avec

$$A^0 = q_2 \frac{\partial}{\partial q_1} + \cdots + q_{n+1} \frac{\partial}{\partial q_n} \quad \text{où } 1 \leqq n \leqq N-1,$$

$$A^1 = \frac{\partial}{\partial q_{n+1}}.$$

On a $A^1 \cdot L^m_{A^0} q_1 \equiv 0$ pour $0 \leqq m \leqq n-1$ et $A^1 \cdot L^n_{A^0} q_1 = 1 \not\equiv 0$.

Ainsi, le système (7) a pour nombre caractéristique:

$$\varphi_1 = n.$$

Les nombres caractéristiques jouent un rôle fondamental dans l'étude des systèmes non linéaires de type (I). On les retrouve dans tous les travaux de découplage comme par exemple dans Sinha [34] et Isidori et al. [22]. Ainsi, en vue d'une commande non interactive (partition scalaire de l'entrée et de la sortie, une entrée agissant sur une seule sortie), Porter [30] définit un *index*, noté ici od_s, d'une sortie y_s d'un système de type (I). Cet ordre différentiel donne l'ordre de dérivation minimal de y_s, permettant d'exprimer $y_s^{(od_s)}$ en fonction de l'entrée $\mathbf{u}$ et des dérivées de y_s d'ordre inférieur à od_s. De même, Tokumaru et al. [37] définissent *l'ordre relatif*, noté ici or_s, et qui est utilisé aussi par Hirschorn [19] dans l'étude de l'inversion des systèmes. Le lien entre les quatre nombres φ_s, i_s, od_s et or_s est donné par:

$$i_s = od_s = or_s = \varphi_s + 1.$$

Dans l'étude de l'inversion des systèmes non linéaires, le théorème I.3 de Rebhuhn [31] permet d'énoncer la propriété importante suivante:

PROPOSITION 3[3]. *Les nombres caractéristiques d'un système de type* (I) *sont toujours majorés strictement par la dimension de la variété d'état quand ils sont définis.*

[3] La démonstration simple et directe qui suit est due à une correspondance fructueuse avec Dominique Cerveau.

Preuve. 1. Il est immédiat que si $\boldsymbol{R}$ est un anneau intègre, le $\boldsymbol{R}$-module $\boldsymbol{R}^k$ ($k \in \mathbb{N}^*$) a la propriété que toute suite de $k+1$ éléments est liée. On peut appliquer cela à l'anneau des germes de fonctions analytiques en un point de la variété Q.

2. On va utiliser le résultat précédent pour montrer que pour chaque s ($s = 1, \cdots, r$) les champs des vecteurs A^{l_0}, $(\text{ad } A^0)(A^{l_0}), \cdots, (\text{ad } A^0)^{\varphi_s}(A^{l_0})$, avec A^{l_0} tel que $A^{l_0} \cdot L^{\varphi_s}_{A^0} h_s \not\equiv 0$, sont génériquement indépendants.

Si tel n'était pas le cas, il existerait sur un ouvert des fonctions analytiques $a_0, \cdots, a_{\varphi_s}$ telles que:

$$a_0 A^{l_0} + a_1(\text{ad } A^0)(A^{l_0}) + \cdots + a_{\varphi_s}(\text{ad } A^0)^{\varphi_s}(A^{l_0}) \equiv 0 \text{ les } a_i \text{ n'étant pas tous nuls.}$$

Soit n le plus grand indice tel que $a_i \not\equiv 0$; alors:

$$a_0 A^{l_0} + \cdots + a_n(\text{ad } A^0)^n(A^{l_0}) \equiv 0 \quad \text{et} \quad n \leqq \varphi_s.$$

On applique ce champ de vecteurs à la fonction $L^{\varphi_s - n}_{A^0} h_s$. Or, il est facilement démontrable par récurrence que $(\text{ad } A^0)^n(C)$ est une combinaison linéaire réelle d'éléments de type

$$(A^0)^\nu C (A^0)^\mu \quad \text{où } \nu + \mu = n \quad (\nu \text{ et } \mu \in \mathbb{N})$$

avec un seul terme $C(A^0)^n$.

De ce fait et en tenant compte de la définition de φ_s, on aurait $a_n A^{l_0} \cdot L^{\varphi_s}_{A^0} h_s \equiv 0$, ce qui est exclu. □

Cette propriété permet en particulier de prendre $N-1$ comme borne universelle dans le calcul algorithmique du découplage présenté dans [7].

Si s est tel que φ_s ne soit pas défini, la série génératrice du système (I) restreinte à la sortie y_s ne contient que des termes associés aux mots de $\mathbf{c^*pa^*} \cup \mathbf{a^*}$.

Les résultats qui suivent justifient le nom de nombres caractéristiques que nous avons choisi:

LEMME 1. *Pour un système bouclé de type* (II) *on a*:

(i) *Si φ_s est défini, alors pour tout $i = 1, \cdots, n$:*

$$\hat{A}^i \cdot L^m_{\hat{A}^0} h_s \equiv 0 \quad \textit{avec } 0 \leqq m < \varphi_s \textit{ soit } \hat{A}^i \in \bigcap_{0 \leqq m < \varphi_s} \mathfrak{L}_{L^m_{A^0} h_s}$$

et

$$L^m_{\hat{A}^0} h_s = L^m_{A^0} h_s \quad \textit{avec } 0 \leqq m \leqq \varphi_s,$$

quel que soit le bouclage de type (∗) *utilisé.*

(ii) *Si de plus, la matrice $\Phi = (\beta^i_l)$ ($l = 1, \cdots, d$, $i = 1, \cdots, n$) est inversible, alors φ_s est le nombre caractéristique de la sortie y_s du système bouclé* (II).

(iii) *Si φ_s n'est pas défini alors pour tout entier m*

$$\hat{A}^i \cdot L^m_{\hat{A}^0} h_s \equiv 0 \quad \textit{et} \quad L^m_{\hat{A}^0} h_s = L^m_{A^0} h_s \qquad (i = 1, \cdots, n).$$

Preuve. (i) 1. On montre d'abord par récurrence que

$$L^m_{\hat{A}^0} h_s = L^m_{A^0} h_s \quad \text{pour } 0 \leqq m \leqq \varphi_s.$$

C'est vrai pour $m = 0$ car $L^0_{\hat{A}^0} h_s = L^0_{A^0} h_s = h_s$. Si $L^m_{\hat{A}^0} h_s = L^m_{A^0} h_s$ avec $m < \varphi_s$ on a

$$L^{m+1}_{\hat{A}^0} h_s = \hat{A}^0 \cdot L^m_{\hat{A}^0} h_s = \hat{A}^0 \cdot L^m_{A^0} h_s = L^{m+1}_{A^0} h_s + \sum_{l=1}^{d} \alpha_l A^l \cdot L^m_{A^0} h_s$$

soit

$$L^{m+1}_{\hat{A}^0} h_s = L^{m+1}_{A^0} h_s$$

d'après la définition de φ_s.

2. D'après le paragraphe 1, on a: $\hat{A}^i \cdot L^m_{\hat{A}^0} h_s = \hat{A}^i \cdot L^m_{A^0} h_s$ pour $i = 1, \cdots, n$ et $0 \leqq m < \varphi_s$.

Soit

$$\hat{A}^i \cdot L^m_{\hat{A}^0} h_s = \sum_{l=1}^{d} \beta^i_l A^l \cdot L^m_{A^0} h_s \equiv 0$$

d'après la définition de φ_s.

(ii) Si $\mathbf{\Phi}$ est inversible et si $\forall i \, \hat{A}^i \cdot L^{\varphi_s}_{\hat{A}^0} h_s \equiv 0$ on a $d = n$ et $\sum_{l=1}^{n} \beta^i_l A^l \cdot L^{\varphi_s}_{A^0} h_s \equiv 0$ pour tout $i = 1, \cdots, n$.

En désignant par $\mathbf{\Omega}_s$ la matrice ligne

$$\mathbf{\Omega}_s = (A^1 \cdot L^{\varphi_s}_{A^0} h_s, \cdots, A^n \cdot L^{\varphi_s}_{A^0} h_s)$$

on peut écrire

$$\mathbf{\Omega}_s \cdot \mathbf{\Phi} = 0$$

d'où $\mathbf{\Omega}_s = 0$ ce qui est contradictoire avec la définition de φ_s.

(iii) Immédiat par récurrence. □

4. La plus grande algèbre de découplage. Les algèbres de découplage permettant par les distributions associées de retrouver localement le découplage structurel (cf. Isidori [21]), il convient de déterminer la plus grande algèbre de découplage $\mathfrak{D}^*$ qui existe au moins localement. Si les nombres caractéristiques φ_s $(s = 1, \cdots, r)$ sont définis, elle peut être donnée, comme on va le voir par:

$$\mathfrak{D}^0 = \bigcap_{s=1}^{r} \mathfrak{L}_s \quad \text{avec } \mathfrak{L}_s = \bigcap_{0 \leqq m \leqq \varphi_s} \mathfrak{L}_{L^m_{A^0} h_s}.$$

PROPOSITION 4. *Si les nombres caractéristiques φ_s $(s = 1, \cdots, r)$ du système* (I) *sont tous définis, on a*:

(a) $\mathfrak{D} \subset \mathfrak{D}^0$, *pour tout algèbre de découplage $\mathfrak{D}$ de la sortie y du système* (II) *par rapport à* **p**.

(b) *Une condition nécessaire pour que $\mathfrak{D}$ soit une algèbre de découplage est que*:

$$\mathfrak{D} \cdot g^i_s \equiv 0, \qquad i = 0, \cdots, n, \quad s = 1, \cdots, r$$

où $g^i_s = \hat{A}^i \cdot L^{\varphi_s}_{A^0} h_s$ avec $\hat{A}^i \in \mathscr{A} = \{\hat{A}^0, \cdots, \hat{A}^n\}$.

(c) *Une condition nécessaire et suffisante pour que $\mathfrak{D}^0$ soit une algèbre de découplage du système* (II) *est qu'il existe un bouclage de type* (*) *tel que*:

$$\mathscr{P} \subset \mathfrak{D}^0 \quad \text{et} \quad \mathfrak{D}^0 \cdot g^i_s \equiv 0, \qquad i = 0, \cdots, n, \quad s = 1, \cdots, r.$$

Preuve. (a) On procède par récurrence et on montre que pour tout $s = 1, \cdots, r$,

$$\mathfrak{D} \subset \mathfrak{L}_s.$$

Si $m = 0$, $\mathfrak{L}_{h_s} = \mathfrak{H}_s$ et on sait d'après le théorème 1 que $\mathfrak{D} \subset \mathfrak{H}_s$.

D'autre part, si $\mathfrak{D} \subset \mathfrak{L}_{L^m_{A^0} h_s}$ avec $m < \varphi_s$, comme $[\hat{A}^0, \mathfrak{D}] \subset \mathfrak{D}$ on a, en utilisant le lemme 1,

$$\mathfrak{D} \cdot L^{m+1}_{A^0} h_s = \mathfrak{D} \cdot \hat{A}^0 \cdot L^m_{A^0} h_s = [\mathfrak{D}, \hat{A}^0] \cdot L^m_{A^0} h_s \equiv 0$$

et ainsi

$$\mathfrak{D} \subset \mathfrak{L}_{L^{m+1}_{A^0} h_s}.$$

(b) Si $\mathfrak{D}$ est une algèbre de découplage par rapport à $\mathbf{p}$ de la sortie y du système non initialisé de type (II), on a bien, $\mathfrak{D} \subset \mathfrak{D}^0$ d'après le paragraphe (a) et pour tout $s = 1, \cdots, r$:

$$[\mathfrak{D}, \hat{A}^i] \cdot L_{A^0}^{\varphi_s} h_s \equiv 0 \equiv \mathfrak{D} \cdot g_s^i \quad \text{pour tout } i = 0, \cdots, n.$$

(c) 1. La condition est nécessaire d'après (b) et le théorème 1.

2. Si $\mathscr{P} \subset \mathfrak{D}^0$ et $\mathfrak{D}^0 \cdot g_s^i \equiv 0$, on va montrer que $\mathfrak{D}^0$ est une algèbre de découplage de la sortie y du système (II), par rapport à $\mathbf{p}$. D'après le paragraphe (a), $\mathfrak{D}^0$ sera la plus grande.

Par définition, $\mathfrak{D}^0 \subset \mathfrak{H}$ et de plus, d'après la définition de φ_s, on a pour tout $s = 1, \cdots, r$:

$$[\hat{A}^i, \mathfrak{D}^0] \cdot L_{A^0}^m h_s = -\mathfrak{D}^0 \cdot \hat{A}^i \cdot L_{A^0}^m h_s = -\delta_{\varphi_s}^m \cdot \mathfrak{D}^0 \cdot g_s^i$$

avec $0 \leqq m \leqq \varphi_s$, δ_k^j le symbole de Kronecker et $i = 0, \cdots, n$.

Ainsi, on a bien

$$[\mathscr{A}, \mathfrak{D}^0] \subset \mathfrak{D}^0 \quad \text{et} \quad \mathscr{P} \subset \mathfrak{D}^0 \subset \mathfrak{H}. \qquad \square$$

Dans le cas où le découplage n'est pas total on peut néanmoins, grâce au découplage fini, préciser le lien entre la perturbation $\mathbf{p}$ et les entrées admissibles.

Si une sortie y_s $(s = 1, \cdots, r)$ du système (I), non initialisé, n'est pas découplée totalement par rapport à $\mathbf{p}$. On considère (cf. fin du chapitre 1) l'ordre de découplage fini maximal δ_s auquel y_s est découplée par rapport à $\mathbf{p}$.

LEMME 2. *Soit un système non initialisé de type* (I) *et* y_s $(s = 1, \cdots, r)$ *une de ses sorties découplée à l'ordre fini maximal* δ_s *par rapport à* $\mathbf{p}$. *On a*

(a) *Si* φ_s *est défini, alors*:

(1) *Soit* $\delta_s \leqq \varphi_s$ *et le système bouclé* (II) *a sa sortie* y_s *découplée à l'ordre* δ_s *quel que soit le bouclage de type* (*) *considéré.*

(2) *Soit* $\delta_s > \varphi_s$ *et le système bouclé* (II) *à sa sortie* y_s *au moins découplée, par rapport à* $\mathbf{p}$, *à un ordre supérieur ou égal à* $\varphi_s + 1$, *quel que soit le bouclage de type* (*) *considéré.*

(3) *Soit* $\delta_s > \varphi_s$ *et une condition nécessaire et suffisante pour que la sortie* y_s *du système* (II) *soit découplée, par rapport à* $\mathbf{p}$, *à un ordre supérieur ou égal à* $\varphi_s + 1 + \mu$ $(\mu \geqq 1)$ *est qu'il existe un bouclage de type* (*) *tel que*:

$$W \cdot g_s^i \equiv 0 \quad \textit{pour tout } \mathbf{w} \in (\mathbf{c}^*\mathbf{p}\mathbf{c}^*)_\mu \textit{ avec } \mu \geqq 1,$$

où

$$g_s^i = \hat{A}^i \cdot L_{A^0}^{\varphi_s} h_s \qquad (i = 0, \cdots, n).$$

(b) *Si* φ_s *n'est pas défini, alors quel que soit le bouclage de type* (*) *utilisé, le système bouclé* (II) *a sa sortie* y_s *découplée à l'ordre maximal* δ_s *par rapport à* $\mathbf{p}$ *et* $\delta_s < N$.

Preuve. (a) φ_s étant défini, on sait d'après le lemme 1 que:

$$\hat{A}^i \cdot L_{\hat{A}^0}^m h_s \equiv 0 \quad \text{si } 0 \leqq m < \varphi_s \text{ avec } i = 1, \cdots, n$$

et

$$L_{\hat{A}^0}^m h_s = L_{A^0}^m h_s \quad \text{si } 0 \leqq m \leqq \varphi_s.$$

1. Si $\delta_s \leqq \varphi_s$, d'après les considérations ci-dessus, on voit que quel que soit le bouclage de type (*), le système bouclé (II) a sa sortie y_s découplée par rapport à $\mathbf{p}$ à l'ordre maximal δ_s puisque $\mathscr{P} \cdot L_{\hat{A}^0}^{\delta_s} h_s = \mathscr{P} \cdot L_{A^0}^{\delta_s} h_s$ n'est pas nul.

2. Si $\delta_s > \varphi_s$, $\mathscr{P} \cdot L_{A^0}^m h_s = \mathscr{P} \cdot L_{A^0}^m h_s \equiv 0$ pour $0 \leqq m \leqq \varphi_s$ et comme $\hat{A}^i \cdot L_{A^0}^m h_s \equiv 0$ pour $0 \leqq m < \varphi_s$ $(i = 1, \cdots, n)$ on a $W \cdot \hat{A}^i \cdot L_{A^0}^m h_s \equiv 0$ pour $\mathbf{w} \in \mathbf{c}^*\mathbf{pc}^*$.

Ainsi, quel que soit le bouclage de type (*) la sortie y_s du système (II) est découplée, par rapport à $\mathbf{p}$, à un ordre fini $\mu \geqq \varphi_s + 1$.

3. D'après le paragraphe 2 seuls sont à considérer les mots $\mathbf{w} = \mathbf{w}_1 \cdot \mathbf{w}_2$ avec $\mathbf{w}_1 \in (\mathbf{c}^*\mathbf{pc}^*)_\mu$ et $\mathbf{w}_2 \in \mathbf{a}^*$, $\mathbf{w}_2$ étant de longueur $\varphi_s + 1$. La conclusion vient alors aisément du Lemme 1 et de la définition des fonctions g_s^i.

(b) Evident d'après le lemme 1 et puisque $\mathscr{P} \cdot L_{\hat{A}^0}^{\delta_s} h_s = \mathscr{P} \cdot L_{A^0}^{\delta_s} h_s$ n'est pas nul.

D'autre, part, l'inégalité $\delta_s < N$ se démontre comme dans la proposition 3 car: $\delta_s = 0$ ou alors, $\delta_s \geqq 1$ et $\delta_s = \sup \{n | \forall m (m < n \Rightarrow \forall j P^j \cdot L_{A^0}^m h_s \equiv 0);$ $j = 1, \cdots, k\}$. □

Les résultats de ce chapitre vont nous permettre de proposer maintenant une méthode de découplage.

5. Méthode de découplage. Les égalités $\hat{A}^i \cdot L_{A^0}^{\varphi_s} h_s = g_s^i$ $(i = 0, \cdots, n,\ s = 1, \cdots, r)$ avec $\hat{A}^i \in \mathscr{A}$ sont équivalentes au système d'équations suivant:

$$(**) \qquad \mathbf{\Omega} \cdot \mathbf{\Delta} = \mathbf{\Gamma}^0; \qquad \mathbf{\Omega} \cdot \mathbf{\Phi} = \mathbf{\Gamma}^1$$

où $\mathbf{\Omega}, \mathbf{\Delta}, \mathbf{\Phi}, \mathbf{\Gamma}^0, \mathbf{\Gamma}^1$ sont des matrices respectivement éléments de ${}^r(C^\omega(Q))^d$, ${}^d(C^\omega(Q))^1$, ${}^d(C^\omega(Q))^n$, ${}^r(C^\omega(Q))^1$ et ${}^r(C^\omega(Q))^n$ avec

$$\mathbf{\Omega}_s^l = A^l \cdot L_{A^0}^{\varphi_s} h_s, \qquad \mathbf{\Delta}_l = \alpha_l, \qquad \mathbf{\Phi}_l^i = \beta_l^i,$$

$$\mathbf{\Gamma}_s^0 = g_s^0 - L_{A^0}^{\varphi_s + 1} h_s \quad \text{et} \quad \mathbf{\Gamma}_s^{1i} = g_s^i$$

avec $s = 1, \cdots, r$ et $i = 1, \cdots, n$.

Il s'agit donc de "forcer" le système (**) à avoir des solutions avec les g_s^i tels que $\mathfrak{D} \cdot g_s^i \equiv 0$ $(i = 0, \cdots, n;\ s = 1, \cdots, r)$ où $\mathfrak{D}$ est une algèbre de découplage.

Les bouclages de type (*) que nous rappelons:

$$(*) \qquad \begin{aligned} \mathbf{u}_l(t) &= \alpha_l(q) + \sum_{i=1}^{n} \beta_l^i(q) \mathbf{a}_i(t) \quad \text{avec } l = 1, \cdots, d, \\ \mathbf{p}_j(t) &= \mathbf{p}_j(t) \quad \text{avec } j = 1, \cdots, k \end{aligned}$$

assurant alors un découplage du système (II) avec l'algèbre de Lie $\mathfrak{D}_h^{\mathbf{a}}$ correspondante.

Notons que la matrice $\mathbf{\Omega}$ n'a pas en général un rang constant sur toute la variété Q et les solutions ont un caractère local (cf. Hirschorn [20]).

THEOREM 2. *Soit un système de type* (I) *pour lequel on souhaite voir sa sortie y indépendante de la perturbation* $\mathbf{p}$ *grâce à l'utilisation des bouclages de type* (*). *Alors*:

(A) *Soit s tel que le nombre caractéristique* φ_s *ne soit pas défini, on a*:

1. *Si* δ_s $(\delta_s < N)$ *est l'ordre maximum de découplage fini de la sortie* y_s *du système* (I) *par rapport à* $\mathbf{p}$, *la sortie* y_s *du système bouclé* (II) *reste découplée à l'ordre maximum* δ_s *par rapport à* $\mathbf{p}$ *et ceci quel que soit le bouclage de type* (*) *utilisé.*

2. *Si la sortie* y_s *du système* (I) *est découplée totalement par rapport à* $\mathbf{p}$, *il en est de même dans le système bouclé, quel que soit le bouclage et la plus grande algèbre de découplage est* $\mathfrak{D}_{h_s}^{\mathbf{a}}$, *où ici*:

$$C \in \mathfrak{D}_{h_s}^{\mathbf{a}} \quad ssi \quad \forall m \geqq 0 \quad C \cdot L_{A^0}^m h_s \equiv 0.$$

(B) *Soit s tel que le nombre caractéristique* φ_s *soit défini mais avec* $\mathscr{P} \not\subset \mathfrak{L}_s$. *La sortie* y_s *est alors découplée à un ordre maximal* $\delta_s \leqq \varphi_s$ *indépendamment du bouclage choisi.*

(C) *Les nombres caractéristiques* φ_s $(s = 1, \cdots, r)$ *sont définis et* $\mathscr{P} \subset \mathfrak{D}^0$.

1. Chaque sortie y_s du système (I) est alors découplée à l'ordre φ_s+1 par rapport à $\mathbf{p}$ et chaque bouclage de type (*) assure un découplage de y_s au moins à l'ordre φ_s+1 par rapport à $\mathbf{p}$.

2. Les relations de compatibilité que doivent satisfaire les g_s^i sont mises en évidence en abordant la résolution du système (**) par mise sous forme triangulaire de $\mathbf{\Omega}$ en se plaçant dans l'anneau intégre $C^\omega(Q)$.

Les bouclages de type (*) cherchés et les fonctions g_s^i correspondantes doivent aussi vérifier:

$$\mathscr{P}\cdot g_s^i=0, \qquad i=0,\cdots,n, \quad s=1,\cdots,r,$$

et

$$\mathscr{P}\cdot(\mathbf{\Omega}\cdot\mathbf{\Delta})=\mathbf{\Gamma}^2, \qquad \mathscr{P}\cdot(\mathbf{\Omega}\cdot\mathbf{\Phi})=0$$

avec Γ^2 une matrice de ${}^r(C^\omega(Q))^1$ définie par

$$\mathbf{\Gamma}_s^2=-\mathscr{P}\cdot L_{A^0}^{\varphi_s+1}h_s, \qquad s=1,\cdots,r.$$

3. Si on peut résoudre le système (**) en prenant pour les g_s^i $(i=0,\cdots,n;\ s=1,\cdots,r)$ des fonctions de la $\mathbb{R}$-algèbre $\mathscr{G}$ des fonctions analytiques sur Q, définie comme suit:

$$g\in\mathscr{G} \quad \text{ssi} \quad \exists\tilde{g}\in C^\omega(\mathbb{R}^\eta) \quad \text{avec } \eta=r+\sum_{s=1}^{r}\varphi_s,$$

et

$$g(q)=\tilde{g}(h_1(q),\cdots,L_{A^0}^{\varphi_1}h_1(q),\cdots,h_r(q),\cdots,L_{A^0}^{\varphi_r}h_r(q)),$$

alors toute solution donne un bouclage permettant de découpler totalement y par rapport à $\mathbf{p}$, avec pour algèbre de découplage $\mathfrak{D}^0$, la plus grande des algèbres de découplage possibles.

Preuve. Immédiat d'après les résultats précédents. □

Remarque 1. En prenant g_s^0 nul et g_s^i $(i=1,\cdots,n)$ constant, on a bien $\mathfrak{D}^0(g_s^i)=0$ $(i=0,\cdots,n;\ s=1,\cdots,r)$ et on retrouve les résultats d'Isidori et al. [22]. Cependant, en prenant l'exemple donné dans [4], on voit que cette classe de bouclages est par trop restreinte. En effet, on considère le système défini sur $\mathbb{R}^2$ par:

$$\begin{aligned}\dot{q}_1(t)&=(\mathbf{u}(t)+\mathbf{p}(t))q_1(t),\\ \dot{q}_2(t)&=-\mathbf{p}(t)q_2(t),\\ y(t)&=q_1q_2.\end{aligned}\tag{8}$$

On a

$$A^0=0, \quad A^1=q_1\frac{\partial}{\partial q_1}, \quad P^1=q_1\frac{\partial}{\partial q_1}-q_2\frac{\partial}{\partial q_2}, \quad h(q)=q_1q_2.$$

Une intégration élémentaire montre que le système est totalement découplé par rapport à $\mathbf{p}$ et ainsi la classe des bouclages $\mathbf{u}=\alpha+\beta\mathbf{a}$ donnant le découplage de y par rapport à $\mathbf{p}$ doit contenir le bouclage $\mathbf{u}=\mathbf{a}$!

Le nombre caractéristique φ est égal à 0 car $A^1\cdot h\not\equiv 0$ et ainsi $\mathfrak{D}^0=\mathfrak{H}$. On a $P^1\in\mathfrak{H}\cdot\mathfrak{H}$ est engendré par P^1 et les fonctions g^i $(i=0,1)$ vérifiant $\mathfrak{H}(g^i)\equiv 0$ sont les fonctions analytiques du genre $\tilde{g}\circ h$ avec $\tilde{g}\in C^\omega(\mathbb{R})$.

Les bouclages donnant un découplage sont donc solution de $\mathbf{\Omega}\cdot\mathbf{\Delta}=(g^0)$ et $\mathbf{\Omega}\cdot\mathbf{\Phi}=(g^1)$ avec $\mathbf{\Omega}=(h)$ soit $\alpha=g^0/h$ et $\beta=g^1/h$.

On voit ainsi que c'est le bouclage déterminé par $g^0 = 0$ et $g^1 = h$ qui redonne le système initial.

Remarque 2. Dans le cas de perturbations mesurables on peut compléter les lois de bouclages de type (*) en posant

$$\mathbf{u}_l = \alpha_l + \sum_{i=1}^{n} \beta_l^i \mathbf{a}_i + \sum_{j=1}^{k} \gamma_l^j \mathbf{p}_j \qquad (l = 1, \cdots, d).$$

Le système (I) bouclé s'écrit alors:

$$\begin{aligned} \dot{q} &= \hat{A}^0 + \sum_{i=1}^{n} \mathbf{a}_i \hat{A}^i + \sum_{j=1}^{k} \mathbf{p}_j \hat{P}^j, \\ y &= h(q), \end{aligned} \tag{9}$$

avec

$$\hat{P}^j = P^j + \sum_{l=1}^{d} \gamma_l^j A^l \qquad (j = 1, \cdots, k).$$

Les fonctions γ_l^j, éléments de $C^\omega(Q)$, sont destinées à permettre aux perturbations $\hat{P}^j$ $(j = 1, \cdots, k)$ de vérifier la condition $\hat{P}^j \in \mathfrak{D}^0$.

On peut voir en [6] une application de ceci à l'étude du rejet de perturbations dans un modèle non linéaire de colonne à distiller proposé par Gauthier et al. [16].

6. Immersion et découplage. Pour des raisons supplémentaires de calcul par l'algorithme introduit en [7] et complété par l'algorithme rapide de Kasinski et Lévine [23] explicité en [17], nous nous intéressons plus spécialement aux bouclages liés à l'algèbre $\mathfrak{D}^0$.

Nous allons maintenant indiquer ce que devient le système (II) quand $\mathfrak{D}^0$ est une algèbre de découplage. Nous donnons le dernier outil qui nous sera nécessaire: *l'immersion* (cf. Fliess [12] et Fliess et Kupka [14]). L'immersion a pour but de préciser la notion de même comportement entrée-sortie, notion qui ne vas pas de soi en non linéaire. On peut déjà voir à travers divers travaux récents, comme la linéarisation (cf. Claude, Fliess et Isidori [8] en temps continu et Monaco et Normand-Cyrot [26] en temps discret) et le préprocesseur (Monaco [25]), que l'immersion apparaît comme un concept bien utile en théorie des systèmes.

Soient Σ et Σ' deux systèmes de type (II) ayant comme espaces d'état respectifs Q et Q' et comme fonctions de sorties h et h'.

DÉFINITION. Une immersion de Σ dans Σ' est une application analytique $\tau: Q \to Q'$ telle que Σ et Σ', respectivement initialisés en q et $q' = \tau(q)$, aient même série génératrice et que $h(q_1) \neq h(q_2) \Rightarrow h'(\tau(q_1)) \neq h'(\tau(q_2))$.

Le théorème suivant précise et étend les résultats de Porter [30] et généralise les résultats sur l'immersion par bouclage dans un système linéaire (cf. [8]).

THÉORÈME 3. *Si les nombres caractéristiques* φ_s $(s = 1, \cdots, r)$ *du système* (I) *existent, si* $\mathscr{P} \subset \mathfrak{D}^0$ *et si le système* (**) *admet des solutions avec* Φ *inversible et les fonctions* g_s^i $(i = 0, \cdots, n)$ *éléments de* $\mathscr{G}$, *alors les bouclages de type* (*) *correspondants permettent, par l'application*

$$\tau: Q \to \mathbb{R}^\eta \quad \text{avec } \eta = r + \sum_{s=1}^{r} \varphi_s$$

et

$$\tau(q) = (h_1(q), \cdots, L_{A^0}^{\varphi_1} h_1(q), \cdots, h_r(q), \cdots, L_{A^0}^{\varphi_r} h_r(q)),$$

d'immerger le système bouclé (II) *dans le système* Λ *défini par*:

$$\Lambda \qquad s=1,\cdots,r \quad \begin{cases} \dot{x}_s^m = x_s^{m+1}, & 0 \leqq m < \varphi_s, \\ \dot{x}_s^{\varphi_s} = \tilde{g}_s^0 + \sum_{i=1}^{n} \tilde{g}_s^i \mathbf{a}_i, \\ z_s = x_s^0, \end{cases}$$

où $\tilde{g}_s^i$ *est une fonction analytique sur* $\mathbb{R}^\eta$ *associée à* g_s^i *par* $g_s^i = \tilde{g}_s^i \circ \tau$.

Preuve. L'application τ est bien analytique et nous allons montrer que la série génératrice du système bouclé (II), initialisé en $q(0)$, est identique à celle du système Λ initialisé en $\tau(q(0))$.

Le système (II) ayant sa sortie y découplée par rapport à $\mathbf{p}$, chaque série génératrice $\mathfrak{h}_s$ $(s=1,\cdots,r)$ est telle que

$$\operatorname{supp} \mathfrak{h}_s \subset \mathbf{a}^*$$

avec $\mathbf{a}^*$, le monoïde engendré par l'alphabet $\mathbf{a}=\{\mathbf{a}_0, \mathbf{a}_1, \cdots, \mathbf{a}_n\}$. D'autre part, le système Λ peut s'écrire:

$$\dot{x} = B^0 + \sum_{i=1}^{n} \mathbf{a}_i B^i, \qquad z = f(x) \tag{10}$$

avec $z=(z_1,\cdots,z_r)$; $f=(f_1,\cdots,f_r)$ et $z_s = f_s(x) = x_s^0$ $(s=1,\cdots,r)$;

où

$$B^0 = \sum_{s=1}^{r} \left(\sum_{m=0}^{\varphi_s - 1} x_s^{m+1} \frac{\partial}{\partial x_s^m} + \tilde{g}_s^0 \frac{\partial}{\partial x_s^{\varphi_s}} \right)$$

et

$$B^i = \sum_{s=1}^{r} \tilde{g}_s^i \frac{\partial}{\partial x_s^{\varphi_s}}$$

(étant entendu que le terme $\sum_{m=0}^{\varphi_s - 1}$ disparaît quand $\varphi_s = 0$).

Un calcul élémentaire montre que $L_{B^0}^m f_s = x_s^m$ pour $0 \leqq m \leqq \varphi_s$ et que $B^i \cdot L_{B^0}^{\varphi_s} f_s = \tilde{g}_s^i$ avec $i=0,\cdots,n$ et $s=1,\cdots,r$.

Ainsi, comme Φ est inversible, les nombres φ_s $(s=1,\cdots,r)$ sont aussi les nombres caractéristiques du système Λ.

De plus, si

$$\tilde{\mathfrak{L}}_s = \bigcap_{0 \leqq m \leqq \varphi_s} \mathfrak{L}_{L_{B^0}^m f_s}$$

on a

$$\tilde{\mathfrak{D}}^0 = \bigcap_{s=1}^{r} \tilde{\mathfrak{L}}_s = \{0\}.$$

Comme $g_s^i \in \mathcal{G}$ $(s=1,\cdots,r,\ i=0,\cdots,n)$, il est clair que $\mathcal{G}$ est stable pour les dérivations de $\mathcal{A}=\{\hat{A}^0,\cdots,\hat{A}^n\}$.

De plus, on a pour $i=0,\cdots,n$:

$$\hat{A}^i \cdot L_{A^0}^m h_s = B^i \cdot x_s^m \circ \tau \quad \text{avec } 0 \leqq m \leqq \varphi_s.$$

Ainsi, si $g \in \mathcal{G}$ on a pour tout $i=0,\cdots,n$

$$\hat{A}^i \cdot g = B^i \cdot \tilde{g} \circ \tau$$

ou, encore, quel que soit $q(0)$ et pour tout $i=0,\cdots,n$

$$\hat{A}^i \cdot g|q(0) = B^i \cdot \tilde{g}|\tau(q(0)).$$

On trouve donc bien pour tout $s=1,\cdots,r$ et pour toute initialisation $q(0)$:

$$h_s|q(0) = f_s|\tau(q(0))$$

$$\text{et si } w \in \mathbf{a}^*,\ W\cdot h_s = W\cdot f_s \circ \tau$$

le premier membre de l'égalité étant relatif au système (II) et le second au système Λ.

Ainsi on a, quel que soit $q(0)$

$$\mathfrak{h}_s|q(0) = \mathfrak{f}_s|\tau(q(0)) \qquad (s=1,\cdots,r)$$

($\mathfrak{f}_s$ désignant la série génératrice du système Λ de sortie z_s), et τ est bien une immersion du système (II) dans le système Λ. □

Cette méthode a été utilisée aussi bien pour le calcul de lois de commande pour le pilotage d'un hélicoptère[4] qu'en neuro-endocrinologie [9].

BIBLIOGRAPHIE

[1] G. BASILE AND G. MARRO, *A state space approach to non-interacting controls*, Ric. Autom., 1 (1970), pp. 68-77.

[2] F. BOURNONVILLE, Thèse de docteur-ingénieur, Université de Nantes, France, 1984.

[3] C. I. BYRNES AND A. J. KRENER, *On the existence of globally (f, g)-invariant distributions*, in Differential Geometric Control Theory, R. W. Brockett, R. S. Millman and H. J. Sussmann, eds., Birkhaüser, Boston, 1983, pp. 209-225.

[4] D. CLAUDE, *Découplage des systèmes non linéaires analytiques ou rejet des perturbations*, C.R. Acad. Sci. Paris, 292, série I (1981), pp. 59-62.

[5] ———, *Decoupling of nonlinear systems*, Syst. Control Lett., 1 (1982), pp. 242-248.

[6] ———, *Découplage des systèmes: du linéaire au non linéaire*, dans Outils et modèles mathématiques pour l'automatique, l'analyse de systèmes et le traitement du signal, vol. 3, I. D. Landau, ed., CNRS, Paris, 1983, pp. 533-555.

[7] D. CLAUDE AND P. DUFRESNE, *An application of Macsyma to nonlinear decoupling*, Lectures Notes in Computer Sciences 144, Springer-Verlag, Berlin, 1982, pp. 294-301.

[8] D. CLAUDE, M. FLIESS AND A. ISIDORI, *Immersion, directe et par bouclage, d'un système non linéaire dans un linéaire*, C.R. Acad. Sci. Paris, 296, série I (1983), pp. 237-240.

[9] D. CLAUDE AND E. BERNARD-WEIL, *Découplage et immersion d'un modèle neuro-endocrinien*, C.R. Acad. Sci. Paris, 299, série I (1984), pp. 129-132.

[10] J. DESCUSSE AND J. M. DION, *On the structure at infinity of linear square decoupled systems*, IEEE Trans. Automat. Contr., 27 (1982), pp. 971-974.

[11] M. FLIESS, *Fonctionnelles causales non linéaires et indéterminées non commutatives*, Bull. Soc. Math. France, 109 (1981), pp. 3-40.

[12] ———, *Finite-dimensional observations-spaces for nonlinear systems*, dans Feedback Control of Linear and Nonlinear Systems, D. Hinrichsen and A. Isidori, eds., Lecture Notes in Control and Information Science 39, Springer-Verlag, Berlin, 1982, pp. 73-77.

[13] ———, *Réalisation locale des systèmes non linéaires, algèbres de Lie filtrées transitives et séries génératrices non commutatives*, Invent. Math., 71 (1983), pp. 521-537.

[14] M. FLIESS AND I. KUPKA, *A finiteness criterion for nonlinear input-output differential systems*, this Journal, 21 (1983), pp. 721-728.

[15] E. FREUND, *The structure of decoupled nonlinear systems*, Int. J. Control, 21 (1975), pp. 443-450.

[16] J. P. GAUTHIER, G. BORNARD, S. BACHA AND M. IDIR, *Rejet des perturbations pour un modèle non-linéaire de colonne à distiller*, dans Outils et modèles mathématiques pour l'automatique, l'analyse de systèmes et le traitement du signal, vol. 3, I. D. Landau, ed., CNRS, Paris, 1983, pp. 659-673.

[4] Contrat D.R.E.T. no. 81492.

[17] F. GEROMEL, J. LEVINE AND P. WILLIS, *A fast algorithm for systems decoupling using formal calculus*, Lectures Notes in Control and Information Sciences 63, Springer-Verlag, Berlin, 1984, pp. 378-390.

[18] R. HERMANN, *On the accessibility problem in control theory*, in International Symposium on Nonlinear Differential Equations and Nonlinear Mechanics, Academic Press, New York, 1963, pp. 325-332.

[19] R. M. HIRSCHORN, *Invertibility of nonlinear control systems*, this Journal, 17 (1979), pp. 289-297.

[20] ———, *$(A-B)$-invariant distributions and disturbance decoupling of nonlinear systems*, this Journal, 19 (1981), pp. 1-19.

[21] A. ISIDORI, *Sur la théorie structurelle et le problème de la réjection des perturbations dans les systèmes non linéaires*, dans Outils et modèles mathématiques pour l'automatique, l'analyse de systèmes et le traitement du signal, Vol. 1, I. D. Landau, ed., CNRS, Paris, 1981, pp. 245-294.

[22] A. ISIDORI, A. J. KRENER, C. GORI-GIORGI AND S. MONACO, *Nonlinear decoupling via feedback: a differential geometric approach*, IEEE Trans. Automat. Contr., 26 (1981), pp. 331-345.

[23] A. KASINSKI AND J. LEVINE, *A fast graph theoretic algorithm for the feedback decoupling problem of nonlinear systems*, 8th MNTS Conference, Beersheva, 1983, pp. 550-562.

[24] C. LOBRY, *Contrôlabilité des systèmes non linéaires*, this Journal, 8 (1970), pp. 573-605.

[25] S. MONACO, *On the reproducibility of a feedback by means of linear and bilinear preprocessors*, rapport 82-16, Institut d'Automatique, Université de Rome, 1982.

[26] S. MONACO AND D. NORMAND-CYROT, *The immersion under feedback of a multidimensional discrete-time non-linear system into a linear system*, Int. J. Control, 38 (1983), pp. 245-261.

[27] H. NIJMEIJER, *The triangular decoupling problem for nonlinear control systems*, Nonlin. Anal. Th. Meth. Applic., 8 (1984), pp. 273-279.

[28] H. NIJMEIJER AND A. VAN DER SCHAFT, *Controlled invariance for nonlinear system*, IEEE Trans. Automat. Contr., 27 (1982), pp. 904-914.

[29] ———, *Controlled invariance for nonlinear systems: two worked examples*, IEEE Trans. Automat. Contr., 29 (1984), pp. 361-364.

[30] W. A. PORTER, *Diagonalization and inverses for nonlinear system*, Int. J. Control, 11 (1970), pp. 67-76.

[31] D. REBHUHN, *Invertibility of C^∞ multivariable input-output systems*, IEEE Trans. Automat. Contr., 25 (1980), pp. 207-212.

[32] W. RESPONDEK, *On decomposition of nonlinear control systems*, Syst. Control Lett., 1 (1982), pp. 301-308.

[33] S. N. SINGH AND A. A. SCHY, *Output feedback nonlinear decoupled control synthesis and observer design for manoeuvring aircraft*, Int. J. Control, 31 (1980), pp. 781-806.

[34] P. K. SINHA, *State feedback decoupling of nonlinear systems*, IEEE Trans. Automat. Contr., 22 (1977), pp. 487-489.

[35] H. J. SUSSMANN, *Lie brackets, real analyticity and geometric control*, in Differential Geometric Control Theory, R. W. Brockett, R. S. Millman and H. J. Sussmann, eds., Birkhaüser, Boston, 1983, pp. 1-116.

[36] T. TAKAMATSU, I. HASHIMOTO AND Y. NAKAI, *A geometric approach to multivariable control system design of a distillation column*, Automatica, 15 (1979), pp. 387-402.

[37] H. TOKUMARU AND Z. IWAI, *Non-interacting control of non-linear multivariable systems*, Int. J. Control, 16 (1972), pp. 945-958.

[38] W. M. WONHAM, *Linear Multivariable Control: A Geometric Approach*, 2nd ed., Springer-Verlag, New York, 1977.

[39] W. M. WONHAM AND A. S. MORSE, *Decoupling and pole assignment in linear multivariable system, a geometric approach*, this Journal, 8 (1970), pp. 1-18.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 3, May 1986

016

ON AN ALGORITHM FOR OPTIMAL CONTROL USING PONTRYAGIN'S MAXIMUM PRINCIPLE*

JOSEPH FRÉDÉRIC BONNANS†

Abstract. Y. Sakawa and Y. Shindo recently proposed an algorithm to solve open-loop optimal control problems, using Pontryagin's maximum principle. It is established here that, under some hypothesis, the algorithm is well-defined and globally converges in some weak sense. When the stepsize of the algorithm tends to zero, the displacements are equivalent to those of a gradient with projection method. If the control enters in a linear way in the state equation and in a quadratic way in the criterion, the algorithm can be interpreted as a gradient plus projection method in some new metric.

Key words. optimal control, optimization algorithms, Pontryagin's principle

AMS(MOS) subject classifications. 49D99, 93C10

1. Introduction. We are concerned with the open-loop optimal control of a system governed by an ordinary differential equation. We assume that there are no state constraints and that the control constraints are local with respect to the time. Y. Sakawa and Y. Shindo [4] proposed the following iterative algorithm: a control and its state being given, compute the associated costate, then compute simultaneously a new control and a new state such that the state equation holds and that, at each time, the new control minimizes some function. This function is the sum of the Hamiltonian and of a quadratic term penalizing the difference between the new and the old controls. It is proved in [4] that, under some hypothesis, the sequence of the criterion decreases and that, if a subsequence of the controls converges a.e. towards some control $\bar{u}$, then $\bar{u}$ satisfies the first-order necessary optimality conditions.

We obtain here the following results:

- Under some regularity hypothesis, the algorithm is well-defined.
- the norm of the projected gradients of the controls computed by the algorithm tends towards zero. If the problem is convex, we deduce from it the weak convergence towards some optimal control.
- If the step size tends towards zero, the controls computed by the algorithm are equivalent in some way to those computed by a gradient plus projection method.
- If the control enters in a linear way in the state equation and in a quadratic way in the criterion, the algorithm reduces to some gradient plus projection method in some new metric.

2. Definition and well-posedness of the algorithm. We consider the following system:

$$\begin{aligned} &\frac{dy}{dt}(t) = f(y(t), u(t), t), \quad \text{a.e. } t \in (0, T), \\ &y(0) = 0; \end{aligned} \tag{2.1}$$

with $T > 0$ given, $y(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^p$. We will denote y_u the solution of (2.1). To this system is associated the following criterion

$$J(u) = \int_0^T 1(y_u(t), u(t), t)\, dt + g(y_u(T)). \tag{2.2}$$

* Received by the editors January 24, 1984, and in revised form February 14, 1985.
† INRIA, Domaine de Voluceau, BP 105-Rocquencourt, 78153 Le Chesnay Cedex, France.

Let $\mathcal{U}_{ad}$ be a closed, bounded, convex subset of $\mathbb{R}^p$. The control problem is

$$\begin{aligned} &\text{Min } J(u), \\ &u(t) \in \mathcal{U}_{ad}, \quad \text{a.e. } t \in (0, T). \end{aligned} \tag{2.3}$$

A control u is said to be admissible if $u(t) \in \mathcal{U}_{ad}$, a.e. $t \in (0, T)$ and $J(u)$ makes sense.

The Hamiltonian of the problem is

$$\begin{aligned} &H: \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}, \\ &(y, u, p, t) \to 1(y, u, t) + (p, f(y, u, t)). \end{aligned} \tag{2.4}$$

The costate equation can be written as

$$\begin{aligned} &-\frac{dp}{dt} = \frac{\partial H}{\partial y}(y_u(t), u(t), p(t), t), \quad \text{a.e. } t \in (0, T), \\ &p(T) = \frac{\partial g}{\partial y}(y_u(T)). \end{aligned} \tag{2.5}$$

We denote by $\|\cdot\|$ the Euclidean norm. Define the augmented Hamiltonian

$$\begin{aligned} &K_\varepsilon: \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^p \times \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}, \\ &(y, u, v, p, t) \to H(y, u, p, t) + \frac{1}{2\varepsilon}\|u - v\|^2. \end{aligned} \tag{2.6}$$

Here is the algorithm proposed by Y. Sakawa and Y. Shindo [4]:

ALGORITHM 1.

0) Let some admissible control $u^0(t)$ and a sequence $\{\varepsilon^k\}$ of positive numbers be given. Set $k = 0$. Compute y^0, the state associated to u^0.
1) Compute the costate associated to $u^k(t)$.
2) $k = k+1$. Compute u^k, y^k such that y^k is the state corresponding to u^k and

$$K_{\varepsilon^k}(y^k, u^k, u^{k-1}, p^{k-1}, t) \leqq K_{\varepsilon^k}(y^k, u, u^{k-1}, p^{k-1}, t) \quad \forall u \in \mathcal{U}_{ad}, \text{ a.e. } t \in (0, T).$$

3) Stop if some convergence test is satisfied. Otherwise, go to 1.

We now state the hypotheses made on f, 1, g:

(2.7) f is twice continuously differentiable.

(2.8) There exists $C_1 > 0$ such that u admissible $\Rightarrow \|y_u(t)\| < C_1 \forall t \in [0, T]$.

(2.9) $1, \dfrac{\partial 1}{\partial y}, \dfrac{\partial 1}{\partial u}, \dfrac{\partial^2 1}{\partial y^2}, \dfrac{\partial^2 1}{\partial y\, \partial u}$ are continuous on $\mathbb{R}^n \times \mathbb{R}^p \times [0, T]$.

(2.10) There exists $\mu > 0$ such that, C_1 being the constant in (2.8):

$$\|z\| < C_1$$

implies that the mapping

$$u \to 1(z, u, t) + \mu\|u\|^2$$

is convex for any t in $[0, T]$.

(2.11) g is twice continuously differentiable from $\mathbb{R}^n$ onto $\mathbb{R}$.

Remark 2.1. These hypotheses are less restrictive than those of [4] in which g is null and $\partial^2 H/\partial u^2$ is supposed to exist and to be a positive matrix.

We note that (see Sakawa and Shindo [4]) (2.8) holds if

$$\begin{aligned} &\exists C_2 > 0 \quad \text{such that} \\ &\|f(y, u, t)\| \leqq C_2(1+\|y\|) \quad \forall (y, u) \in \mathbb{R}^n \times \mathcal{U}_{\text{ad}}. \end{aligned} \tag{2.12}$$

We now give a mathematical sense to (2.1)-(2.5). As $\mathcal{U}_{\text{ad}}$ is bounded, (2.1), (2.7) and (2.8) imply that y is in

$$Y = \left\{ y \in L^\infty(0, T, \mathbb{R}^n); \frac{dy}{dt} \in L^\infty(0, T, \mathbb{R}^n) \right\}.$$

As $Y \subset \mathscr{C}(0, T, \mathbb{R}^n)$, the initial condition of (2.1) makes sense. So also does the criterion. We easily check that if y_u exists, (2.5) has a unique solution p_u in Y and that there exists $\delta > 0$ such that, for any admissible u:

$$\|y_u(t)\| < \delta, \qquad \|p_u(t)\| < \delta \quad \forall t \in [0, T].$$

We denote

$$E = \{(z, q) \in \mathbb{R}^n \times \mathbb{R}^n; \|z\| < \delta \text{ and } \|q\| < \delta\}.$$

THEOREM 2.1. *Under hypotheses* (2.7) *to* (2.11), *there exists* $\varepsilon_0 > 0$ *such that, if* $\varepsilon^k < \varepsilon_0$, *for all* $k = 1, 2, \cdots$, *Algorithm* 1 *defines a uniquely defined sequence* $\{u^k\}$ *of admissible controls.*

Before proving the theorem, we first give:

LEMMA 2.1. *Let* $t \in [0, T]$ *be given. Define the* (*a priori multivalued*) *mapping*

$$\begin{aligned} &E \times \mathcal{U}_{\text{ad}} \times [0, \mathrm{T}] \to \mathbb{R}^p, \\ &u_\varepsilon : (z, q, v, t) \to u_\varepsilon(z, q, v, t). \end{aligned}$$

defined by

$$\begin{aligned} &u_\varepsilon \in \mathcal{U}_{\text{ad}}, \\ &K_\varepsilon(z, u_\varepsilon, v, q, t) \leqq K_\varepsilon(z, u, v, q, t) \quad \forall u \in \mathcal{U}_{\text{ad}}. \end{aligned} \tag{2.13}$$

Under hypotheses (2.7) *to* (2.10), *there exist* $\varepsilon_0 > 0$ *and* $\alpha > 0$ *independent of* t, *such that, if* $\varepsilon < \varepsilon_0$, *problem* (2.13) *has a unique solution* u_ε *continuous with respect to* t *such that*:

$$\|u_\varepsilon(z_2, q_2, v_2, t) - u_\varepsilon(z_1, q_1, v_1, t)\| \leqq \alpha(\|z_2 - z_1\| + \|q_2 - q_1\| + \|v_2 - v_1\|). \tag{2.14}$$

Proof of Lemma 2.1. From (2.7), (2.9) and (2.10) we deduce the existence of $C_2 > 0$ such that, for any $(z, q) \in E$, $t \in [0, T]$, $v, u_1, u_2 \in \mathcal{U}_{\text{ad}}$:

$$\begin{aligned} K_\varepsilon(z, u_2, v, q, t) \geqq K_\varepsilon(z, u_1, v, q, t) + \left(\frac{\partial K_\varepsilon}{\partial u}(z, u_1, v, q, t), u_2 - u_1 \right) \\ + \left(\frac{1}{2\varepsilon} - C_2 \right) \|u_2 - u_1\|^2. \end{aligned} \tag{2.15}$$

Consequently, if $\varepsilon < 1/2C_2$, K_ε is strictly convex with respect to u: this implies that (2.13) then has a unique solution. From now on we suppose that $\varepsilon < \varepsilon_0$, with $\varepsilon_0 < 1/2C_2$.

From (2.15) we deduce that

$$\left(\frac{\partial K_\varepsilon}{\partial u}(z, u_2, v, q, t) - \frac{\partial K_\varepsilon}{\partial u}(z, u_1, v, q, t), u_2 - u_1 \right) \geqq \left(\frac{1}{\varepsilon} - 2C_2 \right) \|u_2 - u_1\|^2. \tag{2.16}$$

We write the first order optimality condition of (2.13):

$$\left(\frac{\partial K_\varepsilon}{\partial u}(z, u_\varepsilon, v, q, t), u - u_\varepsilon\right) \geqq 0 \quad \forall u \in \mathcal{U}_{\text{ad}}. \tag{2.17}$$

For $i = 1, 2$, take z_i, q_i, v_i in $E \times E \times \mathcal{U}_{\text{ad}}$ and denote $u_i = u_\varepsilon(z_i, q_i, v_i, t)$. The above inequality implies

$$\left(\frac{\partial K_\varepsilon}{\partial u}(z_i, u_i, v_i, q_i, t), u_{3-i} - u_i\right) \geqq 0, \qquad i = 1, 2.$$

Adding the inequalities for $i = 1, 2$, we get

$$\left(\frac{\partial K_\varepsilon}{\partial u}(z_2, u_2, v_2, q_2, t) - \frac{\partial K_\varepsilon}{\partial u}(z_1, u_1, v_1, q_1, t), u_2 - u_1\right) \leqq 0,$$

or equivalently

$$\left(\frac{\partial K_\varepsilon}{\partial u}(z_2, u_2, v_2, q_2, t) - \frac{\partial K_\varepsilon}{\partial u}(z_2, u_1, v_2, q_2, t), u_2 - u_1\right)$$
$$\leqq \frac{\partial K_\varepsilon}{\partial u}(z_1, u_1, v_1, q_1, t) - \frac{\partial K_\varepsilon}{\partial u}(z_2, u_1, v_2, q_2, t), u_2 - u_1.$$

Using (2.16) and arranging the right-side member, we get

$$\left(\frac{1}{\varepsilon} - 2C_2\right)\|u_2 - u_1\|^2 \leqq \frac{1}{\varepsilon}(v_2 - v_1, u_2 - u_1)$$
$$+ \left(\frac{\partial H}{\partial u}(z_1, u_1, q_1, t) - \frac{\partial H}{\partial u}(z_2, u_1, q_2, t), u_2 - u_1\right).$$

This implies

$$(1 - 2C_2\varepsilon)\|u_2 - u_1\| \leqq \|v_2 - v_1\| + \varepsilon\left\|\frac{\partial H}{\partial u}(z_1, u_1, q_1, t) - \frac{\partial H}{\partial u}(z_2, u_1, q_2, t)\right\|.$$

Hypotheses (2.7), (2.11) imply that $\partial H/\partial u$ is continuously differentiable with respect to (z, q) in E. This implies that for some $C_3 > 0$ independent of $t \in [0, T]$, we have:

$$(1 - 2C_1\varepsilon)\|u_2 - u_1\| \leqq \|v_2 - v_1\| + C_3(\|z_2 - z_1\| + \|q_2 - q_1\|).$$

This proves (2.14). The continuity with respect to t of u_ε is a consequence of the continuity of $\partial K_\varepsilon/\partial u$ with respect to t and of the sufficiency of the optimality condition (2.17) for $\varepsilon < \varepsilon_0$. □

Proof of Theorem 2.1. Let us define

$$\phi: \mathbb{R}^n \times [0, T] \to \mathbb{R}^n,$$
$$\phi(z, t) = f(x, u_\varepsilon(z, p(t), u^0(t), t), t).$$

Then as $\|p(t)\| < \delta$ and $u^0(t)$ is in $\mathcal{U}_{\text{ad}}$ a.e. $t \in [0, T]$, Lemma 2.1 implies that ϕ is Lipschitzian with respect to z in any bounded neighbourhood $\mathcal{U}$ of 0 in $\mathbb{R}^n$, the Lipschitz constant being independent of t, and uniformly bounded in $\mathcal{U}$. The mapping ϕ is obviously measurable. Then by a theorem of Caratheodory (see [7]) the differential equation

$$\frac{dz}{dt} = \phi(z, t), \qquad t \in [0, T], \qquad z(0) = 0,$$

admits a unique solution for $t \in [0, t_0]$ for some $t_0 > 0$. But condition (2.8) implies that $z(t)$ remain uniformly bounded. This implies that the equation above has a unique solution in $[0, T]$. As $z = y^1$ and $u_\varepsilon(z(t), p(t), u^0(t), t) = u^1$, this proves that u^1 is uniquely defined, and so is u^k by recurrence. □

3. Convergence of the algorithm. This paragraph studies the asymptotic behaviour of the sequences generated by the algorithm. We first state a result similar to one of [4].

THEOREM 3.1. *Under hypotheses* (2.7) *to* (2.11), *there exists* $\alpha > 0$ *such that any sequence generated by the algorithm satisfies*

$$J(u^k) - J(u^{k-1}) \leqq -\left(\frac{1}{\varepsilon^k} - \alpha\right) \|u^k - u^{k-1}\|^2_{L^2(0,T,\mathbb{R}^p)}. \tag{3.1}$$

Proof. We drop the variable t when there is no ambiguity. The decrease of the criterion can be written as

$$\begin{aligned} J(u^k) - J(u^{k-1}) = \int_0^T & [H(y^k, u^k, p^{k-1}, t) - H(y^{k-1}, u^{k-1}, p^{k-1}, t) \\ & - (p^{k-1}, f(y^k, u^k, t) - f(y^{k-1}, u^{k-1}, t))]\, dt \\ & + g(y^k(T)) - g(y^{k-1}(T)). \end{aligned}$$

Define

$$\delta y^k = y^k - y^{k-1}, \qquad \delta u^k = u^k - u^{k-1}.$$

Then, with (2.1):

$$\begin{aligned} J(u^k) - J(u^{k-1}) = & \int_0^T \left[H(y^k, u^k, p^{k-1}, t) - H(y^{k-1}, u^{k-1}, p^{k-1}, t) - \left(p^{k-1}, \frac{d}{dt}(\delta y^k)\right)\right] dt \\ & + g(y^k(T)) - g(y^{k-1}(T)). \end{aligned} \tag{3.2}$$

We obviously have

$$\begin{aligned} H(y^k, u^k, p^{k-1}, t) - H(y^{k-1}, u^{k-1}, p^{k-1}, t) = & H(y^k, u^k, p^{k-1}, t) \\ & - H(y^k, u^k - \delta u^k, p^{k-1}, t) + H(y^{k-1} + \delta y^k, u^{k-1}, p^{k-1}, t) - H(y^{k-1}, u^{k-1}, p^{k-1}, t). \end{aligned}$$

It is easy to check that the hypotheses made imply the existence of $C_4 > 0$, such that, $\forall t \in [0, T]$:

$$\begin{aligned} H(y^k, u^k - \delta u^k, p^{k-1}, t) \geqq & H(y^k, u^k, p^{k-1}, t) \\ & - \left(\frac{\partial H}{\partial u}(y^k, u^k, p^{k-1}, t), \delta u^k\right) - C_4 \|\delta u^k\|^2. \end{aligned}$$

$$\begin{aligned} H(y^{k-1} + \delta y^k, u^{k-1}, p^{k-1}, t) \leqq & H(y^{k-1}, u^{k-1}, p^{k-1}, t) \\ & + \left(\frac{\partial H}{\partial y}(y^{k-1}, u^{k-1}, p^{k-1}, t), \delta y^k\right) + C_4 \|\delta y^k\|^2. \end{aligned}$$

These inequalities with (3.2) imply

$$\begin{aligned} J(u^k) - J(u^{k-1}) \leqq \int_0^T & \left[\left(\frac{\partial H}{\partial u}(y^k, u^k, p^{k-1}, t), \delta u^k\right) + C_4 \|\delta u^k\|^2 \right. \\ & \left. + \left(\frac{\partial H}{\partial y}(y^{k-1}, u^{k-1}, p^{k-1}, t), \delta y^k\right)\right. \end{aligned}$$

$$+C_4\|\delta y^k\|^2-\left(p^{k-1},\frac{d}{dt}(\delta y^k)\right)\Big]\,dt \tag{3.3}$$
$$+g(y^k(T))-g(y^{k-1}(T)).$$

We note that

$$\left(\frac{\partial H}{\partial u}(y^k,u^k,p^{k-1},t),\delta u^k\right)=\left(\frac{\partial K_{\varepsilon^k}}{\partial u}(y^k,u^k,u^{k-1},p^{k-1},t),\delta u^k\right)-\frac{1}{\varepsilon^k}\|\delta u^k\|^2$$

and, by definition of u^k

$$\left(\frac{\partial K_{\varepsilon^k}}{\partial u}(y^k,u^k,u^{k-1},p^{k-1},t),\delta u^k\right)\leqq 0,\quad \text{a.e. } t\in(0,T).$$

On the other hand, using (2.5) and integrating by parts,

$$\int_0^T\left[\left(\frac{\partial H}{\partial y}(y^{k-1},u^{k-1},p^{k-1},t),\delta y^k\right)-\left(p^{k-1},\frac{d}{dt}(\delta y^k)\right)\right]dt$$
$$=-\int_0^T\left[\left(\frac{d}{dt}p^{k-1},\delta y^k\right)+\left(p^{k-1},\frac{d}{dt}(\delta y^k)\right)\right]dt$$
$$=-[(p^{k-1}(t),\delta y^k(t))]_0^T=-\left(\frac{\partial g}{\partial y}(y^{k-1}(T)),\delta y^k(T)\right).$$

As $y(T)$ is uniformly bounded in $\mathbb{R}^n$, we deduce from (2.11) that for some $C_5>0$:

$$g(y^k(T))-g(y^{k-1}(T))-\left(\frac{\partial g}{\partial y}(y^{k-1}(T)),\delta y^k(T)\right)\leqq C_5\|\delta y^k(T)\|^2.$$

Using (3.3), we deduce that

$$J(u^k)-J(u^{k-1})\leqq\int_0^T\left[\left(C_4-\frac{1}{\varepsilon^k}\right)\|\delta u^k(t)\|_p^2+C_4\|\delta y^k(t)\|^2\right]dt \tag{3.4}$$
$$+C_5\|\delta y^k(T)\|^2.$$

We end the proof as in [4]: we have

$$\frac{d}{dt}\|\delta y^k(t)\|\leqq\left\|\frac{d}{dt}\delta y^k(t)\right\|=\|f(y^k,u^k,t)-f(y^{k-1},u^{k-1},t)\|.$$

Hypothesis (2.7) and the fact that y and u are uniformly bounded imply the existence of $C_6>0$ such that

$$\frac{d}{dt}\|\delta y^k(t)\|\leqq C_6(\|\delta y^k(t)\|+\|\delta u^k(t)\|).$$

Using Gronwall's inequality, we deduce that for some $C_7>0$

$$\|\delta y^k(t)\|\leqq C_7\int_0^T\|\delta u^k(t)\|\,dt,$$

and with the Cauchy-Schwarz inequality:

$$\|\delta y^k(t)\|^2\leqq C_8T\int_0^T\|\delta u^k(t)\|^2\,dt.$$

This implies

$$\int_0^T \|\delta y^k(t)\|^2\, dt \leqq C_8(T)^2 \int_0^T \|\delta u^k(t)\|^2\, dt.$$

The two last inequalities with (3.4) yields (3.1). □

We denote by $P_{\mathcal{U}_{ad}}$ the map from $L^2(0, T, \mathbb{R}^p)$ onto itself such that $P_{\mathcal{U}_{ad}}(u)(t)$ is a.e. in t the projection of $u(t)$ on $\mathcal{U}_{ad}$. The gradient of J is denoted ∇J. We state a result of global convergence:

THEOREM 3.2. *We suppose that J is bounded from below and that hypotheses* (2.7) *to* (2.11) *hold. Then there exists $\varepsilon_0 > 0$ such that, if $\varepsilon^k < \varepsilon_0$, any sequence generated by the algorithm satisfies*:

$$J(u^k) \text{ decreases}, \tag{3.5}$$

$$\|u^k - u^{k-1}\|_{L^2(0,T,\mathbb{R}^p)} \to 0, \tag{3.6}$$

$$\|u^k - P_{\mathcal{U}_{ad}}(u^k - \varepsilon^k \nabla J(u^k))\|_{L^2(0,T,\mathbb{R}^p)} \to 0. \tag{3.7}$$

Proof. Relations (3.5) and (3.6) are an easy consequence of Theorem 3.1. Let us verify (3.7). From the definition of u^k we deduce that

$$\left(u^k - u^{k-1} + \varepsilon^k \frac{\partial H}{\partial u}(y^k, u^k, p^{k-1}, t),\ v - u^k\right) \geqq 0 \quad \forall v \in \mathcal{U}_{ad}, \quad \text{a.e. } t \in (0, T).$$

This can be written as

$$u^k = P_{\mathcal{U}_{ad}}\left(u^{k-1} - \varepsilon^k \frac{\partial H}{\partial u}(y^k, u^k, p^{k-1}, t)\right).$$

We note that

$$\nabla J(u^{k-1}) = \frac{\partial H}{\partial u}(y^{k-1}, u^{k-1}, p^{k-1}, t),$$

so that

$$u^{k-1} - P_{\mathcal{U}_{ad}}(u^{k-1} - \varepsilon^k \nabla J(u^{k-1})) = u^{k-1} - u^k + P_{\mathcal{U}_{ad}}\left(u^{k-1} - \varepsilon^k \frac{\partial H}{\partial u}(y^k, u^k, p^{k-1}, t)\right)$$
$$- P_{\mathcal{U}_{ad}}\left(u^{k-1} - \varepsilon^k \frac{\partial H}{\partial u}(y^{k-1}, u^{k-1}, p^{k-1}, t)\right).$$

As $P_{\mathcal{U}_{ad}}$ is a contraction, we obtain

$$\|u^{k-1} - P_{\mathcal{U}_{ad}}(u^{k-1} - \varepsilon^k \nabla J(u^{k-1}))\|_{L^2(0,T,\mathbb{R}^p)} \leqq \|u^{k-1} - u^k\|_{L^2(0,T,\mathbb{R}^p)}$$
$$+ \varepsilon^k \left\| \frac{\partial H}{\partial u}(y^k, u^k, p^{k-1}, t) - \frac{\partial H}{\partial u}(y^{k-1}, u^{k-1}, p^{k-1}, t) \right\|_{L^2(0,T,\mathbb{R}^p)}. \tag{3.8}$$

From (3.6) we deduce that

$$u^k(t) - u^{k-1}(t) \to 0,$$
$$y^k(t) - y^{k-1}(t) \to 0, \quad \text{a.e. } t \in (0, T).$$
$$p^k(t) - p^{k-1}(t) \to 0.$$

As $\partial H/\partial u$ is uniformly continuous on compact sets, this implies that

$$\frac{\partial H}{\partial u}(y^k, u^k, p^{k-1}, t) - \frac{\partial H}{\partial u}(y^{k-1}, u^{k-1}, p^{k-1}, t) \to 0, \quad \text{a.e. } t \in [0, T].$$

Using Lebesgue's theorem, we see that the right side of (3.8) tends towards zero. □

The hypotheses we made do not allow us to prove the convergence of $\{u^k\}$: in fact they do not even insure the existence of an optimal control. However, stronger results are easily deduced from (3.8) under an additional hypothesis: here is an example:

THEOREM 3.3. *We suppose that $J(u)$ is an l.s.c. convex function, defined on a convex domain, and bounded from below, that hypotheses* (2.7) *to* (2.11) *hold, that ε is small enough so that* (3.5), (3.6) *and* (3.7) *hold and that* $\liminf \varepsilon^k > 0$. *Then some subsequence of $\{u^k\}$ (still denoted $\{u^k\}$) satisfies*

$$u^k \to \bar{u} \qquad L^2(0, T, \mathbb{R}^p) \text{ weak},$$

where $\bar{u}$ is an optimal control.

The proof of the theorem is a direct consequence of (3.7) and of the following lemma:

LEMMA 3.1. *Let U be an Hilbert space and K a closed convex subset of U. Let $J: U \to \mathbb{R}$ be an l.s.c. proper convex functional. Let $\{u^k\}$ be a bounded sequence such that J is Gâteaux differentiable on $\{u^k\}$ and $\{\nabla J(u^k)\}$ is bounded. We assume that there exists a sequence ρ^k of positive numbers such that* $\liminf \rho^k \geqq \rho > 0$ *and that*

$$\|u^k - P_K(u^k - \rho^k \nabla J(u^k))\|_U \to 0.$$

Then any weak limit point of $\{u^k\}$ (there exist some) satisfies

$$\bar{u} \in K,$$

$$J(\bar{u}) \leqq J(u) \quad \forall u \in K.$$

Proof of Lemma 3.1. Denote $g^k = \nabla J(u^k)$, and suppose that for some subsequence (still denoted k) $u^k \to \bar{u}$ in U. Then $\bar{u}$ is in K. Denote

$$v^k = P_K(u^k - \rho^k g^k).$$

We have

$$(v^k - u^k + \rho^k g^k, w - v^k) \geqq 0 \quad \forall w \in K,$$

i.e.

$$(v^k - u^k, w - v^k) + \rho^k (g^k, w - u^k) + \rho^k (g^k, u^k - v^k) \geqq 0 \quad \forall v \in K.$$

We made the hypothesis $\|v^k - u^k\|_U \to 0$. Then as $\{g^k\}$ is bounded and as $\rho^k \geqq \rho/2$ for k great enough:

$$\liminf_{k \to +\infty} (g^k, w - u^k) \geqq 0 \quad \forall w \in K. \tag{3.9}$$

As J is convex, we have

$$J(w) \geqq J(u^k) + (g^k, w - u^k) \quad \forall w \in H.$$

Hence with (3.9), for any $w \in K$:

$$J(w) \geqq \liminf \{J(u^k) + (g^k, w - u^k)\} \geqq \liminf J(u^k).$$

The weak lower semi-continuity of J gives the conclusion. □

Remark 3.1. Lemma 3.1. still holds if we only suppose that g^k is an element of the subdifferential of J at u^k.

4. Relations with the gradient plus projection method. We have seen in the proof of Theorem 3.2. that the algorithm can be written as

$$u^k = P_{\mathcal{U}_{ad}}\left(u^{k-1} - \varepsilon \frac{\partial H}{\partial u}(y^k, u^k, p^{k-1}, t)\right). \tag{4.1}$$

The gradient plus projection method (A. A. Goldstein [2]) for a control problem can be written as

$$\hat{u}^k = P_{\mathcal{U}_{ad}}\left(u^{k-1} - \varepsilon \frac{\partial H}{\partial u}(y^{k-1}, u^{k-1}, p^{k-1}, t)\right), \tag{4.2}$$

where ε is the step size. We fix u^{k-1} and show that the two points obtained by (4.1) and (4.2) are very near when ε is small.

PROPOSITION 4.1. *Let $u^{k-1} \in L^2(0, T, \mathbb{R}^p)$ be an admissible control. We denote u_ε^k and $\hat{u}_\varepsilon^k$ the new controls obtained by* (4.1) *and* (4.2). *Then, if $\hat{u}_\varepsilon^k \neq u^{k-1}$ for some $\varepsilon > 0$:*

$$\lim_{\varepsilon \to 0} \frac{\|u_\varepsilon^k - \hat{u}_\varepsilon^k\|_{L^2(0,T,\mathbb{R}^p)}}{\|\hat{u}_\varepsilon^k - u^{k-1}\|_{L^2(0,T,\mathbb{R}^p)}} \to 0.$$

Proof. As $P_{\mathcal{U}_{ad}}$ is a contraction, (4.1), (4.2) imply that (dropping the subscripts ε):

$$\|u^k - \hat{u}^k\|_{L^2(0,T,\mathbb{R}^p)} \leqq \varepsilon \left\| \frac{\partial H}{\partial u}(y^k, u^k, p^k, t) - \frac{\partial H}{\partial u}(y^{k-1}, u^{k-1}, p^{k-1}, t) \right\|_{L^2(0,T,\mathbb{R}^p)}.$$

The continuous function $\partial H/\partial u$ is uniformly continuous on the bounded set of admissible states, controls and costates, for $t \in [0, T]$, so that $\forall \alpha > 0$, there exists ε_1 such that

$$\varepsilon < \varepsilon_1 \Rightarrow \|u^k - \hat{u}^k\|_{L^2(0,T,\mathbb{R}^p)} \leqq \alpha\varepsilon. \tag{4.3}$$

Lemma 1 of D. P. Bertsekas, E. M. Gafni [1] states that

$$\frac{1}{\varepsilon}\|\hat{u}_\varepsilon^k - u^{k-1}\|_{L^2(0,T,\mathbb{R}^p)} \geqq \frac{1}{\varepsilon'}\|\hat{u}_{\varepsilon'}^k - u^{k-1}\|_{L^2(0,T,\mathbb{R}^p)} \quad \forall \varepsilon' > \varepsilon > 0;$$

hence

$$\|\hat{u}^k - u^{k-1}\|_{L^2(0,T,\mathbb{R}^p)} \geqq C_1\varepsilon$$

for some $C_1 > 0$. This with (4.3) proves the proposition. □

We now see that in an important particular case, the algorithm is equivalent to a gradient plus projection method in some metric.

THEOREM 4.1. *We assume that the control enters in a linear way into the state equation and in a quadratic way into the criterion, i.e.*

$$f(y, u, t) = f_1(y, t) + B(t)u, \qquad l(y, u, t) = \tfrac{1}{2}u^t N(t)u + l_1(y, t),$$

where $B(t)$ and $N(t)$ are time-dependent matrices of convenient dimension, and $N(t)$ is symmetric. We denote

$$M_\varepsilon(t) = I + \varepsilon N(t).$$

Then the algorithm reduces to the gradient plus projection method in $L^2(0, T, \mathbb{R}^p)$ endowed

with the new metric

$$((u, v))_{L^2(0,t,\mathbb{R}^p)} = \int_0^T (u(t), M_\varepsilon(t)v(t))\, dt. \tag{4.4}$$

Proof. From (4.1) written with $k=1$ we see that u^1 is characterized by

$$(u^1(t)-u^0(t)+\varepsilon(N(t)u^1(t)+B^*(t)p^0(t)), v-u^1(t))\geqq 0, \quad \forall v\in \mathcal{U}_{\text{ad}}, \text{ a.e. } t\in(0, T).$$

This is equivalent to

$$((I+\varepsilon N(t))u^1(t)-(I+\varepsilon N(t))u^0(t)+\varepsilon(N(t)u^0(t)+B^*(t)p^0(t)), v-u^1(t))\geqq 0$$
$$\forall v\in \mathcal{U}_{\text{ad}}, \text{ a.e. } t\in[0, T].$$

We note that $\nabla J(u^0)(t) = N(t)u^0(t)+B^*(t)p^0(t)$, so that

$$(M_\varepsilon(t)(u^1(t)-u^0(t)+\varepsilon(M_\varepsilon)^{-1}\nabla J(u^0)(t)), v-u^1(t))\geqq 0 \quad \forall v\in \mathcal{U}_{\text{ad}}, \text{ a.e. } t\in(0, T).$$

This is the characterization of the point obtained by the gradient plus projection method associated to the metric (4.4). □

5. Conclusion. In this study of an algorithm of resolution of optimal control problems, we establish some results of well-posedness and of convergence. In addition, we show that this algorithm is strongly related to the gradient plus projection method. For numerical experiments using this algorithm, see [4], [5], [6].

Acknowledgments. I thank Professor Y. Sakawa, from Osaka University, for an interesting discussion concerning the algorithm studied in this paper, and the editor and a referee for their useful remarks.

REFERENCES

[1] D. P. BERTSEKAS AND E. M. GAFNI, *Two-metric projection methods for constrained optimization*, this Journal, 22 (1984), pp. 936-964.

[2] A. A. GOLDSTEIN, *Convex programming in Hilbert space*, Bull. Amer. Math. Soc., 70 (1964), pp. 709-710.

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

[4] Y. SAKAWA AND Y. SHINDO, *On global convergence of an algorithm for optimal control*, IEEE Trans. Automat. Comput, AC-25 (1980), pp. 1149-1153.

[5] ———, *Optimal control of container cranes*, Automatica, 18 (1982), pp. 257-266.

[6] Y. SAKAWA, Y. SHINDO AND Y. HASHIMOTO, *Optimal control of a rotary crane*, J. Optim. Theory Appl., 35 (1981), pp. 535-557.

[7] G. SANSONE AND R. CONTI, *Nonlinear Differential Equations*, Pergamon Press, Oxford, 1964.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 4, July 1986

STOCHASTIC ADAPTIVE CONTROL FOR EXPONENTIALLY CONVERGENT TIME-VARYING SYSTEMS*

GRAHAM C. GOODWIN†, DAVID J. HILL† AND XIE XIANYA‡

Abstract. This paper shows that the standard stochastic adaptive control algorithms for time-invariant systems have an inherent robustness property which renders them applicable, without modification, to time-varying systems whose parameters converge exponentially. One class of systems satisfying this requirement is those having non-steady-state Kalman filter or innovations representations. This allows the usual assumption of a stationary ARMAX representation to be replaced by a more general state space model.

Key words. stochastic control, adaptive control, martingale convergence, passivity, time-varying systems

1. Introduction. A stochastic adaptive controller is an algorithm which combines on-line parameter estimation with on-line control to generate a control law applicable to systems having unknown parameters and random disturbances [1]. Control laws based on this philosophy have been studied for at least three decades [2], but it is only recently that rigorous convergence analyses have appeared. To gain insight into the operation of these algorithms, several special cases have been studied in detail. For example, the authors of [3] have examined the convergence properties of a particular scheme which combines a simple stochastic gradient parameter estimator with a minimum variance control law.

A number of interesting properties of these simple stochastic adaptive control laws have been established. For example, the tracking error is known to converge to a minimum (in a specific sense) in a sample mean square sense [3]. In the case of regulation about a zero desired output, then it has been shown [5] that the parameter estimates converge to a fixed multiple of the true parameter values. However, if the desired output sequence is continuously disturbed and an identifiability condition holds, then the parameters can be shown to converge to their true values [4]. Various extensions of the above results have also been studied. For example convergence results have been established in [7], for least squares based adaptive control algorithms.

The above papers deal with systems having constant parameters. However, in practice one is often confronted with systems whose parameters vary with time in some fashion. This has motivated several authors to investigate special classes of time-varying systems in an effort to gain insights into the convergence properties relevant to this case. For example, Caines [8] has analyzed the performance of the stochastic gradient algorithm of [3] applied to systems with (converging) martingale parameters. Further results for systems having random parameters are described in [9].

The current paper also deals with systems whose parameters are time-varying. Indeed, the work has much in common with the results in [8], [9]. All three papers reduce to treatment of a near-super-martingale equation of a particular form—see equation (3.28) later. However, here the parameter time variations are deterministic and thus a different method of analysis is necessary from that used in [8], [9].

Our analysis has three key steps: a proof that a system which is convergent toward a minimum phase system has an input which grows no faster than the output; a proof that a system which is exponentially convergent toward a strictly passive system is eventually strictly passive in a certain sense (where the strict passivity concepts are

* Received by the editors May 29, 1984, and in revised form February 28, 1985.

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

‡ Department of Computer Science, Shanghai University of Science and Technology, Shanghai, China.

defined later); and a martingale convergence proof along the lines of [3], but using a modified martingale result as first proposed in [10], [11] in a different context.

One application of the results developed here is to systems described by a state space model corresponding to a non-steady-state innovations representation. Subject to the assumption that the system has no uncontrollable modes (in the filtering sense [12]) on the unit circle, then it is known [13] that the parameters in the innovations model are time-varying and converge exponentially fast toward those of the steady-state optimal filter. Thus the results of this paper allow global convergence to be established for the standard adaptive control algorithms when applied to these systems. This represents a relaxation of the usual modelling assumption employed elsewhere in the literature (e.g. [3] to [9]) that the system is described by an ARMAX model or equivalently a steady-state Kalman filter model. This particular robustness to modelling assumption is often implicitly assumed in the literature, and it is thus interesting for technical completeness to have a formal proof that the results go through in this case.

2. Preliminary result on passive systems. We verify a result which will be needed in the subsequent proof of convergence of an adaptive control algorithm. This concerns a passivity property of a system which is exponentially convergent toward an asymptotically stable and input strictly passive system. The definitions of passivity concepts used correspond to those presented in [15, Appendix C]. Consider the extended Hilbert space $l_{2e}^n(\mathbb{Z}_+)$ of sequences $v:\mathbb{Z}_+\to\mathbb{R}^n$ with truncated inner product $\langle u, v\rangle_T := \sum_{k=0}^{T-1} u^T(k)v(k) < \infty$. The main definition for present purposes is as follows.

DEFINITION 2.1. Consider a dynamical system represented by mapping $G: l_{2e}^n(\mathbb{Z}_+)\to l_{2e}^n(\mathbb{Z}_+)$. The system is input strictly passive (ISP) iff $\exists \delta > 0$ and β such that

$$\langle y, u\rangle_T \geqq \delta\|u\|_T^2 + \beta \quad \forall u \in l_{2e}(\mathbb{Z}_+) \text{ and } T \geqq 0. \tag{2.1}$$

Consider the following special case of the time-varying linear system model (A.1), (A.2):

$$x(t+1) = A(t)x(t) + Bu(t), \tag{2.2}$$

$$y(t) = Cx(t). \tag{2.3}$$

Let $A(t)\to A$ exponentially fast and define $\{y^*(t)\}, \{x^*(t)\}$ by

$$x^*(t+1) = Ax^*(t) + Bu^*(t), \tag{2.4}$$

$$y^*(t) = Cx^*(t) \tag{2.5}$$

with (A, B) controllable and A asymptotically stable.

The following result establishes that if (2.4), (2.5) is also ISP then (2.2), (2.3) satisfies a property very close to ISP.

THEOREM 2.1. *Provided* (2.4), (2.5) *is input strictly passive then there exist* $\bar{n}_0$, β *and* $\delta > 0$ *such that*

$$\sum_{t=n_0}^{N} (y(t)u(t) - \delta u(t)^2) + \beta \geqq 0 \tag{2.6}$$

for all $N \geqq n_0 \geqq \bar{n}_0$ *and for all* $\{u(t)\} \in l_{2e}(\mathbb{Z}_+)$. β, δ *depend on* $u(t)$ *for* $t < n_0$.

Proof. Since (2.4), (2.5) is ISP, there exists $\delta^* < 0$ and $\beta^*(x^*(n_0))$ such that

$$\sum_{t=n_0}^{N} (y^*(t)u^*(t) - \delta^* u^*(t)^2) + \beta^*(x^*(n_0)) \geqq 0 \tag{2.7}$$

for all $N \geqq n_0$ and for all $u(t) \in l_{2_e}(\mathbb{Z}_+)$. From (2.7), we have

$$\sum_{t=n_0}^{N} (y^*(t)-y(t))u(t)+\sum_{t=n_0}^{N} y(t)u(t) \geqq \delta^* \sum_{t=n_0}^{N} u(t)^2-\beta^*(x^*(n_0)).$$

Let

$$\alpha(n_0, N) := \sum_{t=n_0}^{N} (y^*(t)-y(t))u(t).$$

Then

$$\sum_{t=n_0}^{N} y(t)u(t) \geqq \delta^* \sum_{t=n_0}^{N} u(t)^2-\alpha(n_0, N)-\beta^*(x^*(n_0)). \tag{2.8}$$

The remainder of the proof involves establishing a bound for $\alpha(n_0, N)$. Now

$$y^*(t) = C\Phi^*(t, n_0)x^*(n_0)+\sum_{i=n_0}^{t-1} C\Phi^T(t, i+1)Bu^*(i),$$

$$y(t) = C\Phi(t, n_0)x(n_0)+\sum_{i=n_0}^{t-1} C\Phi(t, i+1)Bu(i)$$

where $\Phi^*(t, n_0) = A^{t-n_0}$ and $\Phi(t, n_0)$ is given by (A.3). Then

$$\begin{aligned}\alpha(n_0, N) &= \sum_{t=n_0}^{N} u(t)C\left(A^{t-n_0}x^*(n_0)-\prod_{i=n_0}^{t-1} A(i)\right)x(n_0)\\ &\quad+\sum_{t=n_0}^{N} u(t)C\left(\sum_{i=n_0}^{t-1}\left(A^{t-i-1}-\prod_{j=i+1}^{t-1} A(j)\right)Bu(i)\right)\\ &:= \alpha_1(n_0, N)+\alpha_2(n_0, N).\end{aligned}$$

Now without loss of generality choose $u^*(t)$ so that

$$x^*(n_0) = x(n_0) \quad \text{and} \quad y^*(n_0) = y(n_0).$$

This is possible because (2.4), (2.5) is controllable. We can consider $u(t) = u^*(t)$ for $t \geqq n_0$. We have

$$|\alpha(n_0, N)| \leqq |\alpha_1(n_0, N)|+|\alpha_2(n_0, N)|.$$

The remainder of the proof involves bounding α_1 and α_2 in terms of $\|u(t)\|_{n_0}^N$.

We will use $|\cdot|$ for the Euclidean norm.

$$\begin{aligned}|\alpha_1(n_0, N)| &= \left|\sum_{t=n_0}^{N} u(t)C\left(A^{t-n_0}-\prod_{i=n_0}^{t-1} A(i)\right)x(n_0)\right|\\ &\leqq \|u(t)\|_{n_0}^N\left\|C\left(A^{t-n_0}-\prod_{i=n_0}^{t-1} A(i)\right)x(n_0)\right\|_{n_0}^N\end{aligned}$$

using the Schwarz inequality

$$\leqq \|u(t)\|_{n_0}^N \sum_{t=n_0}^{N}\left|C\left(A^{t-n_0}-\prod_{i=n_0}^{t-1} A(i)\right)x(n_0)\right|$$

since $\|\cdot\|_2 \leqq \|\cdot\|_1$

$$\leqq \|u(t)\|_{n_0}^N|C||x(n_0)|\sum_{t=n_0}^{N}\left|A^{t-n_0}-\prod_{i=n_0}^{t-1} A(i)\right|$$

$$\leqq \|u(t)\|_{n_0}^N|C||x(n_0)|\eta^{n_0-1}\sum_{t=n_0}^{N}\chi\eta^{t-n_0}$$

choosing $n_0 \geqq \bar{n}$ and using Lemma A.2

$$\leqq 2\varepsilon_1\|u(t)\|_{n_0}^N|x(n_0)|$$

where

$$\varepsilon_1 = \frac{|C|\chi\eta^{n_0-1}}{2(1-\eta)} \leqq \varepsilon_1(\|u(t)\|_{n_0}^N)^2 + \varepsilon_1|x(n_0)|^2.$$

So $|\alpha_1(n_0, N)| \leqq \varepsilon_1(\|u(t)\|_{n_0}^N)^2 + \varepsilon_1|x(n_0)|^2$.

Now turning to α_2, we have

$$|\alpha_2(n_0, N)| = \left| \sum_{t=n_0}^{N} u(t) C \left(\sum_{i=n_0}^{t-1} \left(A^{t-i-1} - \prod_{j=i+1}^{t-1} A(j) \right) Bu(i) \right) \right|$$

$$\leqq \|u(t)\|_{n_0}^N \left\| C \sum_{i=n_0}^{t-1} \left(A^{t-i-1} - \prod_{j=i+1}^{t-1} A(j) \right) Bu(i) \right\|_{n_0}^N.$$

Observe that

$$\left\| C \sum_{i=n_0}^{t-1} \left(A^{t-i-1} - \prod_{j=i+1}^{t-1} A(j) \right) Bu(i) \right\|_{n_0}^N$$

$$= \left\{ \sum_{t=n_0}^{N} \left(\sum_{i=n_0}^{t-1} C \left(A^{t-i-1} - \prod_{j=i+1}^{t-1} A(j) \right) Bu(i) \right)^2 \right\}^{1/2}$$

$$\leqq \left\{ \sum_{t=n_0}^{N} \left(\sum_{i=n_0}^{t-1} \left(C \left(A^{t-i-1} - \prod_{j=i+1}^{t-1} A(j) \right) B \right)^2 \right) \left(\sum_{i=n_0}^{t-1} u(i)^2 \right) \right\}^{1/2}$$

using the Schwarz inequality

$$\leqq \left\{ \sum_{t=n_0}^{N} \sum_{i=n_0}^{t-1} \left(C \left(A^{t-i-1} - \prod_{j=i+1}^{t-1} A(j) \right) B \right)^2 \right\}^{1/2} \|u(t)\|_{n_0}^N$$

$$\leqq |C||B| \|u(t)\|_{n_0}^N \left\{ \sum_{t=n_0}^{N} \sum_{i=n_0}^{t-1} \chi^2 \eta^{2(t-1)} \right\}^{1/2}$$

using Lemma A.2

$$= |C||B|\chi \|u(t)\|_{n_0}^N \left\{ \sum_{t=n_0}^{N} \sum_{i=n_0}^{t-1} \eta^{2(t-1)} \right\}^{1/2}.$$

Now

$$\sum_{t=n_0}^{N} \sum_{i=n_0}^{t-1} \eta^{2(t-1)} = \sum_{p=0}^{N-n_0} \sum_{q=0}^{p-1} \eta^{2(p+n_0-1)}$$

$$\leqq \eta^{2(n_0-1)} \sum_{p=0}^{N-n_0} p\eta^{2p}$$

$$\leqq \eta^{2(n_0-1)} \frac{\eta^2}{(1-\eta^2)^2}$$

since $|\eta| < 1$

$$= \frac{\eta^{2n_0}}{(1-\eta^2)^2}.$$

So we have

$$|\alpha_2(n_0, N)| \leqq (\|u(t)\|_{n_0}^N)^2 |C||B|\chi \frac{\eta^{2n_0}}{(1-\eta^2)^2}.$$

Let

$$\varepsilon_2 = |C||B|\chi \frac{\eta^{2n_0}}{(1-\eta^2)^2}.$$

So

$$|\alpha_2(n_0, N)| \leqq \varepsilon_2(\|u(t)\|_{n_0}^N)^2.$$

We then have

$$|\alpha(n_0, N)| \leqq (\varepsilon_1+\varepsilon_2)(\|u(t)\|_{n_0}^N)^2+\varepsilon_1|x(n_0)|^2.$$

Note that ε_1, ε_2 can be made arbitrarily small by taking n_0 large enough.

Let

$$\varepsilon := \varepsilon_1+\varepsilon_2.$$

Substituting into (2.8)

$$\sum_{t=n_0}^{N} y(t)u(t) \geqq (\delta^*-\varepsilon)(\|u(t)\|_{n_0}^N)^2-\beta^*(x(n_0))-\varepsilon_1|x(n_0)|^2.$$

Let

$$\delta := \delta^*-\varepsilon, \qquad \beta := \beta^*(x(n_0))+\varepsilon_1|x(n_0)|^2.$$

By taking n_0 large enough, it can be guaranteed that $\delta>0$. □

Remarks. 1. Since system (2.4), (2.5) is both input strictly passive and asymptotically stable, it is in fact very strictly passive (see [15, Appendix C]).

2. The system does not become input strictly passive as usually defined because δ is dependent on $x(n_0)$.

3. The adaptive control algorithm. We are concerned here with the adaptive control of a linear time-varying finite dimensional system admitting an autoregressive moving average representation of the form:

$$\begin{aligned} &y(t)+a_1(t)y(t-1)+\cdots+a_n(t)y(t-n)\\ &\quad=b_0(t)u(t-d)+\cdots+b_m(t)u(t-d-m)\\ &\qquad+\omega(t)+c_1(t)\omega(t-1)+\cdots+c_l(t)\omega(t-l). \end{aligned} \tag{3.1}$$

We shall express (3.1) in compact notation as

$$A(t, q^{-1})y(t) = \acute{B}(t, q^{-1})q^{-d}u(t)+C(t, q^{-1})\omega(t) \tag{3.2}$$

where q^{-1} represents the delay operator and $A(t, q^{-1}) = 1+a_1(t)q^{-1}+\cdots+a_n(t)q^{-n}$; $B(t, q^{-1}) = b_0(t)+b_1(t)q^{-1}+\cdots+b_m(t)q^{-m}$; $C(t, q^{-1}) = 1+c_1(t)q^{-1}+\cdots+c_l(t)q^{-l}$. The corresponding initial condition is $x_0 := \{y(0)\cdots y(1-k);\ u(1-d), \cdots, u(1-k);\ \omega(0), \cdots, \omega(1-k)\}$ where $k = \max\{n, m+d, l\}$.

The process $\{x_0, \omega(1), \omega(2), \cdots\}$ is defined on the underlying probability space $(\Omega, \mathcal{F}, P)$ and we define $\mathcal{F}_0$ to be the σ-algebra generated by x_0. Further, for all $t \geqq 1$ $\mathcal{F}_t$ shall denote the σ-algebra generated by the observations up to time t. The distributions of the random variables $x_0, \omega(1), \omega(2), \cdots$ are assumed mutually absolutely continuous with respect to Lebesgue measure.

We make the following assumptions on the process $\{\omega(t)\}$:

(3.3) N.1 $\quad E\{\omega(t)|\mathcal{F}_{t-1}\} = 0$ a.s. $t \geqq 1$.

(3.4) N.2 $\quad E\{\omega(t)^2|\mathcal{F}_{t-1}\} = \sigma_t^2 \leqq \sigma^2 < \infty, \qquad t \geqq 1.$

(3.5) N.3 $\quad \limsup_{N\to\infty} \frac{1}{N}\sum_{t=1}^{N} \|\omega(t)\|^2 < \infty$ a.s.

We wish to design an adaptive control law to cause $\{y(t)\}$ to track (in some sense) a given desired output sequence $\{y^*(t)\}$ and to ensure that $\{y(t)\}$, $\{u(t)\}$ remain bounded (in some sense). Reference [3] presents further background to this problem as well as giving a convergence analysis for a particular adaptive control algorithm for the case when $A(t, q^{-1})$, $B(t, q^{-1})$, $C(t, q^{-1})$ do not depend on t.

In order to specify the algorithm, we assume:

S.1. d is known.

S.2. Upper bounds for n, m, and l are known.

In addition, the following properties of system (3.2) will be assumed in the analysis of the algorithm:

S.3
$$a_i(t) \to a_i, \qquad i = 1, \cdots, n,$$
$$b_i(t) \to b_i, \qquad i = 0, \cdots, m,$$
$$c_i(t) \to c_i, \qquad i = 1, \cdots, l,$$

exponentially fast.

S.4. $B(z)$ and $C(z)$ have all zeros outside the closed unit circle, where

$$B(z) = b_0 + b_1 z + \cdots + b_m z^m,$$
$$C(z) = 1 + c_1 z + \cdots + c_l z^l.$$

S.5. The system $C(q^{-1})z(t) = b(t)$ is input strictly passive.

For simplicity we shall treat the single input single output unit delay ($d = 1$) case. However, natural extensions exist for the multi-input multi-output nonunit delay as explored for non-time-varying systems in [3], [15], etc.

The model (3.2) can be rearranged into the following predictor form:

$$C(t, q^{-1})[y(t) - \omega(t)] = \alpha(t, q^{-1})y(t-1) + \beta(t, q^{-1})u(t-1) \tag{3.6}$$

where

$$\alpha(t, q^{-1}) := [C(t, q^{-1}) - A(t, q^{-1})]q, \tag{3.7}$$

$$\beta(t, q^{-1}) := B(t, q^{-1})q. \tag{3.8}$$

The adaptive control algorithm which we propose to analyze is the following stochastic gradient minimum variance algorithm:

$$\text{A.1} \qquad \hat{\theta}(t) = \hat{\theta}(t-1) + \frac{\phi(t-1)}{r(t-2) + \phi(t-1)^T \phi(t-1)} e(t) \tag{3.9}$$

where $\hat{\theta}(t)$ is an estimate of $\theta(t)$ and $\theta^T(t) = (\alpha_1(t), \alpha_2(t), \cdots, \alpha_n(t), b_0(t), \cdots, b_m(t), c_1(t), \cdots, c_l(t))$. $\hat{\theta}(0)$ is given such that $\hat{\theta}_{n+1}(0) \neq 0$.

$$\text{A.2} \qquad r(t-1) = r(t-2) + \phi(t-1)^T \phi(t-1), \qquad r(0) > 0 \text{ given.} \tag{3.10}$$

$$\text{A.3} \qquad y^*(t) = \phi(t-1)^T \hat{\theta}(t-1).$$

$$\text{A.4} \qquad \phi(t-1)^T = (y(t-1), \cdots, y(t-\bar{n}), u(t-1), \cdots, u(t-\bar{n}), -\bar{y}(t-1), \cdots, -y(t-\bar{n})) \tag{3.11}$$

where $\bar{n}$ = upper bound on max $(n, m+1, l)$.

$$\text{A.5} \qquad \bar{y}(t) = \phi(t-1)^T \hat{\theta}(t).$$
$$\text{A.6} \qquad e(t) = y(t) - y^*(t). \tag{3.12}$$

This algorithm differs slightly from the time-invariant version in [3] by using a posteriori predictions. A discussion of the significance of this can be seen in [15].

The theoretical possibility of division by zero while evaluating $u(t)$ can be avoided since it can be argued inductively [3] that division by zero is a zero probability event. Since all the results in this paper are almost sure results, then division by zero can only affect the convergence on a set of measure zero. This argument depends on the above assumption of absolute continuity of the distribution functions. Hence, the algorithm is well-posed in the sense that all variables remain bounded in finite time (a.s.).

We then have the following global convergence result:

THEOREM 3.1. *Let Assumptions* N.1-N.3 *and* S.1-S.5 *hold for the system* (3.1) *and the algorithm* A.1-A.6. *Then with probability one, for any initial parameter estimate* $\hat{\theta}(0)$

$$\limsup_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} y(t)^2 < \infty, \tag{3.13}$$

$$\limsup_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} u(t)^2 < \infty, \tag{3.14}$$

$$\lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} [E\{(y(t)-y^*(t))^2 | \mathcal{F}_{t-1}\} - \sigma_t^2] = 0 \tag{3.15}$$

where σ_t^2 *is the minimum mean square control error achievable at time t by* $\mathcal{F}_{t-1}$ *measurable controls.*

Proof. We shall present an outline proof only, highlighting the key departures from the usual proofs for time-invariant systems as in [3], [15].

Set

$$\eta(t) = y(t) - \bar{y}(t). \tag{3.16}$$

We then have the following preliminary properties of the algorithm:

$$\text{P.1} \qquad \eta(t) = \frac{r(t-2)}{r(t-1)} e(t). \tag{3.17}$$

$$\text{P.2} \qquad \lim_{N\to\infty} \sum_{t=1}^{N} \frac{\phi(t-1)^T \phi(t-1)}{r(t-1)r(t-2)} < \infty. \tag{3.18}$$

$$\text{P.3} \qquad C(t, q^{-1}) z(t) = b(t) \tag{3.19}$$

where

$$z(t) = \eta(t) - \omega(t), \tag{3.20}$$

$$b(t) = -\phi(t-1)^T \tilde{\theta}(t), \tag{3.21}$$

$$\tilde{\theta}(t) = \hat{\theta}(t) - \theta(t). \tag{3.22}$$

$$\text{P.4} \qquad E\{b(t)\omega(t) | \mathcal{F}_{t-1}\} = -\frac{\phi(t-1)^T \phi(t-1)}{r(t-1)} \sigma_t^2. \tag{3.23}$$

The above properties are as in [3], [13].

Now subtracting $\theta(t)$ from both sides of (3.9) and using (3.10), (3.17), we have

$$\tilde{\theta}(t) = \tilde{\theta}(t-1) + \theta_e(t) + \frac{\phi(t-1)}{r(t-2)} \eta(t)$$

where

$$\theta_e(t) = \theta(t) - \theta(t-1),$$

$$\tilde{\theta}(t) - \frac{\phi(t-1)\eta(t)}{r(t-2)} = \tilde{\theta}(t-1) + \theta_e(t). \tag{3.24}$$

Define

$$V(t) = \tilde{\theta}(t)^T \tilde{\theta}(t).$$

Then squaring both sides of (3.24) and using (3.21) gives

$$V(t) + \frac{2b(t)\eta(t)}{r(t-2)} + \frac{\phi(t-1)^T\phi(t-1)}{r(t-2)^2}\eta^2(t) = V(t-1) + 2\tilde{\theta}(t-1)^T\theta_e(t) + \|\theta_e(t)\|^2.$$

Hence using (3.20), (3.23) we have

$$\begin{aligned} E\{V(t)|\mathcal{F}_{t-1}\} &= V(t-1) - \frac{2}{r(t-2)} E\{b(t)z(t)|\mathcal{F}_{t-1}\} \\ &\quad + \frac{2\phi(t-1)^T\phi(t-1)}{r(t-2)r(t-1)}\sigma_t^2 - E\left\{\frac{\phi(t-1)^T\phi(t-1)}{r(t-2)^2}\eta(t)^2|\mathcal{F}_{t-1}\right\} \\ &\quad + \|\theta_e(t)\|^2 + 2\tilde{\theta}(t-1)^T\theta_e(t). \end{aligned} \tag{3.25}$$

From Assumption S.3, there exists a G, λ with $0 < G < \infty$, $|\lambda| < 1$ such that

$$\|\theta_e(t)\|^2 \leqq G|\lambda|^t. \tag{3.26}$$

Also, for $0 < a(t) < \infty$, we have

$$2[\tilde{\theta}(t-1)^T\theta_e(t)] \leqq a(t)^2\|\tilde{\theta}(t-1)\|^2 + \frac{1}{a(t)^2}\|\theta_e(t)\|^2.$$

Thus selecting $a(t)^2 = |\lambda|^{t/2}$ and using (3.26), we have

$$\begin{aligned} 2[\tilde{\theta}(t-1)^T\theta_e(t)] &\leqq |\lambda|^{t/2}\|\tilde{\theta}(t-1)\|^2 + G|\lambda|^{t/2} \\ &= |\lambda|^{t/2}V(t-1) + G|\lambda|^{t/2}. \end{aligned} \tag{3.27}$$

Substituting (3.27) into (3.25) gives

$$\begin{aligned} E\{V(t)|\mathcal{F}_{t-1}\} &\leqq V(t-1)[1 + |\lambda|^{t/2}] - \frac{2}{r(t-2)} E\{b(t)z(t)|\mathcal{F}_{t-1}\} \\ &\quad + \frac{2\phi(t-1)^T\phi(t-1)}{r(t-1)r(t-2)}\sigma_t^2 - E\left\{\frac{\phi(t-1)^T\phi(t-1)}{r(t-2)^2}\eta(t)^2\,\middle|\,\mathcal{F}_{t-1}\right\} + 2G|\lambda|^{t/2}. \end{aligned} \tag{3.28}$$

Define

$$S(t) := 2\sum_{j=n_0}^{t}[b(j)z(j) - \delta z(j)^2] + K, \tag{3.29}$$

$$X(t) := V(t) + \frac{S(t)}{r(t-2)} + 2\delta\sum_{j=n_0}^{t}\frac{z(j)^2}{r(j-2)} + \sum_{j=n_0}^{t}\frac{\phi(j-1)^T\phi(j-1)}{r(j-2)^2}\eta(j)^2. \tag{3.30}$$

From Assumptions S.5, S.3, Property P.3 and Theorem 2.1, we know that there exist a n_0, $\delta > 0$ and a K (depending on the conditions at n_0) such that $S(t) \geqq 0$ for all $t \geqq n_0$. Under these conditions $X(t) \geqq 0$.

It is readily seen using (3.28), (3.30) that

$$E\{X(t)|\mathcal{F}_{t-1}\} \leqq X(t-1)[1+|\lambda|^{t/2}] + \frac{2\phi(t-1)^T\phi(t-1)}{r(t-2)r(t-1)}\sigma_t^2 + 2G|\lambda|^{t/2} \quad \text{for } t \geqq n_0.$$

From Property P.2, and Assumption N.2, we have that

$$\sum_{t=0}^{\infty}\left[\frac{\phi(t-1)^T\phi(t-1)}{r(t-2)r(t-1)}\sigma_t^2 + G|\lambda|^{t/2}\right] < \infty.$$

Thus we can apply the martingale convergence theorem (Appendix B) to conclude

$$X(t) \to X < \infty \quad \text{a.s.} \tag{3.31}$$

Using (3.30), we see that

$$\lim_{N\to\infty}\sum_{t=1}^{N}\frac{z(t)}{r(t-2)} < \infty \quad \text{a.s.}, \tag{3.32}$$

$$\lim_{N\to\infty}\sum_{t=1}^{N}\frac{\phi(t-1)^T\phi(t-1)}{r(t-2)^2}\eta(t)^2 < \infty \quad \text{a.s.} \tag{3.33}$$

The lower summation limits in (3.32), (3.33) can be extended from n_0 to 1 because the algorithm ensures all variables remain bounded in finite time (a.s.).

A simple argument by contradiction can now be used to conclude (3.13) to (3.15) using (3.32) together with Assumptions S.3, S.4 and Theorem A.1. The steps are exactly as in [3] and as explained in general in [15]. □

4. Adaptive control with general state space model. Consider a linear finite dimensional system described by the following time-invariant state space model:

$$x(t+1) = Fx(t) + Gu(t) + v_1(t), \tag{4.1}$$

$$y(t) = Hx(t) + v_2(t) \tag{4.2}$$

where $\{v_1(t)\}$, $\{v_2(t)\}$ are zero mean Gaussian white noise sequences satisfying:

$$E\{v_1(t)v_1(t)^T\} = Q = DD^T \geqq 0, \tag{4.3}$$

$$E\{v_2(t)v_2(t)^T\} = R \geqq 0. \tag{4.4}$$

The initial state $x(0)$ is also assumed to have a Gaussian distribution with mean $\bar{x}_0$ and covariance P_0. We make the following assumptions:

S.S.1: (H, F) is observable.

S.S.2(a): (F, D) has no uncontrollable modes on the unit circle and $P_0 > 0$

or

(b): (F, D) is stabilizable and $P_0 \geqq 0$.

Using standard Kalman filtering ideas [12], the innovations model for (4.1), (4.2) is

$$\hat{x}(t+1) = F\hat{x}(t) + Gu(t) + K(t)\omega(t), \qquad \hat{x}(0) = \bar{x}_0, \tag{4.5}$$

$$y(t) = H\hat{x}(t) + \omega(t). \tag{4.6}$$

Here $K(t)$ is obtained from the solution of the following matrix Riccati equation:

$$\Sigma(t+1) = F\Sigma(t)F^T - F\Sigma(t)H^T(H\Sigma(t)H^T + R)^{-1}H\Sigma(t)F^T + Q, \tag{4.7}$$

$$\Sigma(0) = P_0, \tag{4.8}$$

$$K(t) = F\Sigma(t)H^T(H\Sigma(t)H^T + R)^{-1}. \tag{4.9}$$

In view of Assumption S.S.1, we can transform the system state such that (H, F) are in observer canonical form, i.e. (4.5), (4.6) can be written as

$$\hat{x}(t+1)=\begin{bmatrix}-a_1 & 1 & & 0\\ \vdots & & \ddots & \\ \vdots & & & 1\\ -a_n & 0 & \cdots & 0\end{bmatrix}\hat{x}(t)+\begin{bmatrix}b_1\\ \vdots\\ \vdots\\ b_n\end{bmatrix}u(t)+\begin{bmatrix}k_1(t)\\ \vdots\\ \vdots\\ k_n(t)\end{bmatrix}\omega(t), \tag{4.10}$$

$$y(t)=[1 \quad 0 \quad \cdots \quad 0]\hat{x}(t)+\omega(t). \tag{4.11}$$

Using (4.11) in (4.10) gives the following *time varying* ARMAX model:

$$A(q^{-1})y(t)=B(q^{-1})u(t)+C(t, q^{-1})\omega(t) \tag{4.12}$$

where

$$\begin{aligned}A(q^{-1})&=1+a_1q^{-1}+\cdots+a_nq^{-n},\\ B(q^{-1})&=b_1q^{-1}+\cdots+b_nq^{-n},\\ C(t, q^{-1})&=1+(k_1(t-1)+a_1)q^{-1}+\cdots+(k_n(t-n)+a_n)q^{-n}.\end{aligned} \tag{4.13}$$

It is known [13] that Assumptions S.S.1, S.S.2 are sufficient to ensure:

$$K(t)\to\bar{K} \quad \text{exponentially fast} \tag{4.14}$$

and

$$C(q^{-1})=1+(\bar{k}+a_1)q^{-1}+\cdots+(\bar{k}_n+a_n)q^{-n} \text{ is asymptotically stable.} \tag{4.15}$$

We make the following additional assumptions:

S.S.3: $b_1\neq 0$ (corresponding to $d=1$ in § 3).

S.S.4: An upper bound for n is known. (As in S.2, this is an assumption on the data supplied to the algorithm.)

S.S.5: The system

$$C(q^{-1})z(t)=b(t)$$

is input strictly passive.

S.S.6: $B(z)$ has all zeros outside the unit circle and $b_1\neq 0$ (the latter for simplicity only).

We then have the following elementary corollary to Theorem 3.1:

COROLLARY 4.1. *Let Assumptions* S.S.1-S.S.6 *hold for the system* (4.1), (4.2) *and the algorithm* A.1-A.6. *Then with probability one, for any initial parameter estimate* $\hat{\theta}(0)$

$$\limsup_{N\to\infty}\frac{1}{N}\sum_{t=1}^{N}y(t)^2<\infty, \tag{4.16}$$

$$\limsup_{N\to\infty}\frac{1}{N}\sum_{t=1}^{N}u(t)^2<\infty, \tag{4.17}$$

$$\lim_{N\to\infty}\frac{1}{N}\sum_{t=1}^{N}E\{(y(t)-y^*(t))^2|\mathcal{F}_{t-1}\}=\sigma^2 \tag{4.18}$$

where

$$\sigma^2=H\bar{\Sigma}H^T+R \tag{4.19}$$

and $\bar{\Sigma}$ *is the steady state solution of* (4.7).

Proof. The result is immediate from Theorem 3.1 on noting that $\{\omega(t)\}$ is a Gaussian innovations sequence and therefore satisfies N.1–N.3. Also, $\sigma_t^2 \to \sigma^2$ exponentially fast [13]. Hence

$$\lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} \sigma_t^2 = \sigma^2. \tag{4.20}$$

□

Remarks. 1. It should be pointed out that this result can be obtained directly from the corresponding result for the steady-state ARMAX model [3], [15]. Firstly, note that the covariance P_0 for the Kalman filter (with $\bar{x}_0$) defines a Gaussian distribution for x_0. So use of the steady-state gain $\bar{K}$ corresponds to a particular choice of the initial condition distribution. Since the martingale convergence theorem leads to a sample path convergence result, modification of the initial state distribution does not affect this result. The proof for this comes by noting that once a.s. convergence has been established with respect to one distribution, then it is also true for any other distribution which is absolutely continuous with respect to the original one [18]. Thus, having proved global convergence for the distribution corresponding to a steady-state Kalman filter, it also holds for other distributions corresponding to the non-steady-state case. However, it should be realized that the original martingale properties will no longer apply.

2. If a general initial condition distribution is used and one replaces (4.12) by the corresponding steady state ARMAX model, then the prediction error so defined will not satisfy N.1.

3. The result in Corollary 4.1 also applies to degenerate distributions, i.e. when the initial state is exactly known. In this case the argument in Remark 1 above cannot be used and the more complicated machinery of § 3 is necessary to deal with this case.

5. Conclusions. This paper has analyzed a robustness property of the discrete time stochastic adaptive control algorithm based on gradient estimation and minimum variance control. The algorithm is shown to be globally convergent when the system parameters are exponentially convergent to values satisfying the conditions for a globally convergent time-invariant system. This result is applied to the special case where the time variation is derived from a non-steady-state Kalman filter.

Appendix A—Properties of convergent linear systems. We present some properties of time-varying linear systems which are convergent toward an asymptotically stable system.

We consider the following time-varying system:

$$x(t+1) = A(t)x(t) + B(t)u(t), \tag{A.1}$$

$$y(t) = C(t)x(t) + D(t)u(t) \tag{A.2}$$

and we introduce the notation:

$$\Phi(n, n_0) = \prod_{k=n_0}^{n-1} A(k) \tag{A.3}$$

for the state transition matrix.

We will use $|\cdot|$ for the Euclidean norm and $\|\cdot\|$ for the l_2 norm (similarly for the induced norms).

LEMMA A.1. *Let A be asymptotically stable. Suppose $A(k) \to A$. Then there exist $\bar{n}$, $v > 0$, and $0 < \beta < 1$ such that*

$$|\Phi(n, n_0)| \leqq v\beta^{n-n_0} \tag{A.4}$$

for all $n > n_0 \geqq \bar{n}$.

Proof. Since A is asymptotically stable, $\exists P>0$ such that

$$P-A^TPA=Q \quad \text{with } Q>0. \tag{A.5}$$

Now consider the autonomous time-varying system

$$x(k+1)=A(k)x(k) \tag{A.6}$$

and try the Lyapunov function

$$V(x)=x^TPx. \tag{A.7}$$

Then

$$\begin{aligned} V(x(k+1))-V(x(k)) &= x(k)^T[A^T(k)PA(k)-P]x(k) \\ &:= -x(k)^TQ(k)x(k). \end{aligned} \tag{A.8}$$

Now since $Q(k)\to Q>0$, then $\exists \bar{n}$ such that for $k\geqq \bar{n}$, we have some μ such that

$$V(x(k+1))-V(x(k))\leqq -\mu|x(k)|^2. \tag{A.9}$$

Introducing $|x|_P=(x^TPx)^{1/2}$, we have

$$M|x|_P\leqq |x| \quad \text{where } M=(1/\lambda_{\max}P)^{1/2}. \tag{A.10}$$

Hence from (A.9)

$$\begin{aligned} |x(k+1)|_P^2 &\leqq x(k)^TPx(k)-\mu|x(k)|^2 \\ &\leqq (1-\mu M^2)|x(k)|_P^2 \quad \text{for all } x(k). \end{aligned}$$

Thus

$$|A(k)|_P\leqq (1-\mu M^2):=\beta<1.$$

Hence for $n>n_0\geqq \bar{n}$

$$|\Phi(n,n_0)|_P=\left|\prod_{k=n_0}^{n-1}A(k)\right|_P\leqq \prod_{k=n_0}^{n-1}|A(k)|_P\leqq \beta^{n-n_0}.$$

So

$$|\Phi(n,n_0)|\leqq v\beta^{n-n_0} \quad \text{for all } n>n_0\geqq \bar{n}$$

where

$$v=(1/\lambda_{\min}P)^{1/2}. \qquad \square$$

The authors suspect that Lemma 2.1 has been established elsewhere though they have no specific reference. Fuchs [14] states a related (but differing in detail) result. We can now immediately establish:

THEOREM A.1. *Consider the system* (2.1), (2.2). *Then provided* $A(k) \to A$ *with* A *asymptotically stable and* $B(k)$, $C(k)$, $D(k)$ *are bounded*:

(a) *There exist* $\bar{n}$, $0 < K_1 < \infty$, $0 \leqq K_2 < \infty$ *independent of* N *such that*

$$\sum_{t=n_0}^{N} |y(t)|^2 \leqq K_1 \sum_{t=n_0}^{N} |u(t)|^2 + K_2 \quad \text{for all } N \geqq n_0 \geqq \bar{n}. \tag{A.11}$$

(b) *There exist* $0 \leqq m_3 \leqq \infty$, $0 \leqq m_4 < \infty$ *which are independent of* t *such that*

$$|y_i(t)| \leqq m_3 + m_4 \max_{n_0 \leqq \tau \leqq N} |u(\tau)| \quad \text{for all } N \geqq t \geqq n_0. \tag{A.12}$$

Proof. The proof mimics the corresponding proof for the time-invariant case given elsewhere (see for example [15, Appendix B]). □

In the sequel, we use the notation $\|x(t)\|_{n_0}^{N} := (\sum_{t=n_0}^{N} |x(t)|^2)^{1/2}$. If the sum is finite for all $N \geqq n_0$, we say $x(t) \in l_{2e}(\mathbb{Z}_+)$ where $\mathbb{Z}_+$ is the set of integers $n_0, n_0+1, \cdots$.

LEMMA A.2. *Consider a matrix* A *and sequence* $A(k)$ *as in Lemma* A.1. *Suppose* $A(k) \to A$ *exponentially. Then there exist* $\bar{n}$, $\chi > 0$, *and* $0 < \eta < 1$ *such that*

$$\left| A^{n-n_0} - \prod_{k=n_0}^{n-1} A(k) \right| \leqq \chi \eta^{n-1} \tag{A.13}$$

for all $n > n_0 \geqq \bar{n}$.

Proof. Let $K(k) := A - A(k)$. We find the following observation useful

$$\begin{aligned} A^{n-n_0} - \prod_{k=n_0}^{n-1} A(k) &= AA^{n-n_0-1} - (A - K(n-1)) \prod_{k=n_0}^{n-2} A(k) \\ &= A\left(A^{n-n_0-1} - \prod_{k=n_0}^{n-2} A(k) \right) + K(n-1) \prod_{k=n_0}^{n-2} A(k). \end{aligned}$$

Repeated application of this gives

$$\begin{aligned} A^{n-n_0} - \prod_{k=n_0}^{n-1} A(k) &= A^{n-n_0-1} K(n_0) + A^{n-n_0-2} K(n_0+1) A(n_0) \\ &\quad + A^{n-n_0-3} K(n_0+2) \prod_{k=n_0}^{n_0+1} A(k) \\ &\quad + \cdots + AK(n-2) \prod_{k=n_0}^{n-3} A(k) + K(n-1) \prod_{k=n_0}^{n-2} A(k). \end{aligned}$$

Now since A is stable, $\exists \gamma > 0$, $0 < \xi < 1$ such that

$$|A^k| \leqq \gamma \xi^k.$$

Lemma A.1 gives that for $n > n_0 \geqq \bar{n} \exists v > 0$, $0 < \beta < 1$ such that

$$\left| \prod_{k=n_0}^{n-1} A(k) \right| \leqq v \beta^{n-n_0}.$$

Further, exponential convergence of $A(k)$ to A gives that $\exists \Omega > 0$, $0 < \lambda < 1$ such that

$$|K(k)| \leqq \Omega \lambda^k.$$

Let

$$\psi := \max(\gamma, v, \Omega), \qquad \eta := \max(\xi, \beta, \lambda).$$

Then

$$\begin{aligned}\left|A^{n-n_0}-\prod_{k=n_0}^{n-1}A(k)\right| &\leqq |K(n_0)||A^{n-n_0-1}|+|A(n_0)||K(n_0+1)||A^{n-n_0-2}| \\ &\quad +\left|\prod_{k=n_0}^{n_0+1}A(k)\right||K(n_0+2)||A^{n-n_0-3}| \\ &\quad +\cdots+\left|\prod_{k=n_0}^{n-3}A(k)\right||K(n-2)||A| \\ &\quad +\left|\prod_{k=n_0}^{n-2}A(k)\right||K(n-1)| \\ &\leqq \psi^2\eta^{n-1}(1+\psi\eta+\psi\eta^2+\cdots+\psi\eta^{n-n_0-2}+\eta^{n-n_0-1}) \\ &\qquad \text{using the above exponential bounds} \\ &= \psi^2\eta^{n-1}\left(1+\eta^{n-n_0-1}+\psi\eta\sum_{k=0}^{n-n_0-3}\eta^k\right) \\ &\qquad \text{assuming that } n-n_0\geqq 3 \\ &\leqq \psi^2\eta^{n-1}\left(2+\frac{\psi\eta}{1-\eta}\right).\end{aligned}$$

Let

$$\chi := \phi^2\left(2+\frac{\psi\eta}{1-\eta}\right).$$

Then we have the required inequality. (It is easy to see that the result holds for $n>n_0$.) □

Appendix B—Modified martingale convergence theorem. Let $\{X(t)\}$ be a sequence of nonnegative random variables adapted to an increasing sequence of sub σ-algebras $\{\mathcal{F}_t\}$. If

$$E\{X(t+1)|\mathcal{F}_t\}\leqq(1+\gamma(t))X(t)-\alpha(t)+\beta(t) \quad \text{a.s.}$$

where $\alpha(t)\geqq 0$, $\beta(t)\geqq 0$ and $E\{X(0)\}<\infty$, $\sum_{j=1}^{\infty}|\gamma(t)|<\infty$, $\sum_{j=0}^{\infty}\beta(j)<\infty$ a.s. then $X(t)$ converges almost surely to a finite random variable and

$$\lim_{N\to\infty}\sum_{t=0}^{N}\alpha(t)<\infty \quad \text{a.s.}$$

Proof. See Neveu [16]. □

The above theorem was first employed in a stochastic adaptive control proof in [10], [11]. Other applications are given in [15].

REFERENCES

[1] K. J. ASTROM, *Theory and applications of adaptive control—A survey*, Automatica, 19 (1983), pp. 471-487.

[2] C. S. DRAPER AND Y. T. LI, *Principles of Optimalizing Control Systems and Application to Internal Combustion Engines*, ASME, New York, 1951.

[3] G. C. GOODWIN, P. J. RAMADGE AND P. E. CAINES, *Discrete time stochastic adaptive control*, this Journal, 19 (1981), pp. 829-853.

[4] H. F. CHEN AND P. E. CAINES, *The strong consistency of the stochastic gradient algorithm of adaptive control*, Technical Reports, McGill Univ., Montreal, Canada, July, 1983.
[5] P. R. KUMAR, *Adaptive control with the stochastic approximation algorithm: Geometry and convergence*, IEEE Trans. Automat. Control, to appear.
[6] K. S. SIN AND G. C. GOODWIN, *Stochastic adaptive control using a modified least squares algorithm*, Automatica, 18 (1983), pp. 315-321.
[7] H. F. CHEN, *Recursive system identification and adaptive control by use of the modified least squares algorithm*, this Journal, 22 (1984), pp. 758-776.
[8] P. E. CAINES, *Stochastic adaptive control: Randomly varying parameters and continuously disturbed controls*, IFAC Congress, Kyoto, Japan, August 1981.
[9] H. F. CHEN AND P. E. CAINES, *On the adaptive control of stochastic systems with random parameters*, Proc. 23rd Conference on Decision and Control, Las Vegas, NV, December 1984, pp. 33-38.
[10] M. A. HERSH AND M. B. ZARROP, *Stochastic adaptive control of nonminimum phase systems*, Control Systems Centre Technical Report, UMIST, United Kingdom, 1981.
[11] M. A. HERSH, Ph.D. thesis, Control Systems Centre, UMIST, United Kingdom, 1983.
[12] B. D. O. ANDERSON AND J. B. MOORE, *Optimal Filtering*, Prentice-Hall, Englewood Cliffs, NJ, 1979.
[13] S. W. CHAN, G. C. GOODWIN AND K. S. SIN, *Convergence properties of the Riccati difference equation in optimal filtering of non-stabilizable systems*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 110-118.
[14] J. J. FUCHS, *On the good use of the spectral radius of a matrix*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 1134-1135.
[15] G. C. GOODWIN AND K. S. SIN, *Adaptive Filtering Prediction and Control*, Prentice-Hall, Englewood Cliffs, NJ, 1984.
[16] J. NEVEU, *Discrete Parameter Martingales*, North-Holland, Amsterdam, 1975.
[17] G. C. GOODWIN, P. J. RAMADGE AND P. E. CAINES, *Discrete-time multivariable adaptive control*, IEEE Trans. Automat. Control, AC-25 (1980), pp. 449-456.
[18] M. LOEVE, *Probability Theory*, Springer-Verlag, New York, 1963.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 4, July 1986

002

QUASI-VARIATIONAL INEQUALITIES AND ERGODIC IMPULSE CONTROL*

P. L. LIONS† AND B. PERTHAME‡

Abstract. In this paper, we solve the general ergodic control problem for reflected diffusion processes and the associated quasi-variational inequalities. Our method relies on some a priori estimates for the solutions of quasi-variational inequalities where the "discount factor" (i.e., the zero order term) is set to 0.

Key words. quasi-variational inequalities, reflected diffusion processes, impulse control, variational inequalities, Neumann conditions, ergodic control

1. Introduction. It is well known that impulse control problems for reflected diffusion processes may be solved by considering the solution of quasi-variational inequalities with Neumann boundary conditions (we refer to A. Bensoussan and J. L. Lions [3], A. Bensoussan [1] for more details). A typical example of such a quasi-variational inequality (QVI in short) is the following

$$\int_\Omega \nabla u_\alpha \cdot \nabla(v-u_\alpha)\,dx+\alpha\int_\Omega u_\alpha(v-u_\alpha)\,dx \geqq \int_\Omega f(v-u_\alpha)\,dx, \tag{1}$$
$$\forall v\in H^1(\Omega),\quad v\leqq Mu_\alpha \text{ a.e.},\quad u_\alpha\in H^1(\Omega),\quad u_\alpha\leqq Mu_\alpha \text{ a.e.}$$

where Ω is a given bounded smooth open set in $\mathbb{R}^n$, $\alpha>0$, f is a given function and M is an operator defined (for example) on $C(\bar\Omega)$ by

$$Mu=k+\inf\{c_0(\xi)+u(x+\xi)/\xi\geqq 0,\ x+\xi\in\bar\Omega\}, \tag{2}$$

where $k>0$, c_0 is continuous, subadditive, nonnegative and $c_0(0)=0$. Of course $\xi\geqq 0$ means that all the coordinates of ξ are nonnegative. We will assume (see below for a more precise assumption) that M maps $C(\bar\Omega)$ into $C(\bar\Omega)$.

In this particular situation, our main result shows that: denoting by $\langle\varphi\rangle=\text{meas}\,(\Omega)^{-1}\int_\Omega v\,dx$, $(\alpha u_\alpha, u_\alpha-\langle u_\alpha\rangle)$ converge uniformly on $\bar\Omega$ to the unique solution (λ, v_0) of

$$\int_\Omega \nabla u_0\cdot\nabla(v-v_0)\,dx\geqq\int_\Omega (f-\lambda)(v-v_0)\,dx, \tag{3}$$
$$\forall v\in H^1(\Omega),\quad v\leqq Mv_0 \text{ a.e.},\ \lambda\in\mathbb{R},\quad v_0\in H^1(\Omega)\cap C(\bar\Omega),\quad \langle v_0\rangle=0,\quad v_0\leqq Mv_0 \text{ in } \bar\Omega.$$

This result is stated in § 2; the existence part is proved in §§ 3 and 4: in § 3 we obtain H^1 bounds on $u_\alpha-\langle u_\alpha\rangle$ while in § 4 we prove that $u_\alpha-\langle u_\alpha\rangle$ are equicontinuous on $\bar\Omega$. The uniqueness is proved in § 5 and extensions to general second-order operators and general oblique derivative conditions are given in § 6. Finally the interpretation of these results in terms of ergodic impulse control is presented in § 7. Some technical results are collected in the appendix.

The above result (and the extensions obtained in this paper) answers a longstanding question concerning ergodic impulse control theory. And even if we borrow some techniques used in other ergodic problems by P. L. Lions [12], F. Gimbert [7], the specificity of the implicit obstacle Mu requires a particular treatment. We would like to mention a few other works on ergodic stochastic control problems: J. M. Lasry [10];

* Received by the editors September 20, 1984, and in revised form March 29, 1985.
† Ceremade, Université Paris IX, 75775 Paris Cedex 16, France.
‡ Centre de Mathématiques Appliquèes, E.N.S. Ulm, 45, 75230 Paris Cedex 05, France.

M. Robin [18], [19]; A. Bensoussan [2]; F. Gimbert [7]; P. L. Lions [12]; J. M. Lasry and P. L. Lions [11]; I. Capuzzo-Dolcetta and M. G. Garonni [5]...

2. The main result. We consider $u_\alpha \in H^1(\Omega) \cap C(\bar{\Omega})$, the solution of (1) (the existence and uniqueness is deduced from the results of [3], Hanouzet and Joly [8]). Recall that we assume:

$$\forall u \in C(\bar{\Omega}), \quad Mu \in C(\bar{\Omega}); \tag{4}$$

this is the case if Ω is convex or more generally (cf. Perthame [16], [17]) if Ω satisfies:

$$\forall x \in \bar{\Omega}, \quad \{\xi / \xi \geqq 0;\ \xi \neq 0;\ x+\xi \in \partial\Omega;\ \exists \varepsilon > 0, \forall y \geqq 0, \quad x+y \notin \Omega \text{ if } |y-\xi| \leqq \varepsilon\} = \varnothing. \tag{5}$$

We will assume:

$$f \in L^p(\Omega) \quad \text{for some } p > \frac{N}{2} \text{ if } N \geqq 2, \quad p \geqq 1 \text{ if } N = 1. \tag{6}$$

THEOREM 1. *We assume* (5) *and* (6). *Then* αu_α *converges uniformly on* $\bar{\Omega}$ *to some constant* λ, *and* $u_\alpha - \langle u_\alpha \rangle$ *converges uniformly on* $\bar{\Omega}$ *and strongly in* $H^1(\Omega)$ *to* v_0. *Finally* (λ, v_0) *is the unique solution of the system* (3).

Remark. We give below extensions of this result and its stochastic interpretation.

The uniqueness of (λ, v_0) is proved in § 5; we prove in §§ 3-4 that $v_\alpha = u_\alpha - \langle u_\alpha \rangle$ is bounded in $H^1(\Omega)$ and relatively compact in $C(\bar{\Omega})$. If we know that $\alpha \langle u_\alpha \rangle$ is bounded, the convergence of $(\alpha u_\alpha, v_\alpha)$ is then easily deduced from these bounds.

We prove next that αu_α is bounded in $L^\infty(\Omega)$. First of all let us mention that this is an obvious consequence of comparison arguments if $f \in L^\infty(\Omega)$; indeed in this case we have:

$$\alpha \|u_\alpha\|_{L^\infty} \leqq \|f\|_{L^\infty}.$$

In the general case (i.e. (6)) we need to find a supersolution $\bar{u}_\alpha$ and a subsolution $\underline{u}_\alpha$ of the QVI (1) such that $\alpha \bar{u}_\alpha$, $\alpha \underline{u}_\alpha$ are bounded in $L^\infty(\Omega)$. For the supersolution, we just need to take $\bar{u}_\alpha$, the solution of

$$-\Delta \bar{u}_\alpha + \alpha \bar{u}_\alpha = f^+ \text{ in } \Omega, \quad \frac{\partial \bar{u}_\alpha}{\partial n} = 0 \text{ on } \Gamma, \quad \bar{u}_\alpha \in W^{2,p}(\Omega)$$

(where n denotes the unit outward normal to $\partial\Omega = \Gamma$). For the subsolution we first consider v_α^m, the solution of

$$-\Delta v_\alpha^m + \alpha v_\alpha^m = f \wedge (-m) \text{ in } \Omega, \quad \frac{\partial v_\alpha^m}{\partial n} = 0 \text{ on } \Gamma, \quad v_\alpha^m \in W^{2,p}(\Omega).$$

As α goes to 0, it is well known that $v_\alpha^m - \langle v_\alpha^m \rangle$ goes uniformly to v^m, the solution of

$$-\Delta v^m = f \wedge (-m) - \langle f \wedge (-m) \rangle \quad \text{in } \Omega,$$

$$\langle v^m \rangle = 0, \quad \frac{\partial v^m}{\partial n} = 0 \text{ on } \Gamma, \quad v^m \in W^{2,p}(\Omega).$$

Remarking that v^m converges uniformly to 0 as m goes to ∞, we choose m such that $\|v^m\|_{L^\infty} \leqq k/4$. Hence for α small enough

$$|v_\alpha^m(x) - v_\alpha^m(y)| \leqq k \quad \forall x, y \in \bar{\Omega},$$

and thus $v_\alpha^m \leqq M v_\alpha^m$ on $\bar{\Omega}$; and we may take $\underline{u}_\alpha = v_\alpha^m$ for α small (since $v_\alpha^m \leqq 0$, we may

take $\underline{u}_\alpha = v_{\alpha_0}^m$ for $\alpha \geqq \alpha_0 \cdots$). In view of standard results on linear equations $\alpha \underline{u}_\alpha$ is bounded in $L^\infty(\Omega)$, and we conclude since we have

$$\underline{u}_\alpha \leqq u_\alpha \leqq \bar{u}_\alpha.$$

In what follows we set $\lambda_\alpha = \alpha\langle u_\alpha \rangle$. λ_α is bounded, and $v_\alpha = u_\alpha - \langle u_\alpha \rangle$ is the solution of the following QVI

$$\begin{aligned} &a(v_\alpha, v - v_\alpha) + \alpha(v_\alpha, v - v_\alpha) \geqq (f_\alpha, v - v_\alpha) \quad \forall v \in H^1(\Omega), \quad v \leqq M v_\alpha \text{ a.e.}, \\ &v_\alpha \in C(\bar{\Omega}) \cap H^1(\Omega), \quad v_\alpha \leqq M v_\alpha \text{ in } \bar{\Omega} \end{aligned} \tag{7}$$

where (f, g) denotes the L^2 scalar product, $a(\varphi, \psi) = \int_\Omega \nabla\varphi \cdot \nabla\psi \, dx$, $f_\alpha = f - \lambda_\alpha$.

3. Boundedness of v_α in $H^1(\Omega)$. The bounds in $H^1(\Omega)$ rely on the following crucial lemma.

LEMMA 2. *Let $m_\alpha = \min_{\bar{\Omega}} v_\alpha$; then there exists $C \geqq 0$ independent of $\alpha \in]0, 1[$ such that*

$$|m_\alpha| \leqq C \| v_\alpha^- \|_{L^1(\Omega)} + C. \tag{8}$$

Using this lemma (which is proved below), it is now easy to prove the H^1 bounds: indeed using $v \equiv m_\alpha$ as a test function in (7), we find

$$\begin{aligned} \int_\Omega |Dv_\alpha|^2 + \alpha v_\alpha^2 \, dx &\leqq \int_\Omega f_\alpha (v_\alpha - m_\alpha) + \alpha v_\alpha m_\alpha \, dx \\ &\leqq C m_\alpha + C \| v_\alpha \|_{H^1(\Omega)} \leqq C \| v_\alpha \|_{H^1(\Omega)} + C \quad \text{by (8).} \end{aligned}$$

Recalling that $\langle v_\alpha \rangle = 0$, we conclude using the Poincaré inequality.

We now turn to the proof of Lemma 2: we consider the sets

$$\begin{aligned} E &= \{x \in \bar{\Omega}, \, v_\alpha(x) < m_\alpha + k\}, \\ F &= \{x \in \bar{\Omega}, \, v_\alpha(x) < m_\alpha + 2k + C_0\}, \end{aligned}$$

where $C_0 = \sup \{c_0(\xi), \, \xi \geqq 0, \, |\xi| \leqq \operatorname{diam}(\bar{\Omega})\}$.

Since we may assume without loss of generality that $|m_\alpha| = -m_\alpha$ is as large as we wish, (8) will be proved if we show that

$$\operatorname{meas}(F) \geqq \gamma > 0 \tag{9}$$

for some γ independent of $\alpha \in]0, 1[$.

Letting $\delta > 0$, we set $\Gamma_\delta = \{x \in \bar{\Omega}, \operatorname{dist}(x, \Gamma) \leqq \delta\}$. We may assume that E is contained in Γ_δ. Indeed if this were not the case, there would exist $x_0 \in E \cap (\Omega - \Gamma_\delta)$ and for all $x \in \bar{\Omega}$, $x \leqq x_0$ we would have

$$v_\alpha(x) \leqq M v_\alpha(x) \leqq k + c_0(x_0 - x) + v_\alpha(x_0) < m_\alpha + 2k + C_0,$$

i.e., $\{x \in \bar{\Omega}, \, x \leqq x_0\} \subset F$, proving (9) since $x_0 \notin \Gamma_\delta$.

Now if $E \subset \Gamma_\delta$, we observe that on the open set E, $v_\alpha < M v_\alpha$ and thus we obtain easily

$$\begin{aligned} &-\Delta v_\alpha + \alpha v_\alpha = f_\alpha \quad \text{in } E, \qquad \frac{\partial v_\alpha}{\partial n} = 0 \quad \text{on } \partial E \cap \partial\Omega = \Gamma_1', \\ &v_\alpha = m_\alpha + k \quad \text{on } \partial E \cap \Omega, \qquad v_\alpha \in W^{2,p}_{\text{loc}}(E \cup \Gamma_1') \cap C(\bar{E}). \end{aligned}$$

Let $h \in L^p(\Omega)$ satisfy $h \leqq f_\alpha - \alpha k - \alpha m_\alpha$, $\forall \alpha \in]0, 1[$ and let z_α be the solution of

$$-\Delta z_\alpha + \alpha z_\alpha = h \quad \text{in } \Gamma_\delta, \qquad \frac{\partial z_\alpha}{\partial n} = 0 \quad \text{on } \partial\Omega,$$

$$z_\alpha = 0 \quad \text{on } \partial\Gamma_\delta - \partial\Omega, \qquad z_\alpha \in W^{2,p}(\Gamma_\delta).$$

We have

$$\|z_\alpha\|_{L^\infty(\Gamma_\delta)} \leqq C \|h\|_{L^p(\Gamma_\delta)} \tag{10}$$

where C does not depend on α or δ. Thus $|z_\alpha| \leqq k$ for δ small enough and, by the maximum principle, $z_\alpha \leqq v_\alpha - k - m_\alpha$ on E. At the minimum point of v_α we thus have $z_\alpha \leqq -k$ and this contradicts (10). This shows that for δ small enough E cannot be contained in Γ_δ and thus (9) and (8) are proved.

4. Compactness in $C(\bar{\Omega})$. We first show that v_α is bounded in $L^\infty(\Omega)$: notice that we already know from (8) and the H^1 bound that v_α is bounded from below. To show that v_α is bounded from above, we observe that (7) yields

$$a(v_\alpha, \varphi) + (v_\alpha, \varphi) \leqq (f_\alpha + (1-\alpha)v_\alpha, \varphi) \quad \forall \varphi \in H^1(\Omega), \quad \varphi \geqq 0. \tag{11}$$

Hence $v_\alpha \leqq \bar{v}_\alpha^1$, where $\bar{v}_\alpha^1$ is the solution of

$$-\Delta \bar{v}_\alpha^1 + \bar{v}_\alpha^1 = f_\alpha^+ + (1-\alpha)v_\alpha^+ \text{ in } \Omega, \quad \frac{\partial \bar{v}_\alpha^1}{\partial u} = 0 \quad \text{on } \Gamma, \quad \bar{v}_\alpha^1 \in W^{2,p_1}(\Omega)$$

where $p_1 = \min(p, 2N/(N-2))$ (if $N \leqq 2$, $p_1 = p$). If $p_1 > N/2$, $N \geqq 3$ or if $N \leqq 2$, $\bar{v}_\alpha^1$ is bounded in $L^\infty(\Omega)$ and we conclude. Now if $p_1 < N/2$, we deduce that $\bar{v}_\alpha^1$ is bounded in $L^{p_2}(\Omega)$ where $p_2 = Np_1/(N-2p_1)$ and thus v_α is also bounded in $L^{p_2}(\Omega)$. By an easy bootstrap argument we then obtain the L^∞ bound.

Next, we observe that v_α is also a solution of the following QVI:

$$\begin{aligned} &a(v_\alpha, v - v_\alpha) + (v_\alpha, v - v_\alpha) \geqq (f_\alpha + (1-\alpha)v_\alpha, v - v_\alpha) \\ &\forall v \in H^1(\Omega), \quad v \leqq Mv_\alpha \text{ a.e.}, \quad v_\alpha \in C(\bar{\Omega}) \cap H^1(\Omega), \quad v_\alpha \leqq Mv_\alpha \text{ in } \bar{\Omega}. \end{aligned} \tag{12}$$

If we denote by $g_\alpha = f_\alpha + (1-\alpha)v_\alpha$, we observe that g_α is bounded in $L^p(\Omega)$ and thus we conclude that v_α is compact in $C(\bar{\Omega})$ by a simple application of the following result proved in the appendix.

PROPOSITION 3. *Under assumptions* (5)-(6) *we denote by* $Tf = w$ *the solution of the following* QVI:

$$\begin{aligned} &a(w, v - w) + (w, v - w) \geqq (f, v - w) \\ &\forall v \in H^1(\Omega), \quad v \leqq Mw \text{ a.e.}, \; w \in C(\bar{\Omega}) \cap H^1(\Omega), \quad w \leqq Mw \text{ in } \bar{\Omega}. \end{aligned} \tag{13}$$

Then T maps bounded sets of $L^p(\Omega)$ into compact sets of $C(\bar{\Omega})$.

Using Proposition 3 we may now prove our claim (Theorem 1) on the convergence of αu_α, v_α to a subsequence (uniqueness will prove that the whole sequence converges); we may assume that as $\alpha \to 0_+$

$$\lambda_\alpha = \alpha\langle u_\alpha \rangle \to \lambda, \qquad v_\alpha \to v_0 \text{ in } C(\bar{\Omega}) \text{ and weakly in } H^1,$$

(thus $\alpha u_\alpha \to \lambda$ uniformly). Of course we have

$$\langle v_0 \rangle = 0, \qquad Mv_\alpha \to Mv_0 \text{ in } C(\bar{\Omega}) \text{ as } \alpha \to 0_+;$$

and we let $\delta_\alpha = \|Mv_\alpha - Mv_0\|_\infty$. If $v \in H^1(\Omega)$, $v \leqq Mv_0$ a.e., $v - \delta_\alpha \leqq Mv_\alpha$ a.e. and

consequently is an admissible test function for problem (7). Thus we find

$$a(v_\alpha, v-v_\alpha)+\alpha(v_\alpha, v-\delta_\alpha-v_\alpha)\geqq(f-\lambda_\alpha, v-\delta_\alpha-v_\alpha)$$

and passing to the limit we find

$$a(v_0, v)\geqq(f-\lambda, v-v_0)+\overline{\lim_\alpha}\, a(v_\alpha, v_\alpha).$$

Since by weak convergence

$$\underline{\lim_\alpha}\, a(v_\alpha, v_\alpha)\geqq a(v_0, v_0)$$

we deduce that (λ, v_0) solves (3). But in addition choosing $v=v_0$ we obtain $a(v_\alpha, v_\alpha)\to a(v_0, v_0)$ and thus ∇v_α converges strongly in L^2 to $a(v_0, v_0)$, proving the strong convergence in H^1.

5. Uniqueness. We first prove the uniqueness of λ: thus let (λ, v_0), (μ, w_0) be two solutions of (3). If $\lambda\neq\mu$, assume for example $\lambda<\mu$; then for $\varepsilon>0$ small enough

$$\lambda-\varepsilon(v_0-C_1)<\mu-\varepsilon w_0 \quad \text{in } \bar{\Omega}$$

where $C_1>\|v_0-w_0\|_{L^\infty}$. Observe next that w_0 (resp. v_0-C_1) is the solution of the following QVI

$$a(w_0, v-w_0)+\varepsilon(w_0, v-w_0)\geqq(f-\mu+\varepsilon w_0, v-w_0)$$

$$\forall v\in H^1(\Omega), \quad v\leqq Mw_0 \text{ a.e.}, \qquad w_0\in C(\bar{\Omega})\cap H^1(\Omega), \quad w_0\leqq Mw_0 \text{ a.e.}$$

(resp. the same problem with $f-\mu+\varepsilon w_0$ replaced by $f-\lambda+\varepsilon(v_0-C_1)$). Standard comparison results then yield: $v_0-C_1\geqq w_0$ in $\bar{\Omega}$, but in view of the choice of C_1 this is not possible. Hence $\lambda=\mu$.

We now have to prove: $v_0\equiv w_0$. This is more delicate and we are first going to show that v_0-w_0 is constant on an open set ω (not empty) included in Ω. To prove this, we assume $v_0\not\equiv w_0$ and then we claim there exists one maximum point of v_0-w_0 denoted by $\bar{x}$ such that

$$w_0(\bar{x})<Mw_0(\bar{x}).$$

Indeed if x_0 is any maximum point of v_0-w_0 on $\bar{\Omega}$, then if the above inequality does not hold at x_0, there exists $\xi\geqq 0$ such that

$$w_0(x_0)=k+c_0(\xi)+w_0(x_1) \quad \text{where } x_1=x_0+\xi\in\bar{\Omega}.$$

Then necessarily we have also

$$v_0(x_0)=k+c_0(\xi)+v_0(x_1)=Mv_0(x_0)$$

since if it were not the case we would have

$$\max(v_0-w_0)<v_0(x_1)-w_0(x_1)$$

and this is obviously impossible. Now clearly x_1 is a new maximum point of v_0-w_0 and we may choose $\bar{x}=x_1$ since $Mw_0(x_1)>w_0(x_1)$.

Then there exists $\delta>0$ such that on $V=\overline{B(\bar{x}, \delta)}\cap\bar{\Omega}$ we have $w_0<Mw_0$ and thus

$$-\Delta w_0=f-\lambda \quad \text{in } \mathring{V}, \qquad \frac{\partial w_0}{\partial n}=0 \quad \text{on } V\cap\partial\Omega.$$

Two cases occur. If $\bar{x}\in\Omega$ then in a connected neighborhood ω of $\bar{x}$ we have

$$a(v_0-w_0, \varphi)\leqq 0 \quad \forall\varphi\in\mathcal{D}_+(\omega), \qquad v_0-w_0\in H^1(\Omega),$$

$$v_0-w_0 \text{ achieves its maximum at } \bar{x}\in\omega$$

and the strong maximum principle (for weak subsolutions) implies that $v_0 - w_0$ is constant on ω.

On the other hand if $\bar{x} \in \partial\Omega$, then for δ small enough $\omega = B(\bar{x}, \delta) \cap \Omega$ is connected and we have:

$$a(v_0 - w_0, \varphi) \leqq 0 \quad \forall \varphi \in H^1_{\Gamma_0}(\Omega), \quad \varphi \geqq 0,$$

$$\max_{\bar{\omega}} (v_0 - w_0) = (v_0 - w_0)(\bar{x}) > 0$$

where

$$H^1_{\Gamma_0}(\omega) = \{u \in H^1(\omega), u|_{\Gamma_0} = 0\}, \qquad \Gamma_0 = \partial B(\bar{x}, \delta) \cap \Omega.$$

Considering the solution $\bar{u}$ of

$$-\Delta \bar{u} = 0 \quad \text{in } \omega, \qquad \frac{\partial \bar{u}}{\partial n} = 0 \quad \text{on } \partial\Omega \cap B(\bar{x}, \delta),$$

$$\bar{u} - (v_0 - w_0) \in H^1_{\Gamma_0}$$

we have: $v_0 - w_0 \leqq \bar{u} \leqq \max_{\bar{\omega}} (v_0 - w_0) = (v_0 - w_0)(\bar{x})$. Thus $\max_{\bar{\omega}} \bar{u} = \bar{u}(\bar{x})$, and again by the strong maximum principle we conclude. Therefore in both cases we have proved that

$$v_0 - w_0 = \max_{\bar{\Omega}} (v_0 - w_0) \quad \text{in } B(y_0, \delta) \quad \overline{B(y_o, \delta)} \subset \Omega \text{ for some } y_0 \in \Omega, \ \delta > 0. \tag{14}$$

Using (14), we are going to prove that $v_0 \equiv w_0 + \max_{\bar{\Omega}} (v_0 - w_0)$ on $\bar{\Omega}$ and since $\langle v_0 \rangle = \langle w_0 \rangle = 0$ the proof of the uniqueness will be completed. To prove this last claim, we consider the problem in $\Omega_1 = \Omega - \overline{B(y_0, \delta)}$

$$\begin{aligned} &a(u, v - u) \geqq (f - \lambda, v - u) \quad \forall v \in H^1(\Omega_1), \\ &v \equiv \varphi_0 \quad \text{on } \partial B(y_0, \delta), \qquad v \leqq Mu \quad \text{a.e. on } \Omega_1, \\ &u \in C(\bar{\Omega}) \cap H^1(\Omega_1), \quad u \equiv \varphi_0 \text{ in } \overline{B(y_0, \delta)}, \quad u \leqq Mu \text{ a.e. on } \Omega_1 \end{aligned} \tag{15}$$

where $\varphi_0 \equiv w_0 + \max_{\bar{\Omega}} (v_0 - w_0)$. Clearly enough in view of (14) both v_0 and φ_0 are solutions of (15) and hence we just have to prove that (15) has a unique solution. And this will be done by an easy adaptation of the classical arguments on QVI due to A. Bensoussan and J. L. Lions [3], B. Hanouzet and J. L. Joly [8]. Indeed let $\underline{u}$ be a fixed solution of (15), if $v \in H^1(\Omega_1) \cap C(\bar{\Omega}_1)$, $v \geqq \underline{u}$ in $\bar{\Omega}_1$ we set $u = Sv$ the solution of the variational inequality

$$\begin{aligned} &a(u, w - u) \geqq (f - \lambda, w - u) \quad \forall w \in H^1(\Omega_1), \qquad w = \underline{u} \text{ on } \partial B(y_0, \delta), \\ &w \leqq M\tilde{v} \text{ in } \Omega_1, \quad u \in H^1(\Omega_1), \quad u = \underline{u} \text{ on } \partial B(y_0, \delta), \quad u \leqq M\tilde{v} \text{ in } \Omega_1 \end{aligned}$$

where $\tilde{v}, \tilde{u}$ are the extensions of v, u to $\bar{\Omega}$ by φ_0 on $\overline{B(y_0, \delta)}$. Then $u \in C(\bar{\Omega}_1)$ and thus $\tilde{u} \in C(\bar{\Omega})$. Finally we consider u_0 solution of

$$-\Delta u_0 = f - \lambda \text{ in } \Omega_1, \quad \frac{\partial u_0}{\partial n} = 0 \text{ on } \partial\Omega, \quad u_0 = \varphi_0 \text{ on } \partial B(y_0, \delta);$$

and we set $\bar{u} = u_0 + C$. Choosing C large enough, we find $\theta \in]0, 1[$ such that

$$\underline{u} \leqq u_0 \leqq \theta\bar{u} + (1 - \theta)\underline{u} \quad \text{in } \bar{\Omega}_1.$$

Observing that S is nondecreasing and concave, we deduce

$$\underline{u} \leqq u_1 = Su_0 \leqq \theta S\bar{u} + (1 - \theta)\underline{u} \leqq \theta u_0 + (1 - \theta)\underline{u};$$

and iterating this string of inequalities, we find for $n \geqq 1$

$$\underline{u} \leqq u_n = S^n u_0 \leqq \theta u_{n-1} + (1-\theta)\underline{u} \quad \text{in } \bar{\Omega}_1.$$

Therefore for $n \geqq 1$

$$0 \leqq u_n - \underline{u} \leqq \theta^n (u_0 - \underline{u}) \quad \text{in } \bar{\Omega}_1$$

and u_n converges uniformly on $\bar{\Omega}_1$ to $\underline{u}$; since $\underline{u}$ is an arbitrary solution of (15), the uniqueness is proved.

6. Extensions. We will only mention the extension to a general operator

$$A = -a_{ij}\partial_{ij} - b_i \partial_i$$

where we use the convention on repeated indices and where $\partial_i = \partial/\partial x_i$. We will assume (for example)

$$a_{ij} = a_{ji} \in W^{1,\infty}(\Omega), \qquad b_i \in L^\infty(\Omega), \qquad \exists \nu > 0, \forall x \in \bar{\Omega}, \quad (a_{ij}) \geqq \nu I_N. \tag{16}$$

We will also consider a general oblique derivative boundary condition determined by a smooth vector field (say C^2) on Γ denoted by γ satisfying

$$\exists \nu > 0, \quad \forall x \in \Gamma, \quad \gamma(x) \cdot n(x) \geqq \nu. \tag{17}$$

In view of (16) we may also write A in divergence form:

$$A = -\partial_i (a_{ij}\partial_j) - (b_j - \partial_i a_{ij})\partial_j;$$

and we will denote by n_A the conormal

$$(n_A)_i = a_{ij} n_j, \quad \forall i \in \{1, \cdots, N\}, \quad \forall x \in \Gamma;$$

and we decompose γ as follows:

$$n_A = c\gamma + \gamma', \qquad n \cdot \gamma' = 0 \quad \text{on } \Gamma \tag{18}$$

and (17) implies $c \geqq c_0 > 0$ on Γ.

The strong formulation (and essentially heuristic) of the analogue of (1) is now:

$$\begin{aligned} &\max\,(Au_\alpha + \alpha u_\alpha - f,\, u_\alpha - Mu_\alpha) = 0 \quad \text{in } \Omega, \\ &\max\left(u_\alpha - Mu_\alpha, \frac{\partial u_\alpha}{\partial \gamma}\right) = 0 \quad \text{on } \Gamma. \end{aligned} \tag{19}$$

The precise formulation of (19) is

$$\begin{aligned} &a(u_\alpha, v - u_\alpha) + \alpha(u_\alpha, v - u_\alpha) \geqq (f, v - u_\alpha) \quad \forall v \in H^1(\Omega), \quad v \leqq Mu_\alpha \text{ a.e.}, \\ &u_\alpha \in C(\bar{\Omega}) \cap H^1(\Omega), \qquad u_\alpha \leqq Mu_\alpha \quad \text{in } \bar{\Omega} \end{aligned} \tag{20}$$

where the bilinear form is now given by

$$a(u, v) = \int_\Omega a_{ij}\partial_i u \partial_j v - b_i \partial_i uv + \partial_i a_{ij} \partial_j uv \, dx + \int_\Gamma \partial_{\gamma'} u \cdot v \, dS$$

the last integral being well defined (by duality) on $H^1(\Omega)^2$ since γ' is tangential.

Adapting easily the arguments of A. Bensoussan and J. L. Lions [3], one finds that (20) has a unique solution (the fact that $u_\alpha \in C(\bar{\Omega})$ being deduced from the methods of the appendix). Mimicking the arguments of the preceding sections, we obtain:

THEOREM 4. *Under assumptions* (5), (6), (16), (17), αu_α *converges uniformly on* $\bar{\Omega}$ *to some constant* λ, *and* $u_\alpha - \langle u_\alpha \rangle$ *converges uniformly on* $\bar{\Omega}$ *and strongly in* $H^1(\Omega)$ *to*

v_0. *Finally* (λ, v_0) *is the unique solution of*

$$
\begin{aligned}
&a(v_0, v-v_0) \geqq (f-\lambda), v-v_0) \quad \forall v \in H^1(\Omega), \quad v \leqq Mv_0 \text{ a.e.},\\
&v_0 \in C(\bar{\Omega}) \cap H^1(\Omega), \qquad v_0 \leqq Mv_0 \text{ in } \bar{\Omega}.
\end{aligned} \tag{21}
$$

7. The ergodic impulse control of diffusion processes. All throughout this section, we assume (5), (6), (16), (17). We introduce $\sigma_{ij} = \sqrt{2}(a_{ij})^{1/2}$ (σ_{ij} is the positive definite symmetric square root) and we first recall the stochastic interpretation of u_α—the solution of (20). To simplify notation, we will assume that $b_i \in W^{1,\infty}(\Omega)$, $f \in C(\bar{\Omega})$.

We consider (a, F, F_t, P, B_t) a probability space with a F_t-adapted Brownian motion in $\mathbb{R}^N$. The control will consist of a sequence of stopping times $(\theta_n)_{n\geqq 0}$ satisfying

$$
\theta_0 = 0 < \theta_1 < \theta_2 < \cdots < \theta_n < \theta_{n+1} \leqq +\infty, \qquad \theta_n \underset{n}{\uparrow} \infty \quad \text{a.s.}
$$

and of a sequence of random variables $(\xi_n)_{n\geqq 1}$ such that

$$
\xi_n \text{ is } F_{\theta_n}\text{-measurable}, \quad \xi_n \geqq 0, \quad X^n_{\theta_n} + \xi_n \in \bar{\Omega} \quad \text{a.s.},
$$

where X^n_t is defined below. We define inductively X^n_t for $t \in [\theta_n, \theta_{n+1}]$ by the solution of the following stochastic differential equation with reflection:

$$
\begin{aligned}
&dX^n_t = \sigma(X^n_t)\, dB_t + b(X^n_t)\, dt - \gamma(X^n_t)\, dA^n_t \quad \text{for } \theta_{n-1} \leqq t \leqq \theta_n,\\
&X^n_{\theta_{n-1}} = X^{n-1}_{\theta_{n-1}} + \xi_{n-1}, \qquad X^n_t \in \bar{\Omega} \quad \forall t \in [\theta_{n-1}, \theta_n],\\
&A^n_t = \int_{\theta_{n-1}}^{t} 1_\Gamma(X^n_s)\, d|A^n|_s \quad \forall t \in [\theta_{n-1}, \theta_n]
\end{aligned}
$$

where X^n_t, A^n_t are continuous, F_t-adapted; A^n_t has bounded variations and $|A^n|_t$ is the total variation of A^n on $[\theta_{n-1}, t]$. Finally $X^0_{\theta_0} = x$, $\xi_0 = 0$.

For more details on stochastic differential equations with reflection we refer the reader to Gihman and Skorohod [6], Ikeda and Watanabe [9], A. Bensoussan and J. L. Lions [3], P. L. Lions and A. S. Sznitman [15].

In what follows we denote by $\mathcal{S}$ any admissible control system consisting of $(\theta_n)_{n\geqq 0}$, $(\xi_n)_{n\geqq 1}$ with the above conditions. If θ is any stopping time ($\theta \leqq \infty$), we define the cost functions:

$$
\begin{aligned}
&J_\alpha(x, \theta, \mathcal{S}) = E\int_0^\theta f(X_t)\, e^{-\alpha t}\, dt + \sum_{n\geqq 1} \{k + c(\xi_n)\} 1_{\theta_n < \theta}\, e^{-\alpha \theta_n},\\
&J_\alpha(x, \mathcal{S}) = J_\alpha(x, +\infty, \mathcal{S}),\\
&J(x, \theta, \mathcal{S}) = E\int_0^\theta f(X_t)\, dt + \sum_{n\geqq 1} \{k + c(\xi_n)\} 1_{\theta_n < \theta} \quad \text{if } \theta \text{ is bounded},
\end{aligned}
$$

where $X_t = X^n_t$ for $t \in [\theta_{n-1}, \theta_n[$, $X_0 = x$.

THEOREM 5. *With the above conditions and notation, we have*

$$
u_\alpha(x) = \inf_{\mathcal{S}} J_\alpha(x, S).
$$

In addition λ*—the limit of* $\alpha\langle u_\alpha\rangle$*—is given by*

$$
\lambda = \lim_{T\to\infty} \frac{1}{T} \inf_{\mathcal{S}} J(x, T, \mathcal{S}), \quad \textit{uniformly for } x \in \bar{\Omega},
$$

$$
\lambda = \lim_{\alpha\to 0+} \alpha \inf_{\mathcal{S}} J_\alpha(x, \mathcal{S}), \quad \textit{uniformly for } x \in \bar{\Omega}.
$$

Furthermore if $v \in C(\bar{\Omega})$, *we denote by*

$$v(x, t) = \inf_{\mathcal{S}} \{J(x, t, \mathcal{S}) + E(v(X_t))\} \quad \forall x \in \bar{\Omega}, \quad \forall t \geqq 0;$$

then $v(x, t) - \langle v(\cdot, t)\rangle$ *converges uniformly on* $\bar{\Omega}$ *to* v_0 *as* $t \to +\infty$. *In particular if* $v \in C(\bar{\Omega})$ *satisfies for some* $T > 0$

$$v(x) = \inf_{\mathcal{S}} \{J(x, T, \mathcal{S}) + E(v(X_t))\} \quad \forall x \in \bar{\Omega}$$

then $v - \langle v\rangle = v_0$; *in addition* v_0 *satisfies* (*where* θ *is any bounded stopping time which may depend on* $\mathcal{S}$):

$$v_0(x) = \inf_{\mathcal{S}} \{J(x, \theta, \mathcal{S}) + E(v_0(X_\theta))\}.$$

Finally (λ, v_0) *is the unique solution in* $\mathbb{R} \times C(\bar{\Omega})$ (*up to the addition of a constant to* v_0) *of*

$$v_0(x) = \lim_{T \to \infty} \inf_{\mathcal{S}} \left\{ E \int_0^T \{f(X_t) - \lambda\}\, dt + \sum_{n \geqq 1} (k + c(\xi_n)) 1_{\theta_n < T} + v_0(X_T) \right\}$$

or

$$v_0(x) = \lim_{\alpha \to 0+} \inf_{\mathcal{S}} \left\{ E \int_0^\infty \{f(X_t) - \lambda\}\, e^{-\alpha t}\, dt + \sum_{n \geqq 1} (k + c(\xi_n))\, e^{-\alpha \theta_n} \right\}.$$

We skip the proof of this result since it is a straightforward application of our results and methods and of the verification arguments given in A. Bensoussan and J. L. Lions [3]. Let us also mention that, as in [3], all infima above are achieved for "explicit" optimal control systems $\mathcal{S}$ obtained by considering the successive exit times of the open set $\{v_0 < Mv_0\}$ and by choosing at each x in $\{v_0 = Mv_0\}$ an optimal jump i.e. choosing (in a measurable way) $\xi \geqq 0$ such that

$$v_0(x) = k + c_0(\xi) + v_0(x + \xi) \quad \text{if } x \in \{v_0 = Mv_0\}.$$

Appendix. Continuity results for variational inequalities and quasivariational inequalities with general oblique derivative boundary conditions. We begin with results concerning variational inequalities. We thus assume (6), (16), (17) and we let $\psi \in C(\bar{\Omega})$. Then exactly as in [3], [4] (see also P. L. Lions [13]), denoting by

$$b(u, v) = \int_\Omega a_{ij} \partial_i u \partial_j v - b_i \partial_i u v + \partial_i a_{ij} \partial_j u v + cuv\, dx + \int_\Gamma \partial_{\gamma'} u \cdot v\, dS$$

where

$$c \in L^\infty(\Omega), \qquad c \geqq c_0 > 0, \tag{A.1}$$

there exists a unique solution u of

$$b(u, v - u) \geqq (f, v - u) \quad \forall v \in H^1(\Omega), \quad v \leqq \psi \text{ a.e.}, \quad u \in H^1(\Omega), \quad u \leqq \psi \text{ a.e.} \tag{A.2}$$

The following result extends and sharpens results of [3]; this result is proved by the method of P. L. Lions [12]—see also P. L. Lions [14], F. Gimbert [7] for similar results.

THEOREM A.1. *We assume* (6), (16), (17) *and* (A.1) *and* $\psi \in C(\bar{\Omega})$. *Then if* u *is the solution of* (A.2), $u \in C(\bar{\Omega})$. *Furthermore if* $\psi \in W^{1,\infty}(\Omega)$ *and* $f \in L^p(\Omega)$ *with* $p > N$, $u \in W^{1,\infty}(\Omega)$; *and if* $\psi \in C^{0,\alpha}(\bar{\Omega})$ *for some* $\alpha \in]0, 1[$, $f \in L^p(\Omega)$ *with* $p = N/(2 - \alpha)$ *if* $N \geqq 2$ ($p = 1$ *if* $N = 1$) *then* $u \in C^{0,\alpha}(\bar{\Omega})$. *Finally if* ψ *lies in a compact set of* $C(\bar{\Omega})$, f

remains bounded in $L^p(\Omega)$ *for some* $p > N/2$ ($p = 1$ *if* $N = 1$), *then* u *remains in a compact set of* $C(\bar{\Omega})$.

Proof. We just have to prove that if $\psi \in W^{1,\infty}(\Omega)$ and $f \in L^p(\Omega)$ for some $p > N$ then $u \in W^{1,\infty}(\Omega)$. Indeed the remainder then follows from the well-known inequality:

$$\|(u_1 - u_2)^+\|_{L^\infty} = \|(\psi_1 - \psi_2)^+\|_{L^\infty}$$

if u_1, u_2 are the solutions of (A.2) corresponding to ψ_1, ψ_2.

Now if $\psi \in W^{1,\infty}(\Omega)$, substracting the solution of the linear problem:

$$A\bar{u} + c\bar{u} = f \text{ in } \Omega, \quad \partial_\gamma \bar{u} = 0 \text{ on } \Gamma, \quad \bar{u} \in W^{2,p}(\Omega)$$

we may assume without loss of generality that $f \equiv 0$.

Next, we recall (cf. [3]) that u is the limit as ε goes to 0 of the solution u_ε of:

$$Au_\varepsilon + cu_\varepsilon + \beta_\varepsilon(u_\varepsilon - \psi) = 0 \text{ in } \Omega, \quad \partial_\gamma u_\varepsilon = 0 \text{ on } \Gamma, \quad u_\varepsilon \in W^{2,p}(\Omega),$$

where $\beta_\varepsilon(t) = (1/\varepsilon)\beta(t)$, $\beta \equiv 0$ if $t \leqq 0$, $\beta'(t) > 0$ if $t > 0$, β is convex, smooth on $\mathbb{R}$.

Since the bounds on $\|u_\varepsilon\|_{W^{1,\infty}}$ obtained below just depend on (16), (17) and $\|\psi\|_{W^{1,\infty}}$, we may assume that a_{ij}, b_i, c, are smooth and so is u_ε. By a simple use of the maximum principle, we find:

$$\|u_\varepsilon\|_{L^\infty(\Omega)} \leqq C$$

where C denotes various constants independent of ε.

We recall from [12], [14], [15] that there exists a smooth matrix valued function $(\alpha_{ij})_{1 \leqq i,j \leqq N}$ satisfying

$$\exists \nu, \mu > 0, \forall x \in \bar{\Omega}, \quad \nu I_N \leqq (\alpha_{ij}) \leqq \mu I_N, \qquad (\alpha_{ij}) = (\alpha_{ji}),$$

$$\partial_\gamma(\alpha_{ij}\partial_i v \partial_j v) \leqq 0 \text{ on } \Gamma, \quad \forall v \in C^2(\bar{\Omega}), \qquad \partial_\gamma v = 0 \text{ on } \Gamma,$$

and we set

$$w = \alpha_{ij}\partial_i u_\varepsilon \partial_j u_\varepsilon + \lambda(C_0 - u_\varepsilon)^2$$

where $\lambda < 0$ is determined below, and $C_0 \geqq 1 + \|u_\varepsilon\|_{L^\infty}$. To simplify notation, we omit below the subscript ε and we set $\varphi_i = \partial_i \varphi$. Then we compute easily Aw and using the equation we find for some $\delta > 0$

$$\begin{aligned} Aw + 2cw \leqq & -2\delta|D^2u|^2 - 2\lambda\delta|Du|^2 - 2\alpha_{ij}u_i\beta'(u-\psi)(u_j - \psi_j) \\ & + C|D^2u||Du| + C|Du|^2 + 2\alpha_{ij}u_i b_{k,j}u_k \\ & - 2\alpha_{ij}u_i c_j u + 2\lambda(C_0 - u)\beta(u-\psi) + C; \end{aligned}$$

choosing λ large enough we deduce

$$\begin{aligned} Aw + (2c + \lambda\delta)w \leqq & 2(\alpha_{ij}u_i b_k u_k)_j + 2(\alpha_{ij}u_i c)_j + C\lambda^2 \\ & + 2\lambda(C_0 - u)\beta(u-\psi) - \beta'(u-\psi)(\alpha_{ij}u_i u_j - \alpha_{ij}\psi_i\psi_j). \end{aligned}$$

Since β is convex, $\beta(t) = 0$ if $t \leqq 0$; $\beta(u-\psi) \leqq \beta'(u-\psi)(u-\psi)$ and we finally obtain

$$Aw + (2c + \lambda\delta)w \leqq 2(\alpha_{ij}u_i b_k u_k)_j + 2(\alpha_{ij}u_i uc)_j + C\lambda^2 - \beta'(u-\psi)(w - \tilde{w})$$

where $\tilde{w} = \alpha_{ij}\psi_i\psi_j + \lambda(C - \psi)^2 \leqq C$.

Using comparison arguments, we deduce

$$w \leqq C\lambda + \|z_\lambda\|_{L^\infty(\Omega)} \quad \text{in } \bar{\Omega}$$

where z_λ is the solution of

$$Az_\lambda + (2c + \lambda\delta)z = 2(\alpha_{ij}u_i b_k u_k)_j + 2(\alpha_{ij}u_i uc)_j \quad \text{in } \Omega,$$

$$\partial_\gamma z_\lambda = 0 \quad \text{on } \Gamma.$$

But classical bounds on linear equations yield

$$\|z_\lambda\|_{L^\infty} \leqq \mu(\lambda)[\|\nabla u\|^2_{L^\infty(\Omega)} + \|\nabla u\|_{L^\infty(\Omega)}]$$

where $\mu(\lambda) \to 0$ as $\lambda \to +\infty$. Hence we obtain finally

$$\|\nabla u\|^2_{L^\infty(\Omega)} \leqq C\lambda + \mu(\lambda)[\|\nabla u\|^2_{L^\infty(\Omega)} + \|\nabla u\|_{L^\infty(\Omega)}]$$

and we conclude choosing λ large.

Let us mention that if b_i, $c \in W^{1,\infty}$, we conclude directly without introducing z_λ. □

We next state the extension of Proposition 3 to the case of general operators A:

THEOREM A.2. *We assume* (5), (6), (16), (17) *and* (A.1). *Then we denote by u the solution of the* QVI

$$\text{(A.3)} \qquad \begin{aligned} &b(u, v-u) \geqq (f, v-u) \quad \forall v \in H^1(\Omega), \quad v \leqq Mu \text{ a.e.} \\ &u \in C(\bar{\Omega}) \cap H^1(\Omega), \qquad u \leqq Mu \quad \text{in } \bar{\Omega}. \end{aligned}$$

Then if f lies in a bounded set of $L^p(\Omega)$ *with* $p > (N/2)$ *if* $N \geqq 2$, $p = 1$ *if* $N = 1$; *u remains in a compact set of* $C(\bar{\Omega})$.

Proof. We recall first that there exists a subsolution $\underline{u}$ of (A.3) which remains in a compact set of $C(\bar{\Omega})$ if f stays bounded in $L^p(\Omega)$ (see for example the construction of $\underline{u}$ in § 2).

Clearly the usual supersolution $\bar{u}$, solution of the linear problem

$$A\bar{u} + c\bar{u} = f \text{ in } \Omega, \quad \partial_\gamma \bar{u} = 0 \text{ on } \Gamma, \quad \bar{u} \in W^{2,p}(\Omega),$$

has the same property. We let $u_0 = \bar{u}$, and we define a sequence u_n inductively as the solution of (A.2) corresponding to $\psi = Mu_{n-1}$. By induction we prove that u—for each fixed n—remains in a compact set of $C(\bar{\Omega})$ if f stays bounded in $L^p(\Omega)$. Indeed this is true for $n = 0$ and if it is true for u_{n-1}, since M maps $C(\bar{\Omega})$ into $C(\bar{\Omega})$ (assumption (4)) and

$$\|Mu - Mv\|_\infty \leqq \|u - v\|_\infty,$$

Mu_{n-1} remains in a compact set of $C(\bar{\Omega})$ and so does u_n by Theorem A.1. Hence by Ascoli's theorem

$$|u_n(x) - u_n(y)| \leqq \delta_n(|x-y|), \qquad \delta_n(t) \to 0 \quad \text{as } t \to 0_+$$

for all x, y $\bar{\Omega}$; where δ_n does not depend on f provided f remains bounded in L^p.

Next, the Hanouzet-Joly argument [8] shows that u_n converges uniformly to u solution of (A.3) and

$$\|u_n - u\|_\infty \leqq C\theta^n$$

for some $\theta \in]0, 1[$, $C \geqq 0$ independent of f provided f remains bounded in $L^p(\Omega)$.

In conclusion, we find if f is bounded in L^p

$$\underline{u} \leqq u \leqq \bar{u} \text{ in } \bar{\Omega} \text{ and thus } \|u\| \leqq C,$$

$$|u(x) - u(y)| \leqq \inf_{n \geqq 1} \{C\theta^n + \delta_n(|x-y|)\} \quad \forall x, y \in \bar{\Omega}.$$

And by Ascoli's theorem, our claim is proved. □

Acknowledgments. The authors would like to thank A. Bensoussan, M. Robin and J. L. Menaldi for calling their attention to the problem treated in this paper.

REFERENCES

[1] A. BENSOUSSAN, *Stochastic Control by Functional Analysis Methods*, North-Holland, Amsterdam, 1982.

[2] A. BENSOUSSAN AND J. L. LIONS, *On the asymptotic behaviour of the solution of variational inequalities*, in Theory of Nonlinear Operators, Akademie Verlag, Berlin, 1978.

[3] ———, *Contrôle impulsionnel et inéquations quasi-variationnelles*, Dunod, Paris, 1982.

[4] ———, *Applications des inéquations variationnelles en contrôle stochastique*, Dunod, Paris, 1978.

[5] I. CAPUZZO-DOLCETTA AND M. G. GARONNI, *Comportement asymptotique de la solution des problèmes non sous forme divergence avec condition de dérivées obliques au bord*, Comptes Rendus Acad. Sci. Paris, 1984.

[6] I. I. GIHMAN AND A. V. SKOROHOD, *The Theory of Stochastic Processes*, Vols. I, II, III, Springer-Verlag, Berlin, 1975.

[7] F. GIMBERT, *Problèmes de Neumann quasi-linéaires ergodiques*, in Thèse de 3e Cycle, Université Paris IX-Dauphine, 1984; J. Funct. Anal., to appear.

[8] B. HANOUZET AND J. L. JOLY, *Convergence uniforme des itérés définissant la solution d'une inéquation quasi-variationnelle*, Comptes Rendus Acad. Sci. Paris, 286 (1978), pp. 735-738.

[9] N. IKEDA AND S. WATANABE, *Stochastic Differential Equations and Diffusion Processes*, North-Holland-Kodansha, Amsterdam, 1981.

[10] J. M. LASRY, *Contrôle stochastique ergodique*, in Thèse d'Etat, Université Paris IX-Dauphine, 1974.

[11] J. M. LASRY AND P. L. LIONS, *Infinite boundary conditions for nonlinear elliptic equations and optimal stochastic control with state constraints*, in preparation.

[12] P. L. LIONS, *Quelques remarques sur les problèmes elliptiques quasi-linèairès du second ordre*, J. Anal. Math., to appear.

[13] ———, *A remark on some elliptic second-order problems*, Bull. U.M.I., 17 (1980), pp. 267-270.

[14] ———, *Optimal control of diffusion processes and Hamilton-Jacobi-Bellman equations*, Part 1, Comm. Partial Differential Equations, 8 (1983), pp. 1101-1174.

[15] P. L. LIONS AND A. S. SZNITMAN, *Stochastic differential equations with reflecting boundary conditions*, Comm. Pure Appl. Math., 37 (1984), pp. 511-537.

[16] B. PERTHAME, *Quasi-variational inequalities and Hamilton-Jacobi-Bellman equations in a bounded region*, Comm. Partial Differential Equations, to appear.

[17] ———, *Some remarks on quasi-variational inequalities and the associated impulse control problem*, in preparation.

[18] M. ROBIN, *On some impulse control problems with long run average cost*, this Journal, 19 (1981), pp. 333-358.

[19] ———, *Long term average cost control problems for continuous time Markov processes: a survey*, Acta Appl. Math., to appear.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 4, July 1986

003

THE PENCIL $(sE-A)$ AND CONTROLLABILITY–OBSERVABILITY FOR GENERALIZED LINEAR SYSTEMS: A GEOMETRIC APPROACH*

VINÍCIUS A. ARMENTANO†

Abstract. In the first part of the paper we study the pencil $(sE-A)$ by using some geometric concepts given by Bernhard [SIAM J. Control Optim., 20 (1982), pp. 612-633] and some new ones which are introduced here. Such concepts are then used to obtain an alternative characterization for the finite and infinite-zero structure of the pencil and to construct a polynomial basis for $\ker(sE-A)$. Necessary and sufficient conditions for the rows or/and the columns of the pencil to be linearly independent over the ring of the polynomials are also given. The main geometric properties of a regular pencil are presented, including the identification of the subspace in which the impulsive response of the autonomous generalized linear system $E\dot{x}=Ax$ takes place. In the second part we consider the generalized linear system $E\dot{x}=Ax+Bu$; $y=Cx$ and we give necessary and sufficient conditions for the infinite-zeros of the regular pencil $(sE-A)$ to be controllable and observable.

Key words. theory of pencils, linear systems, controllability and observability

AMS(MOS) subject classifications. 15A21, 58F19, 93B

1. Introduction. In the first part of the paper (§§ 2-4) we consider the pencil $(sE-A)$ where s is the complex variable and E and A are given maps.

The pencil $(sE-A)$ is said to be regular if E and A are square maps with $\det(sE-A)\neq 0$ and it is said to be singular otherwise.

Bernhard [3] and Wong [17] have introduced some geometric concepts (subspaces) which are important for a geometric approach to the theory of pencils. In § 2 we introduce some other subspaces and we obtain geometric relations which are very useful for the analysis undertaken in the subsequent sections.

In § 3 we study the singular pencil $(sE-A)$. Bernhard [3] has characterized the finite-zeros, also known as the finite elementary divisors [7], and the minimal column indices of $(sE-A)$ by an indirect method, namely, he has identified them with the control invariants of a certain pair of maps denoted here by (F, G).

It is shown that the above elements can be characterized directly from the geometric concepts introduced by him without resorting to notions of linear systems. A polynomial basis for $\ker(sE-A)$ is also constructed.

We then obtain a block triangular representation for $\ker(sE-A)$ from which we can extract the infinite-zero structure and a polynomial basis for the left kernel of $(sE-A)$. We also refer to [12], [14] for algebraic approaches to infinite-zeros of a pencil.

Necessary and sufficient conditions for the rows or the columns of the singular pencil to be linearly independent over the ring of the polynomials are also given.

The regular pencil is analysed in § 4. This kind of pencil is important in the theory of dynamical linear systems. As an example, in [1] is shown a connection between regular pencils, almost controlled invariant subspaces [16] and the proportional-derivative state feedback law, $u=F_1x+F_2\dot{x}$, for the linear system $\dot{x}=Ax+Bu$.

Associated with a regular pencil we have an autonomous generalized linear system $E\dot{x}=Ax$ which has a unique solution for an arbitrary initial condition $x(0-)$. It has been shown in [13] that the response of such a system to an arbitrary $x(0-)$ consists

* Received by the editors January 3, 1984, and in revised form March 11, 1985.

† Department of Electrical Engineering, FEC-Unicamp, C.P. 6122, Campinas 13100 S.P., Brasil. Part of this work was carried out while the author was with the Department of Electrical Engineering, Imperial College, University of London, London, England.

of a combination of an exponential motion determined by the finite-zeros and an impulsive component due to the infinite-zeros of $(sE - A)$.

The subspace in which the exponential motion occurs has already been characterized in [3], [15], [5]. We identify here the subspace in which the impulsive response takes place and we also summarize the main geometric properties of a regular pencil.

In the second part of the paper (§ 5) we consider the generalized linear system $E\dot{x} = Ax + Bu$; $y = Cx$, such that the associated pencil $(sE - A)$ is regular.

We shall concentrate on the study of controllability and observability of the zeros of the pencil $(sE - A)$. Rosenbrock [10] has been the first author to study this subject and he has given conditions for the infinite-zeros to be controllable and observable. Verghese [13] has pointed out that Rosenbrock's conditions were correct only in a special case and he has then provided tests which are valid in any situation. We have built on such tests to obtain necessary and sufficient conditions for the controllability and observability of the infinite-zeros which are expressed in terms of the geometry of the maps E, A, B and C. Geometric interpretations for the controllable and unobservable finite-zeros are also given.

Notation. We shall use throughout lower case letters for vectors, capitals for matrices and maps, and script for subspaces and vector spaces.

Im and ker denote image (range) and kernel (null space), respectively.

If A is a map and $\mathscr{L}$ and $\mathscr{K}$ are subspaces such that $A\mathscr{L} \subset \mathscr{K}$ we then denote the restriction of A to $\mathscr{L}$ with codomain $\mathscr{K}$ by $\mathscr{K}|A|\mathscr{L}$. Mat A denotes the matrix of the map A.

If $\mathscr{L}$ and $\mathscr{K}$ are subspaces such that $\mathscr{L} \subset \mathscr{K}$ then $\mathscr{K}$ (mod $\mathscr{L}$) or $\mathscr{K}/\mathscr{L}$ denotes the quotient space $\{k + \mathscr{L}, k \in \mathscr{K}\}$, $\dim \mathscr{K}/\mathscr{L} = \dim \mathscr{K} - \dim \mathscr{L}$.

If $\mathscr{X}$ is a vector space then $\mathscr{X}'$ stands for the associated dual vector space.

R denotes the real line and C the complex plane.

If n is a positive integer, then $\underline{n}$ stands for the set of integers $\{1, 2, \cdots, n\}$.

$\sigma(A)$ denotes the spectrum of A, i.e., the eigenvalues of A counting multiplicities.

$R[s]$ denotes the ring of the polynomials.

2. Some fundamental subspaces for a geometric description of the pencil $(sE - A)$. This section defines some subspaces which play a vital role later and establishes useful connections between them.

Let E and A be maps from $\mathscr{X}$ to $\mathscr{Z}$, $\dim \mathscr{X} = n$, $\dim \mathscr{Z} = m$, and let $\mathscr{M}$ be any given subspace of $\mathscr{X}$. Consider the following family F_1 of subspaces

$$F_1 := \{\mathscr{V} \subset \mathscr{X} | A\mathscr{V} \subset E\mathscr{V}, \mathscr{V} \subset \mathscr{M}\}.$$

It can be easily seen that the above family is closed under addition and therefore it contains a supremal element $\mathscr{V}^* \subset \mathscr{M}$ which can be computed through the following nonincreasing sequence [3], [17]

$$\mathscr{V}^* := \mathscr{V}^n, \qquad \mathscr{V}^u = \mathscr{M} \cap A^{-1}(E\mathscr{V}^{u-1}), \quad u \in \underline{n}, \qquad \mathscr{V}^0 = \mathscr{X}. \tag{1}$$

Hereafter we shall assume that $\mathscr{M} = A^{-1}(\text{Im } E)$. Thus $\mathscr{V}^*$ is the supremal element of the family F_1 which is contained in $\mathscr{X}$.

It has been shown in [3], [17] that if λ is a finite-zero of the pencil $(sE - A)$ then there exists a vector $v \in \mathscr{V}^*$ such that $Av = \lambda Ev$.

Let $\mathscr{K}$ be any given subspace of $\mathscr{X}$ and define a family F_2 of subspaces according to

$$F_2 := \{\mathscr{W}_a \subset \mathscr{X} | \mathscr{W}_a = \mathscr{K} \cap E^{-1}(A\mathscr{W}_a)\}.$$

The next proposition states that the above family has a unique least element.

PROPOSITION 1. *There is a unique element* $\mathcal{W}_a^* \in F_2$ *such that* $\mathcal{W}_a^* \subset \mathcal{W}_a$ *for every* $\mathcal{W}_a \in F_2$. *Furthermore* $\mathcal{W}_a^*$ *can be computed by the following sequence*

$$\mathcal{W}_a^* := \mathcal{W}_a^n, \qquad \mathcal{W}_a^u = \mathcal{K} \cap E^{-1}(A\mathcal{W}_a^{u-1}), \quad u \in \underline{n}, \qquad \mathcal{W}_a^0 = 0. \tag{2}$$

Proof. First note that the sequence (2) is nondecreasing. We have $\mathcal{W}_a^1 \supset \mathcal{W}_a^0$ and if $\mathcal{W}_a^u \supset \mathcal{W}_a^{u-1}$, then $\mathcal{W}_a^{u+1} = \mathcal{K} \cap E^{-1}(A\mathcal{W}_a^u) \supset \mathcal{K} \cap E^{-1}(A\mathcal{W}_a^{u-1}) = \mathcal{W}_a^u$.

Thus there exists $k \in \underline{n}$ such that $\mathcal{W}_a^u = \mathcal{W}_a^k$, $u \geqq k$. So $\mathcal{W}_a^* = \mathcal{W}_a^k$ and $\mathcal{W}_a^* \in F_2$. To show that $\mathcal{W}_a^*$ is infimal let $\mathcal{W}_a \in F_2$. Then $\mathcal{W}_a \supset \mathcal{W}_a^0$ and if $\mathcal{W}_a \supset \mathcal{W}_a^u$ we obtain

$$\mathcal{W}_a = \mathcal{K} \cap E^{-1}(A\mathcal{W}_a) \supset \mathcal{K} \cap E^{-1}(A\mathcal{W}_a^u) = \mathcal{W}_a^{u+1}.$$

Thus $\mathcal{W}_a \supset \mathcal{W}_a^u$, $\forall u \in \underline{n}$, so that $\mathcal{W}_a \supset \mathcal{W}_a^*$. □

We shall consider three cases for the subspace $\mathcal{K}$:

(i) $\mathcal{K} = \mathcal{V}^*$. In this situation we denote the infimal element of F_2 by $\mathcal{T}^*$, which is then given by (see (2))

$$\mathcal{T}^* := \mathcal{T}^n, \qquad \mathcal{T}^u = \mathcal{V}^* \cap E^{-1}(A\mathcal{T}^{u-1}), \quad u \in \underline{n}, \qquad \mathcal{T}^0 = 0. \tag{3}$$

We shall see in § 3 that the subspace $\mathcal{T}^*$ is the subspace which gives rise to a polynomial basis for $\ker(sE - A)$.

(ii) $\mathcal{K} = \mathcal{X}$. In this case the infimal element of F_2 is denoted by $\mathcal{W}_b^*$ which in turn is given by

$$\mathcal{W}_b^* := \mathcal{W}_b^n, \qquad \mathcal{W}_b^u = E^{-1}(A\mathcal{W}_b^{u-1}), \quad u \in \underline{n}, \qquad \mathcal{W}_b^0 = 0. \tag{4}$$

It will be shown in §§ 3 and 4 that the subspace $\mathcal{W}_b^*$ is related to the infinite-zero structure of the pencil $(sE - A)$.

(iii) If $\mathcal{K} = A^{-1}(\operatorname{Im} E)$ we let $\mathcal{W}_a^*$ denote the infimum of F_2 (computed by the sequence (2)). It will be shown in § 4 that if $(sE - A)$ is a regular pencil then the subspace $\mathcal{W}_a^*$ defined this way is the subspace in which the impulsive response of the autonomous generalized linear system $E\dot{x} = Ax$ takes place.

The next lemma shows some geometric properties of the sequences (1-4) above defined.

LEMMA 1. *Let* $\mathcal{K} = A^{-1}(\operatorname{Im} E)$ *in the sequence* (2). *Then*

(a) $E\mathcal{T}^u = A\mathcal{T}^{u-1}$,

(b) $\mathcal{T}^u = \mathcal{V}^* \cap \mathcal{W}_b^u$,

(c) $\mathcal{W}_a^u = \mathcal{K} \cap \mathcal{W}_b^u$,

(d) $\mathcal{T}^u = \mathcal{V}^* \cap \mathcal{W}_a^u$,

(e) $E\mathcal{W}_b^u = A\mathcal{W}_a^{u-1}$.

Proof. See Appendix.

The relevance and motivation for the subspaces above defined will be better understood in the following sections.

3. The singular pencil. Let $(sE - A)$ be a singular pencil and let $x(s)$ be a polynomial solution of least degree for the equation

$$(sE - A)x(s) = 0 \tag{5}$$

with

$$x(s) = x_k + sx_{k-1} + s^2x_{k-2} + \cdots + s^kx_0.$$

Substituting the above solution in (5) and equating the coefficients of the same power in s we then obtain

$$Ax_k = 0, \quad Ex_k = Ax_{k-1}, \quad \cdots, \quad Ex_1 = Ax_0, \quad Ex_0 = 0. \tag{6}$$

It has been shown by Gantmacher [7] that the vectors x_i, $i \in \{0, 1, \cdots, k\}$ are linearly independent. Let $\mathscr{D} := \text{span}\{x_i\}$. Then from (6) it follows that $A\mathscr{D} = E\mathscr{D}$. This suggests that $\mathscr{T}^*$ (note that $E\mathscr{T}^* = A\mathscr{T}^*$ by Lemma 1a) is the subspace which provides vectors for a polynomial basis for $\ker(sE - A)$. In other words if $x_i(s)$, $i \in \underline{l}$, is a basis over $R[s]$ for $\ker(sE - A)$ with

$$x_i(s) = x_{i,k_i} + sx_{i,k_i-1} + s^2 x_{i,k_i-2} + \cdots + s^{k_i} x_{i,0},$$

for some set of nonnegative integers $\{k_i\}$, $i \in \underline{l}$, then $\mathscr{T}^* = \text{span}\{x_{i,j}\}$, $i \in \underline{l}$, $j \in \{0, 1, \cdots, k_i\}$.

In order to show this consider the following family of subspaces:

$$F_3 := \{\mathscr{T} | A\mathscr{T} = E\mathscr{T}\}$$

such that $\mathscr{T} = \text{span}\{t_{i,j}\}$ with

$$Et_{i,0} = 0, \quad Et_{i,j} = At_{i,j-1}, \quad At_{i,h_i} = 0, \qquad i \in \underline{r}, \quad j \in \underline{h}_i, \tag{7}$$

where $r := \dim(\ker E \cap \mathscr{T}) = \dim(\ker A \cap \mathscr{T})$ and h_i, $i \in \underline{r}$, is a set of nonnegative integers.

The following theorem states that $\mathscr{T}^*$ is the supremum of F_3.

THEOREM 1. *$\mathscr{T}^* = \sup F_3$ and*

$$\mathscr{T}^* = \bigoplus_{i=1}^{l} \mathscr{T}_i$$

with

$$\mathscr{T}_i = \text{span}\{x_{i,j}\}, \qquad \dim \mathscr{T}_i = k_i + 1$$

such that

$$Ex_{i,0} = 0, \quad Ex_{i,j} = Ax_{i,j-1}, \quad Ax_{i,k_i} = 0, \qquad j \in \underline{k}_i \tag{8}$$

where

$$l := \dim(\ker E \cap \mathscr{T}^*) = \dim(\ker A \cap \mathscr{T}^*)$$

and k_i, $i \in \underline{l}$, is a set of nonnegative numbers uniquely defined from the dimensions of the subspaces $\mathscr{T}^u$ in (3).

Furthermore, for any decomposition of $\mathscr{T}^*$ as the direct sum of independent subspaces such as in (8), it is necessarily true that $\dim \mathscr{T}_i = k_i + 1$, $i \in \underline{l}$.

Proof. See Appendix.

Consider the set $\{x_{i,j}\}$, $i \in \underline{l}$, $j \in \{0, 1, \cdots, k_i\}$ in (8). As an immediate consequence of Theorem 1 we now obtain the following result.

COROLLARY 1. *Let $(sE - A)$ be a singular pencil. Then*

$$\ker(sE - A) = \underset{R[s]}{\text{span}}\{x_i(s)\}, \qquad i \in \underline{l},$$

where

$$l := \dim(\ker E \cap \mathscr{T}^*) = \dim \ker(\ker A \cap \mathscr{T}^*)$$

and

$$x_i(s) = x_{i,k_i} + sx_{i,k_i-1} + \cdots + s^{k_i} x_{i,0}.$$

The set $\{k_i\}$, $i \in \underline{l}$ coincides with the set of minimal column indices defined in [7].

Proof. Since $\mathcal{T}^*$ is the supremum of the family F_3 it follows that $x_i(s)$, $i \in \underline{l}$, is indeed a basis over $R[s]$ for $\ker(sE-A)$. By using (8) we then have

$$\operatorname{Mat}[E\mathcal{T}^* | (sE-A) | \mathcal{T}^*] = P(s) = \operatorname{diag}[P_i(s)], \qquad i \in \underline{l}, \quad \text{where}$$

$$P_i(s) = \begin{bmatrix} s & -1 & \cdot & \cdot & 0 & 0 \\ 0 & s & \cdot & \cdot & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot & -1 & 0 \\ 0 & 0 & \cdot & \cdot & s & -1 \end{bmatrix}_{k_i \times (k_i+1)} \tag{9}$$

Thus $P(s)$ corresponds to the set of canonical blocks associated with the minimal column indices in Gantmacher's decomposition of a singular pencil. The set $\{k_i\}$, $i \in \underline{l}$, is the set of minimal column indices due to the uniqueness of the canonical form of a singular pencil under strict equivalence [7]. □

The next theorem contains a geometric characterization for the finite-zeros of the pencil $(sE-A)$.

THEOREM 2. *Let $(sE-A)$ be a singular pencil and consider the map A_f: $\mathcal{V}^*$ (mod $\mathcal{T}^*$) $\to E\mathcal{V}^*$ (mod $E\mathcal{T}^*$) such that $A_f P = QA_v$ where $A_v := E\mathcal{V}^* | A | \mathcal{V}^*$ and P: $\mathcal{V}^* \to \mathcal{V}^*$ (mod $\mathcal{T}^*$), Q: $E\mathcal{V}^* \to E\mathcal{V}^*$ (mod $E\mathcal{T}^*$) are the canonical projections. Then*

$$\sigma(A_f) = \{\textit{finite-zeros of the singular pencil } (sE-A)\}.$$

Proof. Let $\mathcal{V}_1$ be any subspace such that

$$\mathcal{V}^* = \mathcal{T}^* \oplus \mathcal{V}_1 \tag{10}$$

and note that $\mathcal{V}_1 \cap \ker E = 0$ since $\mathcal{V}^* \cap \ker E \subset \mathcal{T}^*$.

Next we show that $E\mathcal{T}^* \cap E\mathcal{V}_1 = 0$. Suppose that $Et = Ev$, for $t \in \mathcal{T}^*$, $0 \neq v \in \mathcal{V}_1$. Then $v - t \in \ker E \cap \mathcal{V}^*$ and thus $v \in \mathcal{T}^*$, which is impossible. Thus we may consider the direct sum

$$E\mathcal{T}^* \oplus E\mathcal{V}_1 \tag{11}$$

so that in the decompositions (10)-(11)

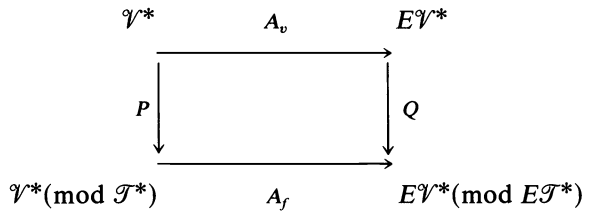
$$\operatorname{Mat}[E\mathcal{V}^* | (sE-A) | \mathcal{V}^*] = \begin{bmatrix} P(s) & -A_{12} \\ 0 & sI - A_{22} \end{bmatrix},$$

where $P(s)$ is as in Corollary 1 and

$$\begin{bmatrix} A_{12} \\ A_{22} \end{bmatrix} = \operatorname{Mat}[E\mathcal{V}^* | A | \mathcal{V}_1].$$

Since an eigenvector v associated with a finite-zero λ belongs to $\mathcal{V}^*$, i.e., $Av = \lambda Ev \Rightarrow v \in \mathcal{V}^*$, it follows that the finite-zeros of the pencil $(sE-A)$ coincide with the eigenvalues of A_{22}.

As a matter of fact, A_{22} is the representation of the map A_f defined above. The following commutative diagram illustrates the definition of A_f.



To show that the diagram commutes, note that $t \in \mathcal{T}^*$ implies $A_v t = Et_1$ for some $t_1 \in \mathcal{T}^*$. Thus $A_f P = QA_v$ and the map A_f is well defined.

Comment. As mentioned previously, in [3] the minimal column indices have been identified with the controllability indices of a certain pair of maps (F, G) while the finite-zeros have been associated with the eigenvalues of the uncontrollable map derived from the pair (F, G) [18]. The aim of Theorem 1, Corollary 1 and Theorem 2 has been to show that we can obtain this information (minimal column indices and finite-zeros) and a polynomial basis for $\ker (sE - A)$ by using only the geometric concepts introduced in the previous section.

We proceed with the geometric description of a singular pencil by identifying its infinite-zero structure and a polynomial basis for $\ker (sE^T - A^T)$ or, equivalently, a polynomial basis for the left kernel of $(sE - A)$.

From Gantmacher's decomposition of a singular pencil [7] it follows that to an infinite-zero of order g_i there corresponds a subspace $\mathcal{W}$, $\dim \mathcal{W} = g_i + 1$ such that

$$E\mathcal{W} \subset A\mathcal{W}, \qquad \mathcal{W} \cap \ker A = 0 \tag{12}$$

and such that the map $N := A\mathcal{W}|E|\mathcal{W}$ has all eigenvalues equal to zero, i.e., there exists a chain of linearly independent vectors $\{w_{i,j}\}$ which span $\mathcal{W}$ (see also [8, Thm. 6]) such that

$$\begin{aligned} &Ew_{i,1} = 0, \\ &Ew_{i,j} = Aw_{i,j-1}, \qquad j \in \{2, \cdots, g_i + 1\}. \end{aligned} \tag{13}$$

By using (13) we obtain the following representation:

$$\text{Mat}\,[A\mathcal{W}|(sE - A)|\mathcal{W}] = sN - I, \quad \text{where}$$

$$N = \begin{bmatrix} 0 & 1 & 0 & \cdot & \cdot & 0 \\ 0 & 0 & 1 & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & 1 \\ 0 & \cdot & \cdot & \cdot & \cdot & 0 \end{bmatrix}_{(g_i+1)\times(g_i+1)} \tag{14}$$

Verghese [14] has shown that the pencil $(sN - I)$ has an infinite-zero of order g_i. The concept of an infinite-zero is a generalization of the concept of a finite-zero in the sense that the pencil $(sN - I)$ loses rank at $s = \infty$.

The next lemma contains some results which will allow us to proceed with the description of structural features of a singular pencil.

LEMMA 2. *Let $(sE - A)$ be a singular pencil for which $\mathcal{V}^* = 0$ in the sequence* (1). *Let $\mathcal{W}_b^*$ be given by the sequence* (4). *Then*:

(a) $\ker_{R[s]} (sE - A) = 0$ *and the pencil has no finite-zeros.*

(b) $E\mathcal{W}_b^* \subset A\mathcal{W}_b^*$ *and the map $N := A\mathcal{W}_b^*|E|\mathcal{W}_b^*$ has all eigenvalues equal to zero. The infinite-zero structure of $(sE - A)$ is determined from the Jordan decomposition*

of N. Moreover

$$\mathcal{W}_b^* = \sup\{\mathcal{W} | E\mathcal{W} \subset A\mathcal{W}\}. \tag{15}$$

(c) *Consider the maps*

$$\hat{E}: \mathcal{X}(\text{mod } \mathcal{W}_b^*) \to \mathcal{X}(\text{mod } A\mathcal{W}_b^*)$$

and

$$\hat{A}: \mathcal{X}(\text{mod } \mathcal{W}_b^*) \to \mathcal{X}(\text{mod } A\mathcal{W}_b^*)$$

such that

$$\hat{E}P = QE, \qquad \hat{A}P = QA,$$

where $P: \mathcal{X} \to \mathcal{X}(\text{mod } \mathcal{W}_b^*)$ *and* $Q: \mathcal{X} \to \mathcal{X}(\text{mod } A\mathcal{W}_b^*)$ *are the canonical projections.*

Then the pencil $(s\hat{E} - \hat{A})$ *has no finite and infinite-zeros and* $\ker_{R[s]}(s\hat{E} - \hat{A}) = 0$. *Furthermore, the rows of* $(s\hat{E} - \hat{A})$ *are linearly dependent over* $R[s]$.

Proof. It is intuitive that if $\mathcal{V}^* = 0$ for the pencil $(sE - A)$ then the structural invariants of such a pencil consist in general of infinite-zeros and minimal row indices only. Therefore $(sE - A)$ is in general a pencil with a number of rows greater or equal to the number of columns. A formal proof of the above theorem is given in the Appendix.

We are, finally, in a position to obtain the infinite-zero structure of the singular pencil $(sE - A)$ as well as a polynomial basis for the left kernel of $(sE - A)$. For this, let $\bar{A}$ and $\bar{E}$ be maps defined by

$$\bar{A}P = QA, \qquad \bar{E}P = QE, \tag{16}$$

where

$$P: \mathcal{X} \to \mathcal{X}(\text{mod } \mathcal{V}^*), \qquad Q: \mathcal{X} \to \mathcal{X}(\text{mod } E\mathcal{V}^*)$$

are the canonical projections. We then have the following theorem.

THEOREM 3. *Let* $(sE - A)$ *be a singular pencil. Then its infinite-zero structure coincides with that of* $(s\bar{E} - \bar{A})$. *Moreover, a polynomial basis for the left kernel of* $(s\bar{E} - \bar{A})$ *is also a polynomial basis for the left kernel of* $(sE - A)$.

Proof. Let $\bar{\mathcal{V}} := \sup\{\mathcal{V} | \bar{A}\mathcal{V} | \bar{E}\mathcal{V}\}$ and write $\mathcal{L} := P^{-1}\bar{\mathcal{V}}$. Then $\bar{A}P\mathcal{L} \subset \bar{E}P\mathcal{L}$, which implies $QA\mathcal{L} \subset QE\mathcal{L}$. Hence

$$A\mathcal{L} \subset E\mathcal{L} + E\mathcal{V}^*$$

which implies

$$A(\mathcal{L} + \mathcal{V}^*) \subset E(\mathcal{L} + \mathcal{V}^*)$$

so that

$$\mathcal{L} + \mathcal{V}^* \subset \mathcal{V}^* \Rightarrow \mathcal{L} \subset \mathcal{V}^* \Rightarrow \bar{\mathcal{V}} = 0.$$

Since $\bar{\mathcal{V}} = 0$ we then conclude from Lemma 2 that the pencil $(s\bar{E} - \bar{A})$ has no finite-zeros and its columns are linearly independent over $R[s]$. By using the results of Corollary 1, Theorem 2 and Lemma 2, it follows that the pencil $(sE - A)$ admits the following triangular decomposition

$$\text{Mat}(sE - A) = \begin{bmatrix} sE_1 - A_1 & x & x & x \\ 0 & sI - A_f & x & x \\ 0 & 0 & sN - I & x \\ 0 & 0 & 0 & sE_4 - A_4 \end{bmatrix} \tag{17}$$

where

$$\begin{bmatrix} sE_1 - A_1 & x \\ 0 & sI - A_f \end{bmatrix} = \text{Mat}\,[E\mathcal{V}^* | (sE - A) | \mathcal{V}^*]$$

and

$$\begin{bmatrix} sN - I & x \\ 0 & sE_4 - A_4 \end{bmatrix} = \text{Mat}\,(s\bar{E} - \bar{A}).$$

The pencil $(sE_1 - A_1)$ is given by $\text{diag}\,[P_i(s)]$, $i \in \underline{l}$, where $P_i(s)$ is as in (9). It can be easily seen that there exists a minor M in $(sE_1 - A_1)$ such that $\det M = s^{k_1 + \cdots + k_l}$, i.e. the rows of $(sE_1 - A_1)$ are linearly independent over $R[s]$.

The pencil $(sI - A_f)$ is regular and the eigenvalues of A_f coincide with the finite-zeros of $(sE - A)$.

The pencil $sN - I$ has only infinite-zeros. The Jordan decomposition of the map N (see Lemma 2b) determines the infinite-zero structure of $(sE - A)$.

By Lemma 2c, the pencil $(sE_4 - A_4)$ has no finite and infinite-zeros, $\ker_{R[s]} (sE_4 - A_4) = 0$ and its rows are linearly dependent over $R[s]$.

It now follows that a polynomial basis for the left kernel of $(sE_4 - A_4)$ is also a basis for the left kernel of $(sE - A)$. □

Comment. The triangular representation (17) is valid for any singular pencil. Van Dooren [11] has obtained a similar representation by means of algorithms which make use of the technique of singular value decomposition. His work is concerned with a stable numerical method which extracts the invariants of a singular pencil whereas this paper deals with the geometric structure of such a pencil.

In the remainder of this section we give geometric criteria for the rows or the columns of a singular pencil to be linearly independent over $R[s]$. First we need a preliminary result.

Consider the maps $\bar{A}$ and $\bar{E}$ in (16) and define the following sequence of subspaces

$$\bar{\mathcal{W}}_b := \bar{\mathcal{W}}_b^n, \qquad \bar{\mathcal{W}}_b^u = \bar{E}^{-1}(\bar{A}\bar{\mathcal{W}}_b^{u-1}), \quad u \in \underline{n}, \quad \bar{\mathcal{W}}_b = 0. \tag{18}$$

The next lemma establishes a relation between the sequence above defined and the sequence (4).

LEMMA 3. *$\bar{\mathcal{W}}_b^u = P\mathcal{W}_b^u$, where $P: \mathcal{X} \to \mathcal{X} (\text{mod } \mathcal{V}^*)$ is the canonical projection.*

Proof. See Appendix.

THEOREM 4. *Let $(sE - A)$ be a singular pencil. Then*:

(a) *Its columns are linearly independent over $R[s]$ if and only if $\mathcal{V}^* \cap \mathcal{W}_b^* = \mathcal{T}^* = 0$ (equivalently, $\mathcal{V}^* \cap \ker E = 0$).*

(b) *Its rows are linearly independent over $R[s]$ if and only if $\mathcal{V}^* + \mathcal{W}_b^* = \mathcal{X}$.*

Proof. (a) Immediate from Corollary 1.

(b) From (17) we have that the rows of $(sE - A)$ are linearly independent over $R[s]$ if and only if $\bar{\mathcal{W}}_b = \mathcal{X} (\text{mod } \mathcal{V}^*)$, i.e., if and only if $\mathcal{V}^* + \mathcal{W}_b^* = \mathcal{X}$ by using Lemma 3. □

4. The regular pencil. In this section we derive several geometric properties for regular pencils. An obvious necessary condition for a pencil $(sE - A)$ to be regular is that E and A are square maps. We then assume that E and A are maps from $\mathcal{X}$ to $\mathcal{X}$, $\dim \mathcal{X} = n$.

We first show geometric criteria for $(sE - A)$ to be regular.

THEOREM 5. *The following statements are equivalent*:

(i) *$(sE - A)$ is a regular pencil.*

(ii) $\mathcal{V}^* \oplus \mathcal{W}_b^* = \mathcal{X}$.

(iii) $\mathcal{V}^* \cap \ker E = 0$.

(iv) $E\mathcal{V}^* \oplus A\mathcal{W}_b^* = \mathcal{X}$.

Proof. (i)$\Leftrightarrow$(ii). From Theorem 4 we have that the rows and the columns of $(sE - A)$ are linearly independent over $R[s]$ if and only if $\mathcal{V}^* \oplus \mathcal{W}_b^* = \mathcal{X}$.

(ii)$\Rightarrow$(iii). Since $\mathcal{W}_b^* \supset \ker E$, we obviously have $\mathcal{V}^* \cap \ker E = 0$.

(iii)$\Rightarrow$(i). Since $\mathcal{V}^* \cap \ker E = 0$, then from Theorem 4 we obtain that the columns of $(sE - A)$ are linearly independent over $R[s]$. Since E and A are square maps it follows that the rows of $(sE - A)$ are also linearly independent over $R[s]$ and hence (i) follows.

(ii)$\Rightarrow$(iv). Note that $\mathcal{V}^* \cap \mathcal{W}_b^* = 0$ implies $\ker E \cap \mathcal{V}^* = 0$ and $\ker A \cap \mathcal{W}_b^* = 0$. Thus $\dim E\mathcal{V}^* = \dim \mathcal{V}^*$ and $\dim A\mathcal{W}_b^* = \dim \mathcal{W}_b^*$. Now suppose that

$$Ev = Aw, \quad v \in \mathcal{V}^*, \quad w \in \mathcal{W}_b^*.$$

This implies $w \in A^{-1}(E\mathcal{V}^*) = \mathcal{V}^*$, which is not possible. Thus $w = 0$ and hence $Ev = 0$. Since $\ker E \cap \mathcal{V}^* = 0$ we then have $v = 0$, so that (iv) is true.

(iv)$\Rightarrow$(i). Consider the decompositions (ii) and (iv) in the statement of the theorem and define the nonsingular map $M^{-1}: \mathcal{X} \to \mathcal{X}$ by

$$M^{-1}x = \begin{cases} Ex, & x \in \mathcal{V}^*, \\ Ax, & x \in \mathcal{W}_b^*. \end{cases} \tag{19}$$

Since $A\mathcal{V}^* \subset E\mathcal{V}^*$ and $E\mathcal{W}_b^* \subset A\mathcal{W}_b^*$ it now follows that the subspaces $\mathcal{V}^*$ and $\mathcal{W}_b^*$ are simultaneously MA- and ME-invariant, with

$$L := MA|\mathcal{V}^*, \qquad J := ME|\mathcal{W}_b^*, \tag{20}$$

$$ME|\mathcal{V}^* = I, \qquad MA|\mathcal{W}_b^* = I. \tag{21}$$

It now follows that in the decomposition (ii) the pencil $M(sE - A)$ admits the representation

$$\operatorname{Mat} M(sE - A) = \begin{bmatrix} sI - L & 0 \\ 0 & sJ - I \end{bmatrix}. \tag{22}$$

Since $\mathcal{V}^*$ is the supremum element of the family F_1 which is contained in $\mathcal{X}$ it follows that the map J is nilpotent ($MEw = \lambda w$, $0 \neq w \in \mathcal{W}_b^*$, $0 \neq \lambda \in C$ implies $Ew = \lambda M^{-1}w = \lambda Aw$, i.e. $w \in \mathcal{V}^*$ and hence $w = 0$).

Since J is nilpotent it follows that $\det(sJ - I) = (-1)^r$, $r := \dim \mathcal{W}_b^*$, and from (22) we conclude that the pencil is regular. □

Remarks. (i) The decomposition Theorem 5(ii) has also appeared in [9] and has been obtained by a different method.

(ii) The subspaces $\mathcal{V}^*$ and $\mathcal{W}_b^*$ in [5] have been described in terms of eigenspaces of $(\lambda E - A)^{-1}E$.

We now turn our attention to the autonomous generalized linear system

$$E\dot{x} = Ax \tag{23}$$

which is closely connected with the pencil $(sE - A)$. This is easily seen by taking the Laplace transform in (23) which yields

$$(sE - A)x(s) = Ex(0-)$$

where $x(0-)$ is the initial condition for (23).

When the pencil $(sE - A)$ is regular then the solution of (23) exists and is unique, namely

$$x(s) = (sE - A)^{-1}Ex(0-).$$

We shall concentrate on systems described by (23) whose associated pencil is regular.

Let M be the map defined from its inverse in (19) and consider the following maps:

$$Q_v: \mathscr{X} \to \mathscr{V}^*, \quad \text{the projection on } \mathscr{V}^* \text{ along } \mathscr{W}_b^* \tag{24}$$

and

$$Q_w: \mathscr{X} \to \mathscr{W}_b^*, \quad \text{the projection on } \mathscr{W}_b^* \text{ along } \mathscr{V}^*. \tag{25}$$

Premultiplying (23) by Q_vM and Q_wM, respectively, it follows that the system (23) can be decomposed as

$$\dot{x}_v = Lx_v, \qquad x_v \in \mathscr{V}^* \tag{26}$$

and

$$J\dot{x}_w = x_w, \qquad x_w \in \mathscr{W}_b^* \tag{27}$$

where L and J are as in (20).

The solution of (26) to an arbitrary initial condition $x_v(0^-)$ is well known and is determined by those finite-zeros of the pencil $(sE - A)$ which are excited (note in (22) that the eigenvalues of L coincide with the finite-zeros of $(sE - A)$).

The solution of (27) to an arbitrary initial condition $x_w(0^-) \in \mathscr{W}_b^*$ is a distribution (delta functional and its higher derivatives) and as pointed out in [13] we should consider the distributional differential equation

$$J\dot{x}_w = x_w + \delta x_w(0^-) \tag{28}$$

where δ denotes the delta functional.

Recall that J is a nilpotent map and let J be taken in the Jordan canonical form, i.e.,

$$J = \operatorname{diag}(J_1, \cdots, J_p, 0), \tag{29}$$

where

$$J_i = \begin{bmatrix} 0 & 1 & 0 & \cdot & \cdot & 0 \\ 0 & 0 & 1 & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & 1 \\ 0 & \cdot & \cdot & \cdot & \cdot & 0 \end{bmatrix}_{(n_i+1)\times(n_i+1)}$$

In (29) the map J is decomposed into $p (p \geqq 0)$ blocks J_i of size $n_i + 1$ and a zero block of dimension $m (m \geqq 0)$ so that

$$\sum_{i \in \underline{p}} (n_i + 1) + m = \dim \mathscr{W}_b^*.$$

It is well known [14] that the pencil $sJ - I$ has p infinite-zeros of respective orders n_i, $i \in \underline{p}$. The integers p, n_i and m will be identified in a moment.

Let $x_w^T = [x_1, \cdots, x_p, x_{p+1}]^T$, where x_i is a column vector of size $n_i + 1$, $i \in \underline{p}$, and x_{p+1} is a column vector of size m.

The block diagonal structure of (29) implies that the equation (28) becomes decomposed as

$$J_i\dot{x}_i = x_i + \delta J_i x_i(0^-), \quad i \in \underline{p} \tag{30}$$

and a trivial equation $x_{p+1} = 0$.

The variables x_{p+1} are called static variables in the sense that they exhibit no dynamical behaviour. It has been shown in [13] that the distributional solution of (30) is given by

$$\begin{aligned} x_{i,1} &= -x_{i,2}(0^-)\delta - \cdots - x_{i,n_i+1}(0^-)\delta^{(n_i-1)}, \\ &\vdots \\ x_{i,n_i} &= -x_{i,n_i+1}(0^-)\delta, \\ x_{i,n_i+1} &= 0, \end{aligned} \tag{31}$$

where $\delta^{(h)}$ denotes the distributional derivative of order h of δ.

The subspace in which the distributional solution (31) takes place will be identified from a result of the next theorem which describes some structural features of the regular pencil $(sE-A)$.

THEOREM 6. *Let $(sE-A)$ be a regular pencil. Then:*

(a) *There are* $\dim \mathcal{V}^*$ *finite-zeros which are the eigenvalues of the map L defined in* (20).

(b) *There are* $p := \dim(\ker E \cap A^{-1}(\operatorname{Im} E))$ *infinite-zeros of respective orders* n_i, $i \in \underline{p}$, *which are determined from the dimensions of the subspaces* $\mathcal{W}_a^u$ *in* (2) (*with* $\mathcal{K} = A^{-1}(\operatorname{Im} E)$). *Further*

$$\sum_{i \in \underline{p}} n_i = \dim \mathcal{W}_a^*.$$

(c)

$$A^{-1}(\operatorname{Im} E) = \mathcal{V}^* \oplus \mathcal{W}_a^* \tag{32}$$

and

$$\operatorname{Im} E = E\mathcal{V}^* \oplus A\mathcal{W}_a^*. \tag{33}$$

(d) *Let* $\mathcal{N} \subset \ker E$ *be any subspace such that*

$$\mathcal{N} \oplus (\ker E \cap A^{-1}(\operatorname{Im} E)) = \ker E. \tag{34}$$

Then there are $m := \dim \mathcal{N}$ *static variables which do not play any role in the determination of the zero structure of* $(sE-A)$.

Proof. (a) See (22).

(b) Since the map J is nilpotent it follows that there exists a basis $\{w_{i,j}\}$ for $\mathcal{W}_b^*$ such that

$$Ew_{i,1} = 0; \quad Ew_{i,j} = Aw_{i,j-1},\ i \in \underline{p}, j \in \{2, \cdots, n_i+1\} \tag{35}$$

and

$$Ew_{i,1} = 0,\ i \in \{p+1, \cdots, p+m\}. \tag{36}$$

It is clear from (35) that for $i \in \underline{p}$, $\operatorname{span}\{w_{i,1}\} = \ker E \cap A^{-1}(\operatorname{Im} E)$. From (35) and (2) it also follows that $w_{i,j} \in \mathcal{W}_a^j$, $i \in \underline{p}$, $j \in \underline{n_i}$. Let $\mathcal{W}_a := \operatorname{span}\{w_{i,j}\}$, $i \in \underline{p}$, $j \in \underline{n_i}$. Hence $\mathcal{W}_a \subset \mathcal{W}_a^*$.

From (35) and (36) it is easy to see that $\dim \mathcal{W}_a = \dim \mathcal{W}_b^* - \dim \ker E$. From Lemma (1e) we obtain that $E\mathcal{W}_b^* = A\mathcal{W}_a^*$ and since the pencil is regular it follows that

$$\dim \mathcal{W}_a^* = \dim \mathcal{W}_b^* - \dim \ker E.$$

Therefore $\mathcal{W}_a = \mathcal{W}_a^*$ and since $w_{i,j} \in \mathcal{W}_a^j$, $i \in \underline{p}$, $j \in \underline{n}_i$, it follows that the orders of the infinite-zeros can be determined in the following way. Let

$$\phi_u := \dim \left[\frac{\mathcal{W}_a^u}{\mathcal{W}_a^{u-1}} \right], \quad u \in \underline{n}$$

and define

$$n_i := \text{number of integers in the set } \{\phi_1, \phi_2, \cdots, \phi_n\} \text{ which are } \geqq i.$$

Then

$$n_1 \geqq n_2 \geqq \cdots \geqq n_p \geqq 1 \quad \text{with} \sum_{i \in \underline{p}} n_i = \dim \mathcal{W}_a^*.$$

(c) First note that $\mathcal{V}^* \oplus \mathcal{W}_a^* \subset A^{-1}(\operatorname{Im} E)$ and let $\tilde{\mathcal{W}}_b$ be any subspace such that $\mathcal{W}_b^* = \mathcal{W}_a^* \oplus \tilde{\mathcal{W}}_b$. From Lemma 1(c), $\tilde{\mathcal{W}}_b \cap A^{-1}(\operatorname{Im} E) = 0$. Thus $\mathcal{X} = \mathcal{V}^* \oplus \mathcal{W}_a^* \oplus \tilde{\mathcal{W}}_b$ and then (32) follows.

To show (33), note from Theorem 5(ii), (iv) and Lemma 1(e), (c) that

$$\operatorname{Im} E = E\mathcal{X} = E\mathcal{V}^* + E\mathcal{W}_b^* = E\mathcal{V}^* \oplus A\mathcal{W}_a^*.$$

(d) It is clear from (36) that for $i \in \{p+1, \cdots, p+m\}$ we have that span $\{w_{i,1}\} = \mathcal{N}$, where $\mathcal{N}$ is as in (34).

Consider $\mathcal{W}_a^1 = \ker E \cap A^{-1}(\operatorname{Im} E)$ and let $\mathcal{C}_1$ be any subspace such that

$$\mathcal{X} = \mathcal{C}_1 \oplus \mathcal{W}_a^1 \oplus \mathcal{N}. \tag{37}$$

Note that $A\mathcal{W}_a^1 \subset \operatorname{Im} E$ and that $A\mathcal{N} \cap \operatorname{Im} E = 0$. Since the pencil is regular we also have that $\dim A\mathcal{N} = \dim \mathcal{N}$.

Let $\mathcal{C}_2$ be any subspace which yields a direct sum for the following decomposition

$$\mathcal{X} = \operatorname{Im} E \oplus \mathcal{C}_2 \oplus A\mathcal{N}. \tag{38}$$

Let $x \in \mathcal{X}$ be represented in the decomposition (37) and Ex and Ax be represented in the decomposition (38). Hence

$$\operatorname{Mat} E = \begin{bmatrix} I & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad \operatorname{Mat} A = \begin{bmatrix} A_{11} & A_{12} & 0 \\ A_{21} & 0 & 0 \\ A_{31} & 0 & I \end{bmatrix} \tag{39}$$

where the identity matrix in Mat E has dim Im E and the identity matrix in Mat A has dim m.

Consider the pencil $(sE - A)$ with E and A as in (39) and postmultiply it by the nonsingular matrix

$$\begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ -A_{31} & 0 & I \end{bmatrix}.$$

This operation does not alter the zero structure of $(sE - A)$ [7] and the resulting pencil is represented as

$$\left[\begin{array}{cc|c} sI - A_{11} & -A_{12} & 0 \\ -A_{21} & 0 & 0 \\ \hline 0 & 0 & I \end{array}\right].$$

It is clear from the above block diagonal structure that the zero structure of $(sE-A)$ can be obtained from the reduced pencil

$$\begin{bmatrix} sI-A_{11} & -A_{12} \\ -A_{21} & 0 \end{bmatrix}. \qquad \square$$

COROLLARY 2. (a) *The number of zeros of the regular pencil is* $\dim E$.

(b) *The distributional response* (31) *belongs to* $\mathcal{W}_a^*$.

Proof. (a) This fact is well known [13], and follows from Theorem 6(a), (b), (c).

(b) By using (35) it follows that the components $x_{i,j}$, $i \in \underline{p}$, $j \in \underline{n}_i$ belong to $\mathcal{W}_a^*$.

Comment. From Corollary 2(b) and (26), it follows that the total response of (23) to an arbitrary $x(0^-)$ belongs to $A^{-1}(\operatorname{Im} E)$ (see (32)). This has a simple interpretation for a discrete time generalized linear system $Ex(k+1)=Ax(k)$, where $x(k)$ must belong to $A^{-1}(\operatorname{Im} E)$ in order for $x(k+1)$ to exist.

Other consequences are stated next.

COROLLARY 3. (a) *The regular pencil* $(sE-A)$ *has no infinite-zeros if and only if* $\ker E \cap A^{-1}(\operatorname{Im} E)=0$.

(b) *Suppose that the regular pencil* $(sE-A)$ *has infinite-zeros. Then there exist no static variables if and only if* $A \ker E \subset \operatorname{Im} E$.

Proof. (a) From Theorem 6(b) we have that $(sE-A)$ has no infinite-zeros if and only if $\mathcal{W}_a^*=0$, i.e. $\ker E \cap A^{-1}(\operatorname{Im} E)=0$.

(b) Since $(sE-A)$ has infinite-zeros then we must have $\ker E \cap A^{-1}(\operatorname{Im} E) \neq 0$. From (34) we then obtain that $(sE-A)$ has no static variables if and only if $\mathcal{N}=0$, i.e. $\ker E \cap A^{-1}(\operatorname{Im} E)=\ker E$, or equivalently $\ker E \subset A^{-1}(\operatorname{Im} E) \Leftrightarrow A \ker E \subset \operatorname{Im} E$. $\square$

Corollary 3 will be useful in the following section.

5. Controllability and observability for generalized linear systems. In this section we consider the generalized linear system Σ:

$$\Sigma: E\dot{x}=Ax+Bu, \qquad y=Cx, \tag{40}$$

such that the associated pencil $(sE-A)$ is regular, $x \in \mathcal{X} := R^n$, $u \in \mathcal{U} := R^m$, $y \in \mathcal{Y} := R^r$.

Consider the maps Q_v and Q_w in (24) and (25) and let M^{-1} be the map defined in (19).

Premultiplying (40) by Q_vM and Q_wM, respectively, it follows that in the decomposition of the Theorem 5(ii) the system (40) becomes decomposed as

$$\dot{x}_v = Lx_v + B_v u, \qquad x_v \in \mathcal{V}^*, \tag{41}$$

$$J\dot{x}_w = x_w + B_w u, \qquad x_w \in \mathcal{W}_b^*, \tag{42}$$

$$y = C_v x_v + C_w x_w,$$

where

$$B_v = Q_v MB, \qquad B_w = Q_w MB,$$

$$C_v = C|\mathcal{V}^*, \qquad C_w = C|\mathcal{W}_b^*.$$

In the following we describe the modal criterion for controllability and observability of the zeros of the pencil $(sE-A)$ [14]. Accordingly, we shall say that the system Σ is controllable or observable.

Test 1. The system Σ is controllable if and only if the pencil $[sE-A \;\; B]$ has no zeros (finite or infinite).

Test 2. The system Σ is observable if and only if the pencil $\begin{bmatrix} sE-A \\ C \end{bmatrix}$ has no zeros (finite or infinite).

From the above tests and the fact that J is nilpotent, it follows immediately that Σ is controllable and observable at its finite-zeros if and only if the pencils $[sI - L \;\; B_v]$ and $\begin{bmatrix} sI-L \\ C_v \end{bmatrix}$ have no finite zeros.

We first give a trivial interpretation for the reachable and unobservable subspaces corresponding to those controllable and unobservable finite-zeros.

PROPOSITION 2. *Let $\mathcal{V}^*$ be the subspace given by the sequence* (1). *Then*:

(a) *The reachable subspace is the least subspace $\mathcal{V}_c \subset \mathcal{V}^*$ such that*

$$A\mathcal{V}_c \subset E\mathcal{V}_c \quad \textit{and} \quad E\mathcal{V}_c \oplus A\mathcal{W}_b^* \supset \mathcal{B}.$$

(b) *The unobservable subspace is the largest subspace $\mathcal{V}_u \subset \mathcal{V}^*$ such that*

$$A\mathcal{V}_u \subset E\mathcal{V}_u \quad \textit{and} \quad \mathcal{V}_u \subset \ker C.$$

Proof. (a) Let $\mathcal{B} := \operatorname{Im} B$. It is well known that the reachable subspace is the least L-invariant subspace $\mathcal{V}_c$ which contains $\mathcal{B}_v = Q_v M\mathcal{B}$, i.e.,

$$L\mathcal{V}_c \subset \mathcal{V}_c \tag{43}$$

and

$$\mathcal{V}_c \supset Q_v M\mathcal{B}. \tag{44}$$

But (43) is equivalent to $A\mathcal{V}_c \subset E\mathcal{V}_c$ and by using (21) and (19) it follows from (44) that

$$Q_v M\mathcal{B} \subset Q_v ME\mathcal{V}_c \Leftrightarrow \mathcal{B} \subset (Q_v M)^{-1} Q_v ME\mathcal{V}_c = E\mathcal{V}_c + M^{-1}\mathcal{W}_b^* = E\mathcal{V}_c \oplus A\mathcal{W}_b^*.$$

(b) The unobservable subspace is the largest L-invariant subspace $\mathcal{V}_u$ which is contained in $\ker C_v$ which is equivalent to $A\mathcal{V}_u \subset E\mathcal{V}_u$ and $\mathcal{V}_u \subset \ker C$. □

Remark. Note that when $\mathcal{V}^* = \mathcal{X}$ (therefore $\mathcal{W}_b^* = 0$) and $E = I$, the above proposition merely gives the reachable and unobservable subspaces for the linear system $\dot{x} = Ax + Bu$; $y = Cx$.

Hereafter we shall concentrate on the study of controllability and observability of the infinite-zeros for the system Σ. Note that the infinite-zeros of the pencil $(sE - A)$ are situated in the pencil $(sJ - I)$. Thus the analysis of dynamic properties of the infinite-zeros, such as controllability, must be carried out on the subsystem

$$J\dot{x}_w = x_w + B_w u. \tag{45}$$

Let q be the index of nilpotency of J, namely, q is the least positive integer such that $J^q = 0$. Campbell [4] has shown that if the control u belongs to the class of the $q-1$ differentiable functions, then the unique solution of (45) in the class of functions is given by

$$x_w(t) = \sum_{i=0}^{q-1} J^i B_w u^{(i)}(t), \qquad t \geqq 0$$

where $u^{(i)}(t)$ denotes the ith derivative of $u(t)$.

Cobb [6] has shown that the reachable subspace $\mathcal{R}_w$ relative to the system (45) is given by

$$\mathcal{R}_w = \langle J | \mathcal{B}_w \rangle := \mathcal{B}_w + J\mathcal{B}_w + \cdots + J^{q-1}\mathcal{B}_w$$

with $\mathcal{B}_w = \operatorname{Im} B_w$. The reachability property implies that any point in $\mathcal{R}_w$ can be reached at any time t by using a suitable control u together with its derivatives.

It is shown next that we may have a situation where $\mathcal{R}_w \subset \mathcal{W}_b^*$ and yet all the infinite-zeros are controllable.

5.1. Controllability of the infinite-zeros. We shall show that controllability of the infinite-zeros is equivalent to reachability of certain quotient spaces. Let $J: \mathcal{W}_b^* \to \mathcal{W}_b^*$ be the map defined in (20) and let $\mathcal{W}_1 \subset \ker J$ be any space such that

$$\mathcal{W}_1 \oplus (\ker J \cap \operatorname{Im} J) = \ker J. \tag{46}$$

Note that $\mathcal{W}_1$ provides only simple eigenvectors to the nilpotent map J, in the sense that if $\{w_i\}$, $i \in \underline{t}$, $t := \dim \mathcal{W}_1$, is a basis for $\mathcal{W}_1$ then $Jw_i = 0$, $i \in \underline{t}$ and there exists no chain of eigenvectors starting from w_i, $i \in \underline{t}$. In other words, the subspace $\mathcal{W}_1$ can be associated with all the simple elementary divisors of J.

Let $\bar{\mathcal{W}} := \mathcal{W}_b^* (\operatorname{mod} \mathcal{W}_1)$ and let $P := \mathcal{W}_b^* \to \bar{\mathcal{W}}$ be the canonical projection. Let $\bar{J}$ be the unique map induced in $\bar{\mathcal{W}}$ such that $\bar{J}P = PJ$ and let $\bar{B}_w := B_w (\operatorname{mod} \mathcal{W}_1): \mathcal{U} \to \bar{\mathcal{W}}$.

Matrix representations for $\bar{J}$ and $\bar{B}_w$ can be readily obtained. For this, let $\mathcal{W}_2$ be any subspace such that $\mathcal{W}_b^* = \mathcal{W}_1 \oplus \mathcal{W}_2$. Then in this decomposition

$$\operatorname{Mat} J = \begin{bmatrix} 0 & J_{12} \\ 0 & J_{22} \end{bmatrix}, \qquad \operatorname{Mat} B_w = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}, \tag{47}$$

with $\operatorname{Mat} \bar{J} = J_{22}$ and $\operatorname{Mat} \bar{B}_w = B_2$.

Note from (46) that $J\mathcal{W}_1 = 0 \subset \mathcal{W}_1$ and there is a subspace $\hat{\mathcal{W}}_2$ such that $J\hat{\mathcal{W}}_2 \subset \hat{\mathcal{W}}_2$ and $\mathcal{W}_1 \oplus \hat{\mathcal{W}}_2 = \mathcal{W}_b^*$. The last statement can be readily verified by thinking of the eigenvector chains of J and (46). Thus by [18, Prop. 0.5], it follows that the elementary divisors of $J|\mathcal{W}_1$ with those of $\bar{J}$, together, give all the elementary divisors of J. Hence $\bar{J}$ is a map which possesses all the elementary divisors of order greater than one of J.

Let $\hat{\mathcal{W}}_2$ be a subspace as described above. Then $J_{12} = 0$ in (47) and the system (45) can be decomposed in the following way:

$$x_{w1} = -B_1 u \tag{48}$$

and

$$J_{22}\dot{x}_{w2} = x_{w2} + B_2 u \tag{49}$$

where $x_{w1} \in \mathcal{W}_1$ and $x_{w2} \in \hat{\mathcal{W}}_2$.

Note that the subsystem (48) is "static", in the sense that at each instant t, $x_{w1}(t)$ is a linear combination of the control variables $u(t)$. We recall that the variables x_{w1} have the same nature of the variables described in the previous section as static variables.

Test 4.2 in [13] gives a necessary and sufficient condition for the infinite-zeros to be observable. The dual of that test gives a necessary and sufficient condition for the controllability of those zeros and it is described next.

Test 3. Apply nonsingular transformations on the right of the pencil $[sE - A \;\; B]$ so as to bring it to the form $[sE_1 - A_1 \;\; A_2 \;\; B_2]$, where E_1 has full column rank and $A = [A_1 \;\; A_2]$. Then the system Σ is controllable at its infinite-zeros, i.e., the pencil $[sE - A \;\; B]$ has no infinite-zeros, if and only if

$$\operatorname{Im} E_1 + \operatorname{Im} A_2 + \mathcal{B} = \mathcal{X}. \tag{50}$$

We are now in a position to link the reachability concept described previously and the modal controllability concept introduced in Test 3.

THEOREM 7. *$\bar{\mathcal{W}}$ is reachable, namely $\langle \bar{J} | \bar{\mathcal{B}}_w \rangle = \bar{\mathcal{W}}$ if and only if the system* (45) *is controllable* (*at infinity*).

Proof. It is well known [18] that $\langle \bar{J} | \bar{\mathcal{B}}_w \rangle = \bar{\mathcal{W}}$ if and only if

$$\operatorname{Im} (\lambda I - \bar{J}) + \bar{\mathcal{B}}_w = \bar{\mathcal{W}} \quad \forall \lambda \in C. \tag{51}$$

Since $\bar{J}$ is nilpotent it is clear that (51) holds for $\lambda \neq 0$. Therefore

$$\langle \bar{J} | \bar{\mathcal{B}}_w \rangle = \bar{\mathcal{W}} \Leftrightarrow \operatorname{Im} \bar{J} + \bar{\mathcal{B}}_w = \bar{\mathcal{W}}. \tag{52}$$

Taking $\bar{J}$ in Jordan canonical form and applying Test 3 to the pencil $[s\bar{J} - I \ \bar{B}_w]$ it is readily seen that the condition (50) reduces to (52). □

Based on the result of Theorem 7, a procedure can be given to obtain the controllable subsystem of a system given by (45).

1. Choose any subspace $\mathcal{W}_1$ according to (46) and take any subspace $\mathcal{W}_2$ such that $\mathcal{W}_1 \oplus \mathcal{W}_2 = \mathcal{W}_b^*$.

2. Compute the pair (J_{22}, B_2).

3. Compute the least J_{22}-invariant subspace $\mathcal{W}_c$ which contains $\operatorname{Im} B_2$, i.e. $\mathcal{W}_c = \langle J_{22} | \operatorname{Im} B_2 \rangle$. By taking $\mathcal{W}_c$ as part of a basis for $\mathcal{W}_2$ we then obtain the following representations

$$\operatorname{Mat} J_{22} = \begin{bmatrix} \hat{J}_{11} & \hat{J}_{12} \\ 0 & \hat{J}_{22} \end{bmatrix}, \qquad \operatorname{Mat} B_2 = \begin{bmatrix} \hat{B}_1 \\ 0 \end{bmatrix},$$

so that the subsystem $J_{22}\hat{x} = \hat{x}_2$ is uncontrollable and $\hat{J}_{11}\hat{x}_1 = \hat{x}_1 + \hat{B}_1 u$ is controllable.

Remark. The procedure above described to obtain a controllable subsystem is not based on a canonical form as in reference [13] whose authors have pointed out the need for a geometric language to describe the concepts of controllability and observability of the infinite-zeros.

The next theorem contains a necessary and sufficient condition for the controllability of Σ at infinity.

THEOREM 8. *The system Σ is controllable at infinity if and only if*

$$\operatorname{Im} E + A \ker E + \mathcal{B} = \mathcal{X}. \tag{53}$$

Proof. Let $\mathcal{W}_1$ be a subspace as in (46) and let $\hat{\mathcal{W}}_2$ be a subspace such that $\mathcal{W}_1 \oplus \hat{\mathcal{W}}_2 = \mathcal{W}_b^*$ and $J\hat{\mathcal{W}}_2 \subset \hat{\mathcal{W}}_2$. This implies $J_{12} = 0$ in (47) and note that (J_{22}, B_2) is the pair induced in $\hat{\mathcal{W}}_2$. The space $\mathcal{X}$ is now decomposed as

$$\mathcal{X} = \mathcal{V}^* \oplus \mathcal{W}_1 \oplus \hat{\mathcal{W}}_2. \tag{54}$$

Let M^{-1} be the map defined in (19). Since $\operatorname{Im} E = E\mathcal{X}$ and $\mathcal{W}_1 \subset \ker J = \ker ME | \mathcal{W}_b^* = \ker E \subset \mathcal{W}_b^*$, it follows that the premultiplication of the left-hand side of (53) by M results in

$$ME(\mathcal{V}^* \oplus \hat{\mathcal{W}}_2) + MA \ker E + M\mathcal{B}.$$

By using (20)-(21) and (46)-(47) it follows that the above expression is equal to

$$\mathcal{V}^* \oplus \operatorname{Im} J_{22} + (\mathcal{W}_1 \oplus \ker J \cap \operatorname{Im} J) + \operatorname{Im} B_2. \tag{55}$$

But $\ker J \cap \operatorname{Im} J = \ker J_{22}$ and since J_{22} is nilpotent it follows that $\ker J_{22} \subset \operatorname{Im} J_{22}$. Thus (55) can be rewritten as

$$\mathcal{V}^* \oplus \mathcal{W}_1 \oplus (\operatorname{Im} J_{22} + \operatorname{Im} B_{22}).$$

Hence $M(\operatorname{Im} E + A \ker E + \mathcal{B}) = \mathcal{X}$ if and only if $\operatorname{Im} J_{22} + \operatorname{Im} B_2 = \hat{\mathcal{W}}_2$ or by (52) if and only if the infinite-zeros are controllable. □

Remarks. (a) Note that when the regular pencil has no infinite-zeros, then by Corollary 3(a), $A \ker E \cap \operatorname{Im} E = 0$ so that $\operatorname{Im} E \oplus A \ker E = \mathcal{X}$.

This simply means that absence of infinite-zeros in the pencil $(sE - A)$, obviously implies absence of infinite-zeros in the pencil $[sE - A \ B]$.

If $(sE-A)$ has infinite-zeros but has no static variables, then by using Corollary 3(b), the condition (53) reduces to $\operatorname{Im} E+\mathscr{B}=\mathscr{X}$, which is the condition given by Rosenbrock [10] and Cobb [5]. This shows that their condition is correct if no static variables are present.

(b) From (53) and (20)-(21) it follows that

$$\operatorname{Im} J+\ker J+\mathscr{B}_w=\mathscr{W}_b^*. \tag{56}$$

Thus conditions (52), (53) and (56) are equivalent necessary and sufficient conditions for controllability of Σ at infinity. The condition (56) has shown up in [5] in connection with the existence of a state feedback map that converts the infinite-zeros into finite ones. (See also [2] on this subject).

5.2. Observability of the infinite-zeros. Consider the system

$$J\dot{x}_w=x_w, \qquad y=C_w x_w. \tag{57}$$

Write again $\mathscr{W}_b^*=\mathscr{W}_1\oplus\hat{\mathscr{W}}_2$, where $\mathscr{W}_1$ is as in (46) and $\hat{\mathscr{W}}_2$ is such that $J\hat{\mathscr{W}}_2\subset\hat{\mathscr{W}}_2$. Then by using this decomposition for $\mathscr{W}_b^*$ it follows that (57) can be written as

$$\begin{bmatrix}0 & 0\\ 0 & J_{22}\end{bmatrix}\begin{bmatrix}\dot{x}_{w1}\\ \dot{x}_{w2}\end{bmatrix}=\begin{bmatrix}x_{w1}\\ x_{w2}\end{bmatrix}, \qquad y=[C_{w1} \quad C_{w2}]\begin{bmatrix}x_{w1}\\ x_{w2}\end{bmatrix}.$$

Hence $x_{w1}=0$ and $y=C_{w2}x_{w2}$. The system (57) is said to be observable at infinity if there are no impulsive motions in $\ker C_{w2}$. This is simply an extension of the observability criterion for exponential modes (caused by the finite-zeros).

Let $\langle\ker C_{w2}|J_{22}\rangle:=\ker C_{w2}\cap J_{22}^{-1}\ker C_{w2}\cap\cdots\cap J_{22}^{-q+1}\ker C_{w2}$, where q is the index of nilpotency of J_{22} (and of J). We then obtain:

THEOREM 9. $\langle\ker C_{w2}|J_{22}\rangle=0$ *if and only if the system* (57) *is observable* (*at infinity*).

Proof. Identical to the proof of Theorem 7 on considering the pair (J_{22}^T, C_{w2}^T) and noting that $\langle\ker C_{w2}|J_{22}\rangle=0$ if and only if $\langle J_{22}^T|\operatorname{Im} C_{w2}^T\rangle=\hat{\mathscr{W}}_2'$. □

Remark. By using a duality argument it follows that the subspace $\langle\ker C_{22}|J_{22}\rangle$ corresponds to those infinite-zeros which are unobservable and this fact can be used in a procedure to extract the observable infinite-zeros.

The next theorem shows how observability can be characterized directly in terms of the maps E, A and C.

THEOREM 10[1]. *The system* Σ *is observable at infinity if and only if*

$$\ker E\cap A^{-1}(\operatorname{Im} E)\cap\ker C=0. \tag{58}$$

Proof. Let $H(s):=\begin{bmatrix}sE-A\\ C\end{bmatrix}$. Since the zero structure of $H(s)$ coincides with the zero structure of $[sE^T-A^T \quad C^T]$ [14], we then obtain from Theorem 8 that the pencil $H(s)$ has no infinite-zeros if and only if

$$\operatorname{Im} E^T+A^T\ker E^T+\operatorname{Im} C^T=\mathscr{X}'.$$

By dualizing this last relation we then obtain (58). □

Remarks. (a) Note from Corollary 3(b) that if the pencil has no static variables then (58) reduces to $\ker E\cap\ker C=0$ and if the pencil has no infinite-zeros then by Corollary 3(a), $\ker E\cap A^{-1}(\operatorname{Im} E)=0$, which means that absence of infinite-zeros in the pencil $(sE-A)$ implies the same in the pencil $H(s)$.

[1] A referee has pointed out he believes that Theorems 8 and 10 are contained in a paper by D. J. Cobb (to appear in IEEE Trans. Automat. Control). Such theorems have appeared in the Ph.D. thesis (September, 1983) of the author of this paper, and since then the author has not been aware of any published work containing the mentioned theorems.

(b) Let $\{w_i\}$, $i \in \underline{p}$, be a basis for $\mathcal{W}_a^1 = \ker E \cap A^{-1}(\operatorname{Im} E)$. Then from (58) it follows that the system Σ is observable at infinity if and only if the vectors Cw_i, $i \in \underline{p}$, are linearly independent. This is equivalent to the test given by Verghese [14] which states that the columns of C corresponding to the first position of each Jordan block of order greater than one of J are linearly independent if and only if Σ is observable at infinity.

6. Appendix.

Proof of Lemma 1. (a) It is proven in [3].

(b) Note that $\mathcal{T}^1 = \mathcal{V}^* \cap \ker E = \mathcal{V}^* \cap \mathcal{W}_b^1$.

Assume that $\mathcal{T}^{u-1} = \mathcal{V}^* \cap \mathcal{W}_b^{u-1}$. Then

$$\begin{aligned}
\mathcal{T}^u &= \mathcal{V}^* \cap E^{-1}(A(\mathcal{V}^* \cap \mathcal{W}_b^{u-1})) \\
&= \mathcal{V}^* \cap E^{-1}(A(A^{-1}(E\mathcal{V}^*) \cap \mathcal{W}_b^{u-1})) \\
&= \mathcal{V}^* \cap E^{-1}(A\mathcal{W}_b^{u-1} \cap E\mathcal{V}^*) \\
&= \mathcal{V}^* \cap E^{-1}(A\mathcal{W}_b^{u-1}) \cap E^{-1}(E\mathcal{V}^*) \\
&= \mathcal{V}^* \cap (E^{-1}(A\mathcal{W}_b^{u-1}) \cap (\mathcal{V}^* + \ker E)) \\
&= \mathcal{V}^* \cap (E^{-1}(A\mathcal{W}_b^{u-1}) \cap \mathcal{V}^* + \ker E) \\
&= E^{-1}(A\mathcal{W}_b^{u-1}) \cap \mathcal{V}^* + \mathcal{V}^* \cap \ker E \\
&= E^{-1}(A\mathcal{W}_b^{u-1}) \cap \mathcal{V}^* = \mathcal{V}^* \cap \mathcal{W}_b^u.
\end{aligned}$$

(c) The proof is analogous to (b) if we replace $\mathcal{V}^*$ by $\mathcal{K} = A^{-1}(\operatorname{Im} E)$.

(d) From (b) and (c), $\mathcal{T}^u = \mathcal{V}^* \cap \mathcal{K} \cap \mathcal{W}_b^u = \mathcal{V}^* \cap \mathcal{W}_a^u$.

(e) By using (c) we obtain

$$A\mathcal{W}_a^{u-1} = A(A^{-1}(\operatorname{Im} E) \cap \mathcal{W}_b^{u-1}) = A\mathcal{W}_b^{u-1} \cap \operatorname{Im} E$$

and

$$E\mathcal{W}_b^u = E(E^{-1}(A\mathcal{W}_b^{u-1})) = A\mathcal{W}_b^{u-1} \cap \operatorname{Im} E. \qquad \square$$

Proof of Theorem 1. We first show that any subspace of the family F_3 is contained in $\mathcal{T}^*$.

Note that

$$t_{i,0} \in \ker E \cap \mathcal{T} \subset \ker E \cap \mathcal{V}^* \subset \mathcal{T}^*$$

and suppose that $t_{i,j-1} \in \mathcal{T}^*$, $j \in \{0, 1, \cdots, h_i - 1\}$. Then

$$t_{i,j} \in E^{-1}(A\mathcal{T}^*) \cap \mathcal{T} \subset E^{-1}(A\mathcal{T}^*) \cap \mathcal{V}^* = \mathcal{T}^*.$$

Finally,

$$t_{i,h_i} \in E^{-1}(A\mathcal{T}^*) \cap \ker A \subset E^{-1}(A\mathcal{T}^*) \cap \mathcal{V}^* = \mathcal{T}^*.$$

To show that the subspace $\mathcal{T}^*$ is the supremum of the family F_3 it then suffices to write $\mathcal{T}^*$ as in (8). For this, consider the sequence $\mathcal{T}^u$ in (3) and define

$$\rho_u := \dim\left[\frac{\mathcal{T}^u}{\mathcal{T}^{u-1}}\right], \qquad u \in \underline{n}. \tag{A.1}$$

By Lemma 1(a), $E\mathcal{T}^u = A\mathcal{T}^{u-1}$, which implies

$$\rho_u = \dim(\ker E \cap \mathcal{T}^u) - \dim(\ker A \cap \mathcal{T}^{u-1}). \tag{A.2}$$

From Lemma 1(b), it follows that $\ker E \cap \mathcal{T}^u = \ker E \cap \mathcal{V}^* = \ker E \cap \mathcal{T}^*$, so that $\dim(\ker E \cap \mathcal{T}^u) = l$, $u \in \underline{n}$. By using (A.2) we also obtain

$$\rho_u = \rho_{u+1} + \dim\left[\frac{\ker A \cap \mathcal{T}^u}{\ker A \cap \mathcal{T}^{u-1}}\right],$$

and since $\mathcal{T}^u \supset \mathcal{T}^{u-1}$, the result is $\rho_u \geqq \rho_{u+1}$.

Let $q \in \underline{n}$ be such that $\mathcal{T}^q \cap \ker A = \mathcal{T}^* \cap \ker A$. Since $\dim(\ker E \cap \mathcal{T}^*) = \dim(\ker A \cap \mathcal{T}^*)$ we then obtain from (A.1) and (A.2) that

$$\rho_1 + \rho_2 + \cdots + \rho_q = \dim \mathcal{T}^* \quad \text{and} \quad \rho_{q+1} = \rho_{q+2} = \cdots = \rho_n = 0.$$

Let $\mathcal{M}_u := \ker A \cap \mathcal{T}^u$, $u \in \underline{q}$. Since $\mathcal{T}^{u-1} \subset \mathcal{T}^u$, then $\mathcal{M}_1 \subset \mathcal{M}_2 \subset \cdots \subset \mathcal{M}_q$. Let $\mathcal{M}'_u \subset \mathcal{M}_u$ be such that

$$\mathcal{M}_{u-1} \oplus \mathcal{M}'_u = \mathcal{M}_u, \qquad u \in \{2, 3, \cdots, q\}.$$

Hence we can decompose $\mathcal{M}_u$ as

$$\mathcal{M}_u = \mathcal{M}'_1 \oplus \mathcal{M}'_2 \oplus \cdots \oplus \mathcal{M}'_u, \qquad u \in \underline{q} \tag{A.3}$$

with $\mathcal{M}'_1 := \mathcal{M}_1$.

Let $d_u := \dim \mathcal{M}'_u$, $u \in \underline{q}$ and let $X_{u,u-1}$ be a matrix of dimension $n \times d_u$ such that

$$\operatorname{Im} X_{u,u-1} = \mathcal{M}'_u, \qquad u \in \underline{q}. \tag{A.4}$$

Since $E\mathcal{T}^u = A\mathcal{T}^{u-1}$, it follows that the equations

$$EX_{u,j} = AX_{u,j-1}, \qquad u \in \{2, \cdots, q\}, \quad j \in \{1, \cdots, u-1\} \tag{A.5}$$

can be solved backwards for matrices $X_{u,j-1}$ of dimension $n \times d_u$ such that $\operatorname{Im} X_{u,j-1} \subset \mathcal{T}^{j-1}$ with

$$EX_{u,0} = 0 \quad \text{and} \quad AX_{u,u-1} = 0. \tag{A.6}$$

We need the following result.

LEMMA A. *The matrices $X_{u,j-1}$, $u \in \{2, \cdots, q\}$, $j \in \{1, \cdots, u-1\}$ are monic. The subspaces $\operatorname{Im} X_{u,j}$, $u \in \underline{q}$, $j \in \{0, \cdots, u-1\}$ are independent and*

$$\mathcal{T}^* = \sum_{u=1}^{q} \sum_{j=0}^{u-1} \operatorname{Im} X_{u,j}.$$

Proof. We shall show that $\ker E \cap \operatorname{Im} X_{u,j} = 0$. This will imply that the solution matrices $X_{u,j-1} = 0$ are monic and $\ker A \cap \operatorname{Im} X_{u,j-1} = 0$.

The proof is by induction. Suppose there exists $x_{u-1} \in \operatorname{Im} X_{u,u-1} = \mathcal{M}'_u$ such that $Ex_{u-1} = 0$.

Since $x_{u-1} \in \ker A$ we then obtain $x_{u-1} \in \ker E \cap \mathcal{T}^* \cap \ker A = \mathcal{M}_1$. But $x_{u-1} \in \mathcal{M}'_u$ and $\mathcal{M}_1 \cap \mathcal{M}'_u = 0$. Thus $x_{u-1} = 0$.

Now assume that $\ker E \cap \operatorname{Im} X_{u,j} = 0$, $j \in \{u-1, u-2, \cdots, 2\}$ and suppose there exists a vector $x_{j-1} \in \operatorname{Im} X_{u,j-1}$ such that $Ex_{j-1} = 0$. This implies that there exist vectors $x_j, \cdots, x_{u-1}$ such that

$$\begin{aligned} Ex_j &= Ax_{j-1}, \qquad x_j \in \operatorname{Im} \mathcal{X}_{u,j}, \\ &\vdots \\ Ex_{u-1} &= Ax_{u,2}, \\ Ax_{u-1} &= 0, \qquad x_{u-1} \in \mathcal{M}'_u, \end{aligned}$$

which implies $x_{u-1} \in \mathcal{M}'_u \cap \ker A \cap \mathcal{T}^{u-j+1} = \mathcal{M}'_u \cap \mathcal{M}_{u-j+1} = 0$. Thus $x_{u-1} = 0$ and therefore $x_{u-2} = x_{u-3} = \cdots = x_j = x_{j-1} = 0$.

In the following we prove that the subspaces $\operatorname{Im} X_{u,j}$, $u \in \underline{q}$, $j \in \{0, 1, \cdots, u-1\}$ are independent, i.e., that

$$\sum_{u=1}^{q} \sum_{j=0}^{u-1} X_{u,j}\alpha_{u,j} = 0$$

implies $\alpha_{u,j} = 0$, where $\alpha_{u,j}$ is a vector of dimension $d_u \times 1$ over the field of the reals.

Note that it suffices to prove that the subspaces $A \operatorname{Im} X_{u,j}$, $u \in \{2, 3, \cdots, q\}$, $j \in \{0, 1, \cdots, u-2\}$ are independent because $AX_{u,u-1} = 0$ and the subspaces $\operatorname{Im} X_{u,u-1} = \mathcal{M}'_u$, $u \in \underline{q}$, are independent.

Thus consider

$$\sum_{u=2}^{q} \sum_{j=0}^{u-2} AX_{u,j}\alpha_{u,j} = 0.$$

By using (A.5) we then obtain a sequence of vectors $(x_0, x_1, \cdots, x_{q-2})$ such that

$$\text{(A.7)} \qquad Ex_0 = 0, \quad Ex_1 = Ax_0, \quad \cdots, \quad Ex_{q-2} = Ax_{q-3}, \quad Ax_{q-2} = 0$$

with

$$\text{(A.8)} \quad \begin{aligned} x_{q-2} &= X_{q,q-1}\alpha_{q,0}, \\ x_{q-3} &= (X_{q-1,q-2}\alpha_{q-1,0}) + (X_{q,q-2}\alpha_{q,0} + X_{q,q-1}\alpha_{q,1}), \\ x_{q-4} &= (X_{q-2,q-3}\alpha_{q-2,0}) + (X_{q-1,q-3}\alpha_{q-1,0} + X_{q-1,q-2}\alpha_{q-1,1}) \\ &\quad + (X_{q,q-3}\alpha_{q,0} + X_{q,q-2}\alpha_{q,1} + X_{q,q-1}\alpha_{q,2}) \end{aligned}$$

$$\text{(A.9)} \quad \vdots$$

$$\begin{aligned} x_0 &= (X_{2,1}\alpha_{2,0}) + (X_{3,1}\alpha_{3,0} + X_{3,2}\alpha_{3,1}) + \cdots \\ &\quad + (X_{j,1}\alpha_{j,0} + X_{j,2}\alpha_{j,1} + \cdots + X_{j,j-1}\alpha_{j,j-2}) + \cdots \\ &\quad + (X_{q,1}\alpha_{q,0} + X_{q,2}\alpha_{q,1} + \cdots + X_{q,q-1}\alpha_{q,q-2}). \end{aligned}$$

Now (A.7) implies $x_{q-2} \in \ker A \cap \mathcal{T}^{q-1} = \mathcal{M}_{q-1}$ and from (A.8) we obtain $x_{q-2} \in \mathcal{M}_{q-1} \cap \mathcal{M}'_q = 0$. Since $X_{q,q-1}$ is monic we then have

$$\text{(A.10)} \qquad \alpha_{q,0} = 0.$$

Again from (A.7) we obtain $Ax_{q-3} = 0$, which implies

$$x_{q-3} \in \ker A \cap \mathcal{T}^{q-2} = \mathcal{M}_{q-2}.$$

By using (A.9), (A.10) and (A.4) we then obtain

$$x_{q-3} \in \mathcal{M}_{q-2} \cap (\mathcal{M}'_q \oplus \mathcal{M}'_{q-1}) = 0.$$

Since $\mathcal{M}'_q$ and $\mathcal{M}'_{q-1}$ are independent subspaces, the result is $\alpha_{q-1,0} = \alpha_{q,1} = 0$.

Proceeding this way, we obtain at the last step

$$x_0 \in \mathcal{M}_1 \cap (\mathcal{M}'_2 \oplus \mathcal{M}'_3 \oplus \cdots \oplus \mathcal{M}'_q) = 0$$

and consequently $\alpha_{u,j} = 0$, $u \in \{2, \cdots, q\}$, $j \in \{0, \cdots, u-2\}$, which implies that the subspaces $\operatorname{Im} X_{u,j}$, $u \in \underline{q}$, $j \in \{0, \cdots, u-1\}$ are independent.

Since the subspaces $\operatorname{Im} X_{u,0}$, $u \in \underline{q}$, are independent with $\dim \bigoplus_{u=1}^{q} \operatorname{Im} X_{u,0} = \sum_{u=1}^{q} d_u$ and since $\dim (\ker E \cap \mathcal{T}^*) = \dim (\ker A \cap \mathcal{T}^*)$, it follows that

$$\mathcal{T}^1 = \ker E \cap \mathcal{T}^* = \bigoplus_{u=1}^{q} \operatorname{Im} X_{u,0}.$$

From (A.2) we also obtain that for $c \in \underline{q}$

$$\mathcal{T}^c = \bigoplus_{u=1}^{q} \operatorname{Im} X_{u,0} \bigoplus_{u=2}^{q} \operatorname{Im} X_{u,1} \bigoplus \cdots \bigoplus_{u=c}^{q} \operatorname{Im} X_{u,c-1}$$

and then

$$\mathcal{T}^q = \mathcal{T}^* = \sum_{u=1}^{q} \sum_{j=0}^{u-1} \operatorname{Im} X_{u,j};$$

which concludes the proof. □

Continuation of Theorem 1. Let $\gamma_i := \sum_{u=i}^{q} d_u$ and consider the following set of vectors $\{x_{i,j}\}$, $j \in \{0, 1, \cdots, q-1\}$:

$$\begin{aligned}
\{x_{i,0}\} &:= [X_{1,0}; X_{2,0}; \cdots; X_{q,0}], \qquad i \in \underline{\gamma}_1, \\
\{x_{i,1}\} &:= [X_{2,1}; X_{3,1}; \cdots; X_{q,1}], \qquad i \in \underline{\gamma}_2, \\
&\vdots \\
\{x_{i,q-1}\} &:= [X_{q,q-1}], \qquad i \in \underline{\gamma}_q,
\end{aligned}$$

i.e., $\{x_{i,j}\}$, $i \in \underline{\gamma}_{j+1}$, is a set formed from the union of the columns of the matrices $[X_{j+1,j}; \cdots; X_{q,j}]$.

Analogously to the definition of the controllability indices of a pair (A, B) in [18], let

$$r_i := \text{number of integers in the set } \{\rho_1, \rho_2, \cdots, \rho_q\} \text{ which are } \geqq i.$$

Then

$$r_1 \geqq r_2 \geqq \cdots \geqq r_l \geqq 1.$$

Write $k_i := r_i - 1$, $i \in \underline{l}$, and note that $k_1 = q - 1$. Since the vectors $\{x_{i,j}\}$ above defined, constitute a basis for $\mathcal{T}^*$, it then follows that

$$\mathcal{T}^* = \bigoplus_{i=1}^{l} \mathcal{T}_i$$

with $\mathcal{T}_i = \operatorname{span}\{x_{i,j}\}$, $\dim \mathcal{T}_i = k_i + 1$, such that

$$Ex_{i,0} = 0, \quad Ex_{i,j} = Ax_{i,j-1}, \quad Ax_{i,k_i} = 0, \qquad j \in \underline{k}_i.$$

Now suppose that $\mathcal{T}^* = \bigoplus_{i=1}^{l} \mathcal{T}_i$, where $\mathcal{T}_i = \operatorname{span}\{x_{i,j}\}$, $\dim \mathcal{T}_i = m_i + 1$ and such that

$$Ex_{i,0} = 0, \quad Ex_{i,j} = Ax_{i,j-1}, \quad Ax_{i,m_i} = 0, \qquad j \in \underline{m}_i.$$

Fix an index $j \in \{0, 1, \cdots, n\}$ and let S_j be a subset of $\underline{l}$, such that $x_{i,j}$, $i \in S_j$ is a vector of the above basis. Then we must have $\dim \operatorname{span}\{x_{i,j}\} = \rho_{j+1}$, $i \in S_j$, which immediately implies that $m_i = k_i$, $i \in \underline{l}$. □

Proof of Lemma 2. (a) The statement of this item follows from Corollary 1 and Theorem 2. Also note that $\mathcal{V}^* = 0$ implies $\ker A = 0$.

(b) From (4) we have that $\mathcal{W}_b^* = E^{-1}(A\mathcal{W}_b^*)$ and thus $E\mathcal{W}_b^* \subset A\mathcal{W}_b^*$.

Suppose that N has an eigenvalue $\lambda \neq 0$. Then there exists a vector $0 \neq w \in \mathcal{W}_b^*$ such that $Ew = \lambda Aw$, i.e., $Aw = \lambda^{-1} Ew$ which implies $w \in \mathcal{V}^* = 0$.

Therefore all eigenvalues of N are equal to zero and N admits the following Jordan decomposition

$$\operatorname{Mat} N = \operatorname{diag}(N_1, \cdots, N_p, 0)$$

where N_i, $i \in \underline{p}$, is a matrix as in (14) with dim $N_i = g_i + 1$, $g_i \geqq 1$. It now follows that the pencil $(s\bar{E} - A)$ has p infinite-zeros of order g_i, $i \in \underline{p}$.

In order to show the statement (15) consider the family F_4 of all subspaces $\mathcal{W}$ described by (12), i.e.

$$F_4 = \{\mathcal{W} \subset \mathcal{X} | E\mathcal{W} \subset A\mathcal{W}\}$$

and such that the map $N := A\mathcal{W}|E|\mathcal{W}$ has all eigenvalues equal to zero.

We show next that $\mathcal{W}_b^*$ is the supremum of the above family. This can be easily seen by using induction in (13), i.e. $w_{i,1} \in \mathcal{W}_b^1$ and if $w_{i,j-1} \in \mathcal{W}_b^{j-1}$ then $w_{i,j} \in E^{-1}(A\mathcal{W}_b^{j-1}) = \mathcal{W}_b^j \subset \mathcal{W}_b^*$.

Next suppose that $\mathcal{W}_b^*$ is a proper subspace of $\tilde{\mathcal{W}}_b := \sup\{\mathcal{W} | E\mathcal{W} \subset A\mathcal{W}\}$. Then the map $A\tilde{\mathcal{W}}_b|E|\tilde{\mathcal{W}}_b$ has at least one eigenvalue $\lambda \neq 0$ which implies that there exists $0 \neq w \in \tilde{\mathcal{W}}$ but $w \notin \mathcal{W}_b^*$ such that $Ew = \lambda Aw$ which again implies $w \in \mathcal{V}^* = 0$, which proves (15).

(c) Note that for $w \in \mathcal{W}_b^*$ we have $Ew = Aw_1$, $w_1 \in \mathcal{W}_b^*$, and thus the map $\hat{E}$ is well defined.

We now show that the maps $\hat{E}$ and $\hat{A}$ are monic, i.e. $\ker \hat{E} = \ker \hat{A} = 0$. Suppose there exists $\hat{x} \in \mathcal{X} (\text{mod } \mathcal{W}_b^*) \in \ker \hat{E}$. Then

$$\hat{E}\hat{x} = \hat{E}Px = QEx = 0,$$

whence

$$Ex \in A\mathcal{W}_b^* \Rightarrow x \in E^{-1}(A\mathcal{W}_b^*) = \mathcal{W}_b^*$$

and then $\hat{x} = 0$. Analogously, suppose that $\hat{x} \in \ker \hat{A}$. Then

$$\hat{A}\hat{x} = \hat{A}Px = QAx = 0 \Rightarrow Ax \in A\mathcal{W}_b^*$$

and since A is monic, it follows that $x \in \mathcal{W}_b^*$ and hence $\hat{x} = 0$.

Since $\hat{E}$ and $\hat{A}$ are monic it follows that the pencil $(s\hat{E} - \hat{A})$ has no infinite-zeros (see (13)) and $\ker_{R[s]}(s\hat{E} - \hat{A}) = 0$, because the subspace $\mathcal{T}^*$ relative to the maps $\hat{E}$ and $\hat{A}$ is zero.

In the sequel we show (as expected) that the pencil $(s\hat{E} - \hat{A})$ has no finite-zeros. Note that since A is monic, then $\lambda = 0$ is not a finite-zero of $(sE - A)$. Now suppose there exist $0 \neq \lambda \in C$ and $\hat{v}$ such that $\hat{A}\hat{v} = \lambda \hat{E}\hat{v}$. Then

$$\hat{E}\hat{v} = \hat{E}Pv = QEv = \lambda^{-1}QAv. \tag{A.11}$$

Let $\nu := \text{span}\{v\}$. Then from (A.11) it follows that $E\nu \subset A\nu + A\mathcal{W}_b^*$ and hence

$$E(\nu + \mathcal{W}_b^*) \subset A(\nu + \mathcal{W}_b^*).$$

From (15) we then have $\nu + \mathcal{W}_b^* \subset \mathcal{W}_b^*$ whence $v \in \mathcal{W}_b^*$ and therefore $\hat{v} = 0$.

Since the pencil $(s\hat{E} - \hat{A})$ has no finite and infinite-zeros and $\ker_{R[s]}(s\hat{E} - \hat{A}) = 0$, it then follows that its rows must be linearly dependent over $R[s]$, i.e., there exists a vector $x^T(s)$ with components in $R[s]$ such that $x^T(s)(s\hat{E} - \hat{A}) = 0$. □

Remark. A polynomial basis for $\ker(sE^T - A^T)$ can be computed as indicated in the proof of Theorem 1.

Proof of Lemma 3. For $u = 1$, $\bar{\mathcal{W}}_b^1 = \ker \bar{E}$ and $\mathcal{W}_b^1 = \ker E$. We have to show that $\ker \bar{E} = P \ker E$.

Let $\bar{x} \in \mathcal{X} (\text{mod } \mathcal{V}^*) \in \ker \bar{E}$. Then by using (16) we obtain $0 = \bar{E}\bar{x} = \bar{E}Px = QEx$, which implies $Ex \in E\mathcal{V}^*$. Hence, $x \in \mathcal{V}^* + \ker E \Rightarrow \bar{x} = Px \in P \ker E$, i.e., $\ker \bar{E} \subset P \ker E$. Now, let $x \in \ker E$. Then $0 = QEx = \bar{E}Px$, so that $P \ker E \subset \ker \bar{E}$.

Assume that $\bar{W}_b^{u-1} = P\mathcal{W}_b^{u-1}$ and let $\mathcal{S} := P^{-1}\bar{\mathcal{W}}_b^u$. Hence from (18) we obtain $\bar{E}P\mathcal{S} \subset \bar{A}P\mathcal{W}_b^{u-1}$, or $QE\mathcal{S} \subset QA\mathcal{W}_b^{u-1}$, which implies, $E\mathcal{S} \subset A\mathcal{W}_b^{u-1} + E\mathcal{V}^*$. Hence

$$\mathcal{S} \subset E^{-1}(A\mathcal{W}_b^{u-1}) + \mathcal{V}^*$$

and then $P\mathcal{S} = \bar{\mathcal{W}}_b^u \subset P\mathcal{W}_b^u$.

Now consider the subspace $\mathcal{W}_b^u$ in (4). Then

$$QE\mathcal{W}_b^u \subset QA\mathcal{W}_b^{u-1} \Rightarrow \bar{E}P\mathcal{W}_b^u \subset \bar{A}P\mathcal{W}_b^{u-1}$$

which implies

$$P\mathcal{W}_b^u \subset \bar{E}^{-1}(\bar{A}\bar{\mathcal{W}}_b^{u-1}) = \bar{\mathcal{W}}_b^u. \qquad \square$$

REFERENCES

[1] V. A. ARMENTANO, *Almost invariant subspaces and generalized linear systems*, Ph.D. thesis, Electrical Engineering Dept., Imperial College, Univ. London, London, 1983.

[2] ———, *Eigenvalue placement for generalized linear systems*, System Control Lett., 4 (1984), pp. 199-202.

[3] P. BERNHARD, *On singular implicit dynamical systems*, this Journal, 20 (1982), pp. 612-633.

[4] S. L. CAMPBELL, *Singular Systems of Differential Equations*, Pitman, London, 1980.

[5] J. D. COBB, *Feedback and pole placement in descriptor variable systems*, Internat. J. Control, 6 (1981), pp. 1135-1146.

[6] ———, *Descriptor variable and generalized singularly perturbed systems: a geometric approach*, Ph.D. dissertation, Dept. Electrical Engineering, Univ. Illinois, Urbana, IL, 1980.

[7] F. R. GANTMACHER, *The Theory of Matrices*, Vol. II, Chelsea, New York, 1959.

[8] N. KARCANIAS AND G. E. HAYTON, *Generalized autonomous dynamical systems, algebraic duality and geometric theory*, 8th World IFAC Congress, Kyoto, Japan, 1981.

[9] F. L. LEWIS, *Descriptor systems: decomposition into forward and backward subsystems*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 167-170.

[10] H. H. ROSENBROCK, *Structural properties of linear dynamical systems*, Internat. J. Control, 2 (1974), pp. 191-202.

[11] P. VAN DOOREN, *The computation of Kronecher's canonical form of a singular pencil*, Linear Algebra Appl., 27 (1979), pp. 103-141.

[12] A. I. G. VARDULAKIS AND N. KARCANIAS, *Relations between strict equivalence invariants and structure at infinity of matrix pencils*, IEEE Trans. Automat. Control, AC-28 (1983), pp. 514-516.

[13] G. C. VERGHESE, B. C. LÉVY AND T. KAILATH, *A generalized state-space for singular systems*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 811-831.

[14] G. C. VERGHESE, *Infinite frequency behaviour in generalized dynamical systems*, Ph.D. dissertation, Dept. Electrical Engineering, Stanford Univ., Stanford, CA, 1978.

[15] ———, *Further notes on singular descriptions*, Joint Automatic Control Conference, TA4, Charlottesville, VA, 1981.

[16] J. C. WILLEMS, *Almost invariant subspaces: an approach to high gain feedback design—Part* I: *Almost controlled invariant subspaces*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 235-252.

[17] K-T. WONG, *The eigenvalue problem* $\lambda Tx + Sx$, J. Differential Equations, 16 (1974), pp. 270-281.

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

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 4, July 1986

004

INTRACTABLE PROBLEMS IN CONTROL THEORY*

CHRISTOS H. PAPADIMITRIOU† AND JOHN TSITSIKLIS†

Abstract. This paper is an attempt to understand the apparent intractability of problems in decentralized decision-making, using the concepts and methods of computational complexity. We first establish that the discrete version of an important paradigm for this area, proposed by Witsenhausen, is NP-complete, thus explaining the failures reported in the literature to attack it computationally. In the rest of the paper we show that the computational intractability of the discrete version of a control problem (the team decision problem in our particular example) can imply that there is no satisfactory (continuous) algorithm for the continuous version. To this end, we develop a theory of continuous algorithms and their complexity, and a quite general proof technique, which can prove interesting by themselves.

Key words. complexity, team theory, decentralized control

1. Introduction. Most classical problems arising in the fields of optimization and control are, in a very real sense, "easy to solve". By this we mean that there are computational procedures with satisfactory performance, which can be used to compute the solution of such problems. Naturally, a lot of effort is being devoted to finding more and more efficient algorithms which exploit any special structure present, but usually there is *no fundamental intractability* to be overcome. For example, in a nonlinear optimal control problem (under some smoothness assumptions) a solution can always be obtained by discretizing the problem with a dense enough grid and then using the discrete dynamic programming algorithm. Roughly speaking, the accuracy ε of the solution so obtained is inversely proportional to the number of points in the grid and such algorithms require time which is a polynomial function of $1/\varepsilon$. The situation is similar in many other classical problems such as nonlinear optimization or numerical integration of partial differential equations. In fact, in some extremely favorable cases (when, for example, the problem can be reduced to the evaluation of some analytic function), the computation time is polynomial in the *logarithm* of $1/\varepsilon$, or, even better, the solution can be expressed *in closed form.*

On the other hand, certain problems that arise in the field of decentralized decision making and control have defied all attempts for the development of realistic algorithms or representations of their solution. (It has been customary to refer to such problems as *nonclassical* control problems.) Witsenhausen's counterexample in decentralized control [Wi] is a paradigm. This problem can be viewed as a simple two-stage stochastic optimal control problem without perfect recall of the measurements. In contrast to related control problems with perfect recall, for which optimal decision rules are linear and easy to compute, the optimal decision rules for Witsenhausen's problem are provably nonlinear, and it is nontrivial to even show that they exist [Wi]. Despite persistent efforts, a representation of the optimal solution to this problem or an efficient algorithm to compute its solution has never been found. Ho and Chang [HC] took a closer look at the *discrete version* of this problem. They considered the "most reasonable" approaches to the construction of an efficient algorithm, and provided a discussion explaining why such approaches fail. However, this could not rule out the possibility that some other approach might lead to an efficient algorithm, or, more importantly, that an efficient solution for the continuous problem is possible.

* Received by the editors September 7, 1984, and in revised form April 12, 1985. This research was supported in part by the National Science Foundation, and by an IBM Faculty Development Award.

† Department of Computer Science, Stanford University, Stanford, California 94305.

This increase in difficulty in going from the centralized to the distributed problem is usually attributed to a loss of convexity; however, no formal explanation of this phenomenon had been attempted. On the other hand, some recent work has related the complexity of decentralized control, somewhat loosely, with the Theory of Computational Complexity [GJ], [PS]. These results indicate that the discrete versions of some seemingly simple problems in decentralized decision making (unfortunately, so far excluding Witsenhausen's counterexample) are computationally intractable (NP-complete or worse) [PT], [TA], [GJW], [Pa], [Ts], thus providing objective measures for the difficulty of the *discrete* problems. Nevertheless, the above research left open the issue of the intractability of the (more interesting) *original continuous problems.* In general, it is not automatically true that if a discrete version of a problem is hard, then the continuous problem is also hard. A classical example here could be linear programming, which can be solved in polynomial time, despite the fact that its discrete version—integer programming—is much harder.

In this paper we address and in many ways settle the issues raised above. In § 2 we discuss the few available results on the complexity of discrete nonclassical control problems. More importantly, we prove that the discrete version of Witsenhausen's counterexample is NP-complete, thus explaining the lack of progress on it, and the failures reported in [HC]. The goal of the remaining sections is to relate the complexity of discrete and continuous problems. In particular, we show that complexity results for a discrete problem can be used to prove the nonexistence of realistic (i.e., polynomial in the desired accuracy) algorithms for classes of continuous problems. We chose to proceed in terms of a specific example, the static team decision problem [MR], [Ra]; however, our proofs define a methodology by which similar results can be proved for other problems as well. In § 3 we make precise the notion of an algorithm that solves a continuous problem. We observe that there are several possible such notions. We also describe the main construction used in the rest of the paper, whereby from any instance of the discrete version of a decision problem we construct an instance of its continuous counterpart, which is provably closely related to the discrete one. In § 4 we show our main results, linking the difficulty of nonclassical control problems (the team problem in particular) to the theory of computational complexity. For three different notions of "efficiently solvable continuous problem" we present evidence that the team problem is not. These negative results depend on $P \neq NP$ and some related conjectures from Complexity Theory. Finally, in § 5 we discuss our results; we also place them into perspective by contrasting them to other theories of complexity for continuous problems [TW], [TWW], [YN], [Ko].

2. The complexity of discrete nonclassical problems. In this section we consider the computational complexity of the discrete versions of some representative nonclassical control problems: the static team decision problem [MT], [Ra], the discrete version of Witsenhausen's counterexample in stochastic control [Wi], as well as some nonclassical control problems in Markov chains. The main new result is that the discrete version of Witsenhausen's problem is NP-complete. For convenience, we restrict to problems involving two agents only; problems with more agents are bound to be at least as hard.

The discrete static team decision problem. We define below the discrete version of the team decision problem of Marschak and Radner, called DTEAM. The problem is the following: Each one of two agents observes a separate integer random variable k_i, $i = 1, 2$, $1 \leqq k_i \leqq N$ and makes a decision $u_i = \gamma_i(k_i)$, $u_i \in \{1, \cdots, M\}$ based on his own information only. Then a cost $c(k_1, k_2, \gamma_1(k_1), \gamma_2(k_2))$ is incurred. The problem consists

of finding decision rules that minimize the expected cost. For simplicity, we take all pairs (k_1, k_2) in the given range to be equiprobable.

An instance $I = (N, M, c, K)$ of DTEAM consists of positive integers N, M (the cardinalities of the observation and decision sets), a nonnegative integer K, and an integer valued cost function $c: \{1, \cdots, N\}^2 \times \{1, \cdots, M\}^2 \to Z$. For any pair γ_1, γ_2 of functions $\gamma_i: \{1, \cdots, N\} \to \{1, \cdots, M\}$, define their cost to be

$$J(\gamma_1, \gamma_2) = \sum_{k_1=1}^{N} \sum_{k_2=1}^{N} c(k_1, k_2, \gamma_1(k_1), \gamma_2(k_2)).$$

The optimal cost is defined to be

$$J^*(I) = \min_{\gamma_1, \gamma_2} J(\gamma_1, \gamma_2).$$

By "solving" this instance, we mean deciding whether $J^*(I) \leqq K$, or not.

We let SDTEAM (for *Simple* DTEAM) be the special case of DTEAM restricted to instances for which $K = 0$, $M = 4$ and the range of c is $\{0, 1\}$.

Complexity theory. At this point is seems appropriate to introduce some basic notions from complexity theory. See [GJ], [HU], [PS] for more complete and formal treatments.

Most of the discrete problems that we deal with in this paper will be of the *language recognition* kind, that is, problems of deciding whether a given string (encoding some combinatorial object) belongs to a fixed set of strings or not. In DTEAM, for example, the string encodes the integers M, N, and K, and the table of the cost function. The question is whether this string is in the set of strings (language) that encode instances of DTEAM in which the optimum cost is below K.

Our precise choice of a model of computation is not very critical. We could choose any variant of the Turing machine, or the random access machine models which appear to be much closer to actual computers [AHU]. All such choices are essentially equivalent (modulo a polynomial), as long as they are basically *realistic*. This latter clause excludes models which, for example, assume real arithmetic with infinite precision at unit cost per operation. Any model, whose units of computation can be achieved within a constant amount of time with constant hardware, is "realistic" in the above sense.

In the interest of differentiating between "easy" and "hard" problems, let us define P to be the class of all such problems that can be solved by an algorithm in a number of steps which is a polynomial in the length of the input string. Some well-known "hard" problems, including the satisfiability problem for Boolean formulas and the traveling salesman problem (with a limit on the cost of the tour, as in the definition of DTEAM), are not known, neither believed, to be in P; they belong, however, in another class, called NP (for nondeterministic polynomial). A problem is in NP if, whenever a string encodes a "yes" instance, there is a polynomially short and polynomially easy to check "certificate" that testifies to this. A "no" instance has no such certificate. For example, in the traveling salesman problem, the certificate is the shortest tour, of cost less than the set limit; in DTEAM the optimum decision rule that achieves cost K or less; and so on. Another, equivalent way to define NP is in terms of problems that can be solved in polynomial time by *nondeterministic* Turing machines (hence the name NP).

Is P = NP? This turns out to be the central open question in Complexity Theory today. It is widely believed that P $\neq$ NP, that is, that P is a proper subset of NP, but no proof exists (or is in sight). However, even in the absence of a definite answer to this question, for certain problems in NP we have quite convincing evidence that they

are indeed intractable. What has been shown is that these problems are NP-*complete.* This means that all problems in NP reduce in polynomial time to these. Hence, NP-complete problems are "the hardest problems in NP", in the sense that, if P is not NP, then the NP-complete problems will be the first to be intractable, of nonpolynomial complexity. A great variety of some of the hardest and most stubborn computational problems from combinatorics, optimization, logic, number theory and graph theory have been shown to be NP-complete (including the traveling salesman problem and the satisfiability of Boolean formulas; see [GJ] for a complete census, circa 1979). The usual way that a new problem is shown NP-complete is to reduce a known NP-complete problem to it. We shall see a rather involved example shortly.

Problem SDTEAM is known to be NP-complete [PT]. In fact, it follows easily from our proof that SDTEAM remains NP-complete even if the instances are restricted so that we know that the optimum cost is either zero or one, and we must decide which of the two. (This is done by taking any instance with $M = 3$—a case which is already NP-complete [PT]—and adding to each pair of observations a choice which can guarantee an overall cost of one). We shall use this fact in our proofs.

In our analysis of the complexity of nonclassical control problems, we shall also refer to complexity classes *above* P and NP. In analogy to polynomial-time computation, one can study the *exponential-time* analog, that is, problems solvable within a number of steps that grows as 2^{cn}, for some constant c. We let EXP and NEXP denote the corresponding deterministic and nondeterministic complexity classes. Also, we let DEXP and NDEXP be the analogous classes for *doubly exponential* complexities, that is, growths of the form 2^{c2^n}. These complexity classes are not, of course, nearly as practically important as P and NP, but they too are unresolved puzzles: It is not known whether EXP = NEXP or DEXP = NDEXP (although we expect that inequality holds). What *is* known, however, is that P = NP implies EXP = NEXP, which in turn implies DEXP = NDEXP (see [HU] for the standard arguments needed to show this).

Witsenhausen's counterexample revisited. Witsenhausen's counterexample is the following problem [Wi]:

$$\text{minimize } E[K(\gamma(x))^2+(\delta(x+\gamma(x)+v)+x+\gamma(x))^2],$$

with respect to all measurable real valued functions γ, δ of a single variable, where x, v are independent, normal, zero mean random variables (with given variance) and K a nonnegative constant. (Notice that this is not a (discrete) computational problem of the kind we introduced in the previous subsections. For more formal treatment of continuous computational problems, see the next section.) As was pointed out in the introduction, a representation of an optimal solution to this problem or an efficient algorithm has never been found. Of course, an algorithm can always be constructed as follows: discretize the densities of the random variables x, v and constrain the decision rules γ, δ to have finite range; then solve the discretized problem by exhaustive enumeration. However, this is unsatisfactory because the number of decision rules that have to be enumerated is exponential in the cardinality of the allowed range of the decision rules. It is this discrete problem that was studied by Ho and Chang [HC] with very little success. We explain this persistent record of failures by proving below that the discretized version of Witsenhausen's problem, as defined by Ho and Chang, is NP-complete.

Let us now define formally the discrete problem of interest:

Problem WITSENHAUSEN: Given probability mass functions f, $g: Z \to Q$ for integer variables x, v and integer constants K, B are there functions γ, $\delta: Z \to Z$ such

that

$$J(\gamma, \delta) = E[\gamma^2(x) + K(x + \gamma(x) + \delta(x + \gamma(x) + v))^2] \leqq B?$$

THEOREM 1. WITSENHAUSEN *is* NP-*complete.*

Proof. We first introduce a variation of the problem of three-dimensional matching (3DM) [GJ]:

3DM: Given a set S and a family F of subsets of S—of cardinality three—can we subdivide F into three subfamilies C_0, C_1, C_2 such that a) subsets in each family are disjoint; b) the union of the subsets in C_0 equals S?

LEMMA 1. 3DM *is* NP-*complete.*

Sketch. We basically use the construction in the standard proof that the (less restricted) version of 3DM, in which the sets in C_1, C_2 are not required to be disjoint, is NP-complete [GJ], [PS]. In that proof we construct, for each Boolean formula with three literals per clause, an instance of 3DM, such that there is a subfamily C_0 as described in 3DM iff the formula was satisfiable. It is not hard to observe, however, that, once a subfamily C_0 exists, the remaining sets of the instance can be subdivided into two subfamilies of disjoint sets. □

To prove Theorem 1, we reduce 3DM to WITSENHAUSEN. Suppose that we are given an instance S, F of 3DM, where $S = \{1, \cdots, m\}$, $F = \{S_1, \cdots, S_n\}$. Without loss of generality, assume that $n \leqq m$. We now construct an instance of WITSENHAUSEN. There will be $3n$ values of the random variable x with nonzero probability and $M = 1 + 4n\nu + 3n$ such values for v, where $\nu = \lceil \sqrt{3n - m + 1} \rceil$. All these values will be taken equiprobable. To complete the construction, we need to specify the sets $X = \{x_1, \cdots, x_{3n}\}$, $V = \{v_1, \cdots, v_M\}$ of values with nonzero probability. Concerning the constants B, K, we let $B = (3n - m)/3nM$, $K = 3nM(B + 1)$. To define the actual integers with nonzero probabilities, we need a lemma:

LEMMA 2. *There are n distinct integers* $0 \leqq z_1 \leqq \cdots \leqq z_n \leqq 3n^4$ *such that*

(a) *All the differences* $z_p - z_q$ *are distinct.*

(b) *Any difference* $(z_{i+1} - z_i) - (z_{j+1} - z_j)$ *is distinct from any difference in* (a).

Proof. We define z_k, $1 \leqq k \leqq n$, recursively. Let $z_1 = 0$ and assume that $z_1, \cdots, z_k$, $k < n$, have been constructed and $z_k \leqq 3k^4$. In order to pick a value for z_{k+1}, notice that it has to obey only the following constraints: (i) $z_{k+1} > z_k$, (ii) $z_{k+1} - z_p \neq z_i - z_j$, $(1 \leqq j, p, q \leqq k)$, (iii) $(z_{k+1} - z_k) - (z_{j+1} - z_j) \neq z_p - z_q$, $(1 \leqq j, p, q \leqq k + 1)$. (Some of the constraints in (iii) hold automatically.) So, z_{k+1} has to avoid at most $3k^4 + k^3 + (k+1)^3 < 3k^4 + 9k^3 < 3(k+1)^4$ values. Therefore, there exists an integer less than or equal to $3(k+1)^4$ whose value can be assigned to z_{k+1}. □

Notice that, given n, the integers $z_1, \cdots, z_n$ can be constructed recursively in polynomial time by means of the procedure suggested by the proof of Lemma 2. Let us assume that such a sequence $z_1, \cdots, z_{3n}$ has been constructed. Moreover, let us multiply each element of the sequence by 4, so that the expressions which are distinct by Lemma 2 are different by at least 4.

We now complete the construction of the sets X, V. The set X contains $3n$ elements; each element $x \in X$ is associated to a set $S_i \in F$ and an element $j_{ik} \in S_i$, where j_{ik} $(i = 1, \cdots, n; 1, 2, 3)$ denotes the kth element of S_i. We then let

$$x_{3(i-1)+k} = 3mz_{3(i-1)+k} + 3j_{ik}. \tag{2.1}$$

The set V contains the element 0; also for any consecutive elements x_i, x_{i+1} of X *corresponding to the same set* (that is, $i = 1, 2 \pmod 3$), V contains the numbers $x_{i+1} - x_i + \rho$, $\rho \in U = \{-\nu, -\nu + 1, \cdots, -2, -1, 1, 2, \cdots, \nu - 1, \nu\}$. Finally, V contains the numbers $3m(A + z_i)$, $i = 1, 2, \cdots, 3n$, where $A = 3z_{3n}$; this completes the construction.

Let us put together a few facts, for future reference:

LEMMA 3. (a) *If for some* γ, δ, *we have* $J(\gamma, \delta) \leqq B$, *then* $|\gamma(x)| \leqq \nu$, $\forall x \in X$.

(b) *The expressions* $|x_i - x_j|$, $|x_i - x_j + x_p - x_q|$, $|x_{i+1} - x_i - x_{j+1} + x_j + x_p - x_q|$, $(1 \leqq i - 1, j - 1, p, q \leqq n)$ *are either zero or no smaller than* $3m$. *For large enough* m, $3m > 4\nu + 2$.

(c) $|x_i - x_j| \leqq 3mA - 2 - \nu$, *for large enough* m.

Proof. (a) If for some $x \in X$ we have $|\gamma(x)| > \nu$, then

$$J(\gamma, \delta) > \frac{\nu^2}{3nM} \geqq \frac{3n - m}{3nM} = B.$$

(b) We use (2.1), the inequality $j_{ik} \leqq m$ and the fact that the magnitudes of the corresponding expressions involving the z's instead of the x's are either zero or at least four; we obtain in the second case, for example, $|x_{i+1} - x_i - x_{j+1} + x_j + x_p - x_q| \geqq 12m - 9m = 3m$. Finally notice that ν increases only as the square root of m, which also proves part (c). □

The following lemma completes the proof of the theorem.

LEMMA 4. (S, F) *is a "yes" instance of* 3DM *if and only if there exist* γ, δ *such that* $J(\gamma, \delta) \leqq B$ *for the above constructed instance of* WITSENHAUSEN.

Proof. If. Suppose that there exist γ, δ such that $J(\gamma, \delta) \leqq B$. Then, in particular, $K(x + \gamma(x) + \delta(x + \gamma(x) + v))^2 / 3Mn \leqq B$, $\forall x \in X$, $\forall v \in V$. Therefore, $|x + \gamma(x) + \delta(x + \gamma(x) + v)|^2 \leqq B/(B+1) < 1$, which implies that $x + \gamma(x) + \delta(x + \gamma(x) + v) = 0$, $\forall x \in X$, $\forall v \in V$. Let $x_i \neq x_j$. Then, using Lemma 3(a, b), we have

$$\begin{aligned} |\delta(x_i + \gamma(x_i) + v) - \delta(x_j + \gamma(x_j) + v')| &= |x_i + \gamma(x_i) - x_j - \gamma(x_j)| \\ &\geqq |x_i - x_j| - |\gamma(x_i) - \gamma(x_j)| \\ &\geqq 3m - 2\nu > 0, \end{aligned}$$

which shows that

$$x_i + \gamma(x_i) + v \neq x_j + \gamma(x_j) + v' \quad \forall v, v' \in V, \quad \forall x_i, x_j \in X, \quad x_i \neq x_j. \tag{2.2}$$

Let x_i, x_{i+1} be two consecutive elements of X corresponding to the same set S_j. Inequality (2.2) must hold for $v' = 0$ and $v = x_{i+1} - x_i + \rho$, $\forall \rho \in U$. Thus, $\gamma(x_i) + \rho \neq \gamma(x_{i+1})$, $\forall \rho \in U$. Consequently, either $\gamma(x_i) = \gamma(x_{i+1})$, or $|\gamma(x_i) - \gamma(x_{i+1})| > \nu$, which would contradict Lemma 3(a). Therefore γ takes the same value on those elements of X corresponding to the same set S_j. We denote this value by $\gamma(S_j)$.

Inequality (2.2) must also hold when x_i, x_j correspond to the same element $k \in S$ belonging to different subsets S_p, S_q; that is, $x_i = 3mz_i + 3k$, $x_j = 3mz_j + 3k$. Let $v = 3m(A + z_j)$, $v' = 3m(A + z_i)$. Inequality (2.2) becomes $\gamma(x_i) \neq \gamma(x_j)$, which implies $\gamma(S_p) \neq \gamma(S_q)$, whenever $S_p \neq S_q$, $S_p \cap S_q \neq \emptyset$.

Notice that (by our choice of B), $\gamma(x)$ can be nonzero for at most $3n - m$ elements of X. Moreover, at most one $\gamma(x)$ per element of $S = \{1, \cdots, m\}$ can be zero; thus, $\gamma(x)$ must be nonzero for exactly $3n - m$ elements; for those elements, $|\gamma(x)| = 1$. Let C_0 (respectively, C_1, C_2) be the family of subsets of F for which $\gamma(S_p) = 0$ (respectively, $\gamma(S_p) = 1$, $\gamma(S_p) = -1$). By the discussion in the last paragraph, subsets within the same family have to be disjoint. Moreover, $\gamma(x) = 0$ for exactly m elements, which shows that C_0 covers S exactly and we have a "yes" instance of 3DM.

Only If. Conversely, suppose that we have a "yes" instance of 3DM and let C_0, C_1, C_2 be the desired families of subsets. We construct γ by letting $\gamma(x_i) = 0$ (respectively, 1, -1) if x_i corresponds to an element of a subset $S_p \in C_0$ (respectively, C_1, C_2). Since C_0 is a cover to S, $\gamma(x) = 0$ for exactly m elements $x \in X$. Consequently, $E[\gamma^2(x)] = 1/(3nM)(3n - m) = B$. It remains to show that δ can be chosen so that

$x+\gamma(x)+\delta(x+\gamma(x)+v)=0$, $\forall x\in X$, $\forall v\in V$. For this it is sufficient to prove that $x_i+\gamma(x_i)+v\neq x_j+\gamma(x_j)+v'$, whenever $x_i\neq x_j$ and for all v, $v'\in V$. So, suppose that the desired inequality does not hold for some x_i, x_j, v, v'. We will derive a contradiction, but we will have to consider the various possible cases for v and v'.

(i) $v=v'=0$. If $x_i+\gamma(x_i)=x_j+\gamma(x_j)$, then $|x_i-x_j|\leqq 2$, which contradicts Lemma 3(b).

(ii) $v=0$, $v'=x_{l+1}-x_l+\rho$, $\rho\in U$. Then

$$|(x_i-x_j)-(x_{l+1}-x_l)|=|\gamma(x_j)-\gamma(x_i)+\rho|\leqq 2\nu+2<3m, \tag{2.3}$$

which implies, by Lemma 3(b), that $x_i-x_j=x_{l+1}-x_l$. It follows that $i=l+1$, $j=l$; therefore, $\gamma(x_i)=\gamma(x_j)$ and, using (2.3), $\rho=0$, which is a contradiction.

(iii) $v=x_{p+1}-x_p+\rho$, $v'=x_{l+1}-x_l+\rho'$, $\rho,\rho'\in U$. Then

$$|x_i-x_j+x_{p+1}-x_p-x_{l+1}+x_l|=|\gamma(x_j)-\gamma(x_i)-\rho+\rho'|<2\nu+2<3m,$$

which implies that $x_i-x_j+x_{p+1}-x_p-x_{l+1}+x_l=0$. So, one of the following must hold: $x_i=x_p$, $x_j=x_l$ or $x_p=x_l$. If $x_i=x_p$, it follows that $x_j=x_l$ (and conversely); in either case, we obtain $x_{p+1}=x_{l+1}$ and $x_p=x_l$; therefore, $x_i=x_j$, which is a contradiction.

(iv) $v=0$, $v'=3m(A+z_l)$. Then, $|x_i-x_j|=|\gamma(x_j)-\gamma(x_i)+3m(A+z_l)\geqq 3mA-2$, which contradicts Lemma 3(c).

(v) $v=x_{l+1}-x_l+\rho$, $\rho\in U$, $v'=3m(A+z_p)$. Then,

$$|x_i-x_j+x_{l+1}-x_l|=|\gamma(x_j)-\gamma(x_i)-\rho'+3m(A+z_p)|\geqq 3mA-2-\nu,$$

which contradicts Lemma 3(c).

(vi) $v=3m(A+z_p)$, $v'=3m(A+z_q)$. Let $x_i=3mz_i+k$, $x_j=3mz_j+k'$. Then, $3m|z_i-z_j+z_p-z_q|=|3(k'-k)+\gamma(x_j)-\gamma(x_i)|$. If $z_i-z_j+z_p-z_q=0$, then $2\geqq|\gamma(x_i)-\gamma(x_j)|=3|k-k'|$, which implies $k=k'$. Therefore, $\gamma(x_i)=\gamma(x_j)$, which is a contradiction because γ takes different values when x_i, x_j correspond to the same element of S. Therefore, $12m\leqq|z_i-z_j+z_p-z_q|\leqq 3m+2$, which is also a contradiction. This completes the proof of the lemma and the theorem. □

Decentralized and output control of imperfectly observed Markov chains. By simply observing that Witsenhausen's counterexample and the static team decision problem are at the root of several problems in decentralized control, we obtain some interesting Corollaries of Theorem 1. For example, one might be interested in formulating and studying problems of decentralized control of Markov chains. However, a single stage of such a problem would require the solution of a static team decision problem and NP-completeness (or worse) follows.

One could also formulate a problem of output control of a Markov chain, similar to the problem studied in [LA] under linear quadratic Gaussian assumptions: that is, the decision at time k would be constrained to be a function only of the observation made at time k (no recall). In fact, problem WITSENHAUSEN is a two-stage output control problem for a Markov chain and NP-completeness follows. The two-stage output control problem can be also easily seen to contain as a special case the problem of minimum distortion quantization which is also NP-complete [GJW]. Infinite horizon average cost versions of that problem can be also easily shown to be NP-complete. Finally, problems of causal coding and control of Markov chains, as defined in [WV], are also NP-complete for the same reasons.

3. Continuous problems and algorithms.

Continuous problems and their complexity. Our final aim is to derive complexity results for continuous problems. Unfortunately, there is no standard model of computa-

tion—or complexity measure—for such problems. In this subsection we shall discuss various notions of computation and complexity pertaining to continuous problems. A comparison of our framework and other existing work on the complexity of continuous problems appears in the last section.

In an instance of a typical continuous problem, we are given a finite set $F = \{f_1, \cdots, f_n\}$ of real functions (without loss of generality, with domains some power of the unit interval) and we are asked to evaluate (usually approximately) a functional $G(F) \in \Re$ of these functions. For example, $f_1, \cdots, f_n$ may be the boundary conditions for a partial differential equation and $G(f_1, \cdots, f_n)$ the value of the corresponding solution at a specific point. Closer to our concerns in this paper, we can define the continuous counterpart of the DTEAM problem mentioned in the previous section. In an instance of this problem, we are given a function $c: [0, 1]^4 \to [0, 1]$, assigning a cost to each combination of observations $y_1, y_2 \in [0, 1]$ and decisions $\gamma_1(y_1), \gamma_2(y_2) \in [0, 1]$. (Notice that we are assuming, for simplicity, that the probability distribution is uniform over $[0, 1]^2$.) The goal is to compute the functional $J^*(c)$ defined by

$$J^*(c) = \inf_{\gamma_1, \gamma_2} \int_0^1 \int_0^1 c(y_1, y_2, \gamma_1(y_1)\gamma_2(y_2))\, dy_1\, dy_2.$$

We shall be interested in the special case of this problem in which the function c is *Lipschitz continuous* with Lipschitz constant 1 (with respect to the max norm on $\Re^4$), as a representative of those special cases that we can hope to solve efficiently. Without such "smoothness" conditions, no realistic solution of continuous problems is possible, for simple information-theoretic considerations. We call the continuous version of the DTEAM problem with the Lipschitz condition the *Lipschitz continuous team problem*, or LCTEAM. That is, LCTEAM is the set of all instances, as described and restricted above. It should be obvious that a host of problems of continuous nature are amenable to similar formalization.

There are several possible notions of what it means for an algorithm to solve such a problem, and, equally important, the complexity of its operation. The subtle part is defining the sense in which the continuous functions f_i are "given". We examine a number of such approaches below.

Oracle algorithms. Continuous problems of the type defined above can often be solved by an algorithm which operates as follows: The input of the algorithm is a positive real ε, and the output is an approximation of the functional with error at most ε. Every time that the algorithm needs the value of a function f_i at some point x, this is done as follows: The algorithm submits x (a rational point), and an integer k to an *oracle* for f_i, and the oracle gives back the k most significant digits of the answer $f_i(x)$. The algorithm is "charged" for this service k steps, plus of course the time it took to construct x up to the desired precision. We say that an oracle algorithm solves a continuous problem Π in polynomial time if, for every instance I of Π there is a polynomial p_I such that the algorithm solves I within accuracy ε in time $p_I(1/\varepsilon)$.

Uniformly polynomial oracle algorithms. There is a stronger notion of efficiency, which requires that the polynomial be independent of the instance I. We call algorithms with this property *uniformly polynomial.* Notice that is a much stronger notion than that of plain polynomial-time oracle algorithms.

Note: The distinction between polynomial and uniformly polynomial algorithms has no counterpart in the context of combinatorial (discrete) problems, since in discrete problems the instance plays the role of both ε and I in the above definitions. It is, however, meaningful for continuous problems. For example, consider the problem in

which we are given one Lipschitz continuous function f over $[0, 1]$, and we are asked to compute $G(f) = \inf_{x\in[0,1]} f(x)$. A straightforward discretization leads to an algorithm with time requirements $O(K_f/\varepsilon)$, where K_f is the Lipschitz constant of f; so, this is a polynomial algorithm. On the other hand, $O(K_f/\varepsilon)$ is also a lower bound and since this problem contains instances with arbitrarily large Lipschitz constants, no uniformly polynomial algorithm exists.

Uniformly hard instances. One way to show that a continuous problem has no polynomial-time oracle algorithm at all is to exhibit an instance for which no polynomial-time oracle algorithm exists; such instances are called *uniformly hard.* Naturally, for discrete problems there are no hard single instances.

Instance-specific algorithms. We obtain an interesting variant of the concept of oracle algorithms by considering single instances of the problem Π. In each instance I, we just wish to compute a number, namely $G(F)$. We could ask the question, is this number *polynomial-time computable,* in the sense that we can compute it within accuracy ε in time polynomial in $1/\varepsilon$ by an ordinary algorithm (involving no oracles). This is a meaningful question only if the functions f_i are themselves polynomial-time computable, in that the value $f_i(x)$ can be computed in time which is polynomial in the accuracy in which x is given, and the desired accuracy.

Iterative algorithms. In numerical analysis or mathematical programming we are often interested in convergent *iterative algorithms.* These differ from the class of algorithms we introduced above in that they do not take ε as an input, and they never halt. Rather, from time to time they produce output values $G_i(F)$, $i = 1, 2, \cdots$ which are increasingly accurate approximations of $G(F)$. We may call an iterative algorithm *polynomial* if there is a polynomial p such that, for every instance, at time $p(1/\varepsilon)$, the most recent output value is accurate, within ε. It is clear that if a polynomial iterative algorithm exists, there also exists a uniformly polynomial algorithm for the problem. In fact the converse also holds [9]: take a uniformly polynomial algorithm and run it with $\varepsilon = 2^{-k}$, $k = 1, 2, \cdots$. The resulting algorithm is a polynomial iterative algorithm. For this reason, we shall not consider iterative algorithms any further.

3.7. The basic construction. Our method of connecting the complexity of the continuous version of the TEAM problem to the (much better understood) complexity of the discrete one, is based on the following lemma and construction:

LEMMA 5. *For each instance I of the* SDTEAM *problem we can define a function $c_I: [0, 1]^4 \to [0, 1]$ such that:*

(i) *Function c_I is Lipschitz continuous (with Lipschitz constant* 1*), and thus it defines an instance of* LCTEAM.

(ii) *The optimum $J^*(c_I)$ equals $1/20N^4$ if the optimum of I was* 1*, and* 0 *if it was* 0 *(recall for that instances of* SDTEAM *these are the only possibilities; N is the number of possible observations in I).*

(iii) *For any I and k-bit numbers y_1, y_2, u_1, u_2, and any $l > 0$, the l most significant bits of $c_I(y_1, y_2, u_1, u_2)$ can be computed in time polynomial in k, l, and the size of I.*

Proof. Let us first define a function $\alpha: [0, N] \to [0, 1/N]$ as follows:

$$\alpha(x) = \begin{cases} x - \lfloor x \rfloor & \text{if } x - \lfloor x \rfloor \leqq \dfrac{1}{N}, \\ \lceil x \rceil - x & \text{if } \lceil x \rceil - x \leqq \dfrac{1}{N}, \\ \dfrac{1}{N} & \text{otherwise} \end{cases}$$

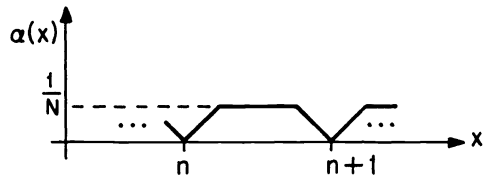


FIG. 1

(see Fig. 1), and define $q(y_1, y_2) = (1/(1-1/N)^2)\alpha(y_1)\alpha(y_2)$. Notice that q has Lipschitz constant $4/N$, and that its integral over $[0, N]^2$ is one.

Let us now recall the cost function of the discrete instance I, call it $d: \{1, 2, \cdots, N\} \times \{1, 2, 3, 4\} \to \{0, 1\}$. For $1 \leqq x_1, x_2 \leqq N$, integers, and $v_1, v_2 \in [1, 4]$, let $h(x_1, x_2, v_1, v_2)$ be the smallest $\delta \leqq 1$ such that there are u_1, u_2 with $|u_1 - v_1|, |u_2 - v_2| \leqq \delta$ and $d(x_1, x_2, u_1, u_2) = 0$, or 1 if no such δ exists. Then, define, for $0 \leqq y_1, y_2, u_1, u_2 \leqq 1$ the cost function $c_I(y_1, y_2, u_1, u_2)$ to be

$$\frac{q(Ny_1, Ny_2)}{20}[h(\lceil Ny_1 \rceil, \lceil Ny_2 \rceil, 3u_1+1, 3u_2+1) + 2p(3u_1+1) + 2p(3u_2+1)],$$

where $p(x)$ is the distance between x and its closest integer.

Let us verify the properties of c_I. To verify (i), function h has discontinuities at integral values of its first two arguments, but q is zero there, so c_I is continuous. To check that the Lipschitz constant is 1, recall that if the functions f_i have Lipschitz constants L_i and maxima m_i, $i = 1, 2$, then the function $f_1 f_2$ has Lipschitz constant $L_1 m_2 + L_2 m_1$. Within each "rectangle" of constant $\lceil Ny_1 \rceil$, $\lceil Ny_2 \rceil$ h has maximum 1 and Lipschitz constant 3, and $p(3u_1+1) + p(3u_2+1)$ has Lipschitz constant 6 and maximum 2, so their sum has maximum 3 and Lipschitz constant 9. Also, q has maximum $4/N^2$ and Lipschitz constant at most 8, and so their product has Lipschitz constant at most 20. It follows that c_I has indeed Lipschitz constant at most 1, as required.

For (ii), let us denote by $\mathcal{D}$ the set of all functions $\gamma: [0, 1] \to [0, 1]$ which are piecewise constant, with discontinuities at points i/N, and taking the values $\{0, \frac{1}{3}, \frac{2}{3}, 1\}$. We claim that, for fixed γ_2, the decision function $\gamma_1(y_1)$ which minimizes $J(\gamma_1, \gamma_2)$ is in $\mathcal{D}$. Notice that $\inf_\gamma J(\gamma_1, \gamma_2)$ is equal to

$$\frac{1}{(1-1/N)^2}\int_0^1 \alpha(Ny_1) \cdot \min_{u_1} \int_0^1 \alpha(Ny_2)[h + 2p(3u_1+1) + 2p(3\gamma_2(y_2)+1)]\, dy_2\, dy_1.$$

To carry out this minimization, it is sufficient to minimize, with respect to u_1,

$$\int_0^1 \alpha(Ny_2)[h(\lceil Ny_1 \rceil, \lceil Ny_2 \rceil, 3u_1+1, 3\gamma_2(y_2)+1)]\, dy_2 + \int_0^1 2p(3u_1+1)\, dy_2$$

for each y_1. The first term does not depend on y_1 within the interval $[i/N, (i+1)/N]$, and thus the optimum u_1 is indeed constant within this interval. Secondly, notice that the second term, together with the Lipschitz condition, ensures that the minimum is achieved at integer values of $3u_1+1$, that is, at values of γ_1 in $\{0, \frac{1}{3}, \frac{2}{3}, 1\}$.

The same argument shows that γ_2 may be constrained to be in $\mathcal{D}$ as well. Once we have shown that the optimizing decision functions are in $\mathcal{D}$, we have essentially shown that the continuous LCTEAM problem defined by c_I is in fact "isomorphic" to the discrete one I, and it has optimum which is $J^*(c_I) = (1/20N^4)J^*(I)$, where $J^*(I)$ is the optimum of I, either 0 or 1. Part (iii) is trivial. □

4. The main results. In this section we present our evidence that the Lipschitz continuous team problem is indeed intractable. We prove *three* such theorems, corresponding to three different notions of complexity of continuous problems introduced in the previous section, namely uniformly polynomial oracle algorithms, polynomial instance-specific algorithms, and uniformly hard instances. In all three cases, we show the intractability of LCTEAM, assuming that a very likely conjecture in Complexity Theory is true. Naturally, the stronger our notion of intractability of LCTEAM, the stronger the complexity-theoretic conjecture needed.

4.1. Nonexistence of uniformly polynomial algorithms.

THEOREM 2. *There is a uniformly polynomial algorithm for* LCTEAM *if and only if* P = NP.

Proof. If. Suppose that P = NP. We shall describe a uniformly polynomial algorithm for LCTEAM. The algorithm works by discretizing the problem, and obtaining appropriately approximate solutions by solving discrete instances of DTEAM (which is possible, once P = NP).

Let R correspond to a *truncation* operation: given some $x \in [0, 1]$ and some $\varepsilon > 0$, $R(x, \varepsilon)$ retains the $\lceil \log(1/\varepsilon) \rceil$ most significant bits of x; consequently, $|x - R(x, \varepsilon)| \leqq \varepsilon$ and if $\log(1/\varepsilon)$ is an integer, then $(1/\varepsilon)R(x, \varepsilon)$ is also an integer.

Given the cost function c of an instance of LCTEAM and some $\varepsilon > 0$ such that $\log(1/\varepsilon)$ is an integer, we construct an instance I of DTEAM by letting $\Delta = \varepsilon/8$, $N = 1/\Delta$, $M = 1/\Delta$ and cost function

$$d(i, j, k, l) = \frac{1}{\Delta} R(c(i\Delta, j\Delta, k\Delta, l\Delta), \Delta), \qquad (i, j, k, l) \in \left\{1, \cdots, \frac{1}{\Delta}\right\}^4. \tag{4.1}$$

Clearly, d is integer-valued and $C = \max d \leqq 1/\Delta$. Let $J^c(\gamma_1, \gamma_2)$, $J^d(\delta_1, \delta_2)$ denote the costs of pairs of decision rules for the continuous (c) and discrete (d) instances, respectively. J^{*c}, J^{*d} are the corresponding optimal costs.

LEMMA 6. $\|J^{*c} - \Delta^3 J^{*d}| \leqq \varepsilon$.

Proof. Let δ_1, δ_2 be the optimal for d. Let $\gamma_i(y_i) = \Delta\delta_i(k_i)$ for $y_i \in [(k_i - 1)\Delta, k_i\Delta]$, $k_i = 1, \cdots, N$, $i = 1, 2$. Using the definition of d and the Lipschitz continuity of c, we obtain

$$\begin{aligned} J^{*c} \leqq J(\gamma_1, \gamma_2) &= \sum_{k=1}^{N} \sum_{m=1}^{N} \int_{(k-1)\Delta}^{k\Delta} \int_{(m-1)\Delta}^{m\Delta} c(y_1, y_2, \gamma_1(y_1), \gamma_2(y_2))\, dy_1\, dy_2 \\ &\leqq \Delta^2 \sum_{k=1}^{N} \sum_{m=1}^{N} (\Delta d(k, m, \delta(k), \delta(m)) + 2\Delta) \qquad (4.2) \\ &= \Delta^3 J^d(\delta_1, \delta_2) + 2\Delta = \Delta^3 J^{*d} + 2\Delta \leqq \Delta^3 J^{*d} + \varepsilon. \end{aligned}$$

In order to prove the converse inequality, suppose that $\bar{\gamma}_1, \bar{\gamma}_2 : [0, 1] \to [0, 1]$, are such that $J^c(\bar{\gamma}_1, \bar{\gamma}_2) \leqq J^{*c} + \Delta$. Let $f(y_1, y_2) = \int_0^1 c(y_1, y_2, u_1, \bar{\gamma}_2(y_2))\, dy_2$. Then f is also Lipschitz continuous with Lipschitz constant 1. It follows that

$$|\inf_u f(y, u) - \inf_u f(y', y)| \leqq 2|y - y'| \quad \forall y, y' \in [0, 1].$$

Let $\hat{\gamma}_1(y_1) = \operatorname{argmin}_{u \in [0,1]} f(k\Delta, u)$, for $y_1 \in ((k-1)\Delta, k\Delta)$. Then,

$$\begin{aligned} J^c(\hat{\gamma}_1, \bar{\gamma}_2) &= \int_0^1 f(y_1, \hat{\gamma}_1(y_1))\, dy_1 \leqq \int_0^1 \left(\inf_{u \in [0,1]} f(y_1, u_1) + 2\Delta \right) dy_1 \\ &= \inf_{\gamma_1} \int_0^1 f(y_1, \gamma_1(y_1))\, dy_1 + 2\Delta = \inf_{\gamma_1} J^c(\gamma_1, \bar{\gamma}_2) + 2\Delta \\ &\leqq J^c(\bar{\gamma}_1, \bar{\gamma}_2) + 2\Delta \leqq J^{c*} + 3\Delta. \end{aligned}$$

In a similar way, we may construct a piecewise constant function $\hat{\gamma}_2: [0, 1] \to [0, 1]$ such that

$$J^c(\hat{\gamma}_1, \hat{\gamma}_2) \leqq J^c(\hat{\gamma}_1, \bar{\gamma}_2) + 2\Delta \leqq J^{c*} + 5\Delta.$$

The decision rules $\hat{\gamma}_i$, $i = 1, 2$, being piecewise constant determine corresponding decision rules $\hat{\delta}_i$, $i = 1, 2$, for the discrete cost function. Then, a chain of inequalities similar to (4.2) leads to

$$\Delta^3 J^{d*} \leqq \Delta^3 J^d(\hat{\delta}_1, \hat{\delta}_2) \leqq J^c(\hat{\gamma}_1, \hat{\gamma}_2) + 2\Delta \leqq J^{c*} + 7\Delta \leqq J^{c*} + \varepsilon. \qquad \square$$

Since P = NP there exists an algorithm for the problem of computing the optimal cost of any instance of DTEAM (based on the algorithm for DTEAM and binary search), which is polynomial in M, N, $\log C$, where C is the largest integer appearing in the cost function. Consider then the following algorithm for LCTEAM:

(i) Decrease ε (at most by a factor of 2) so that $\log(1/\varepsilon)$ is an integer.

(ii) Use the oracle to read the $\log(8/\varepsilon)$ most significant bits of $c(i\Delta, j\Delta, k\Delta, m\Delta)$, $1 \leqq i, j, k, m \leqq N$, where $N\Delta = 1$, $\Delta = \varepsilon/8$.

(iii) Run the assumed algorithm on the resulting instance of DTEAM, as defined by (4.1). Multiply the output by Δ^3 and return it.

This is clearly a uniformly polynomial algorithm, and the proof of the *if* part is complete.

Only If. If we had a uniformly polynomial algorithm for LCTEAM, we could solve any instance I of SDTEAM of size N as follows:

(i) Construct the corresponding instance c of LCTEAM, as in Lemma 5.

(ii) Simulate the assumed algorithm for LCTEAM on it, with desired accuracy $\varepsilon = 1/40N^4$. The time required is polynomial in N, including the computations of c, which, by Lemma 5(iii), are polynomially related to the "charges" for oracle calls of the corresponding computation of the algorithm.

(iii) If the result is less than $\varepsilon = 1/40N^4$, then the optimum cost of SDTEAM was 0, otherwise 1.

Since this is a polynomial-time algorithm for SDTEAM, an NP-complete problem, it follows that P = NP. $\square$

4.2. A hard instance with efficiently computable cost.

THEOREM 3. *If* DEXP $\neq$ NDEXP *then there exists an instance c of* LCTEAM *such that*

(i) *The cost* $c(y_1, y_1, u_1, u_2)$, *where* y_1, y_2, u_1, u_2 *are k-bit numbers between* 0 *and* 1 *can be computed with accuracy* ε *in time polynomial in* 2^k *and* $1/\varepsilon$, *whereas*

(ii) *The optimum value* J^* *is not polynomially computable; that is, it cannot be computed within accuracy* ε *in time polynomial in* $1/\varepsilon$.

Proof. We first need a lemma concerning the existence of certain "hard" sequences of instances of NP-complete problems, in the spirit of [HSI].

LEMMA 7. *If* DEXP $\neq$ NDEXP *then there is a sequence* $I_1, I_2, \cdots$ *of instances of* SDTEAM *such that*:

(a) *Instance* I_i *has size (that is,* N*) equal to* 2^i.

(b) *There is an algorithm which, given i, constructs* I_i *in time polynomial in* 2^i *(the size of the instance produced).*

(c) *There is no polynomial-time algorithm that solves all instances* I_i.

Sketch. Consider a problem L in NDEXP $-$ DEXP. Without loss of generality, instances of L are encoded in binary, and therefore an instance i also represents an integer, in binary. For each instance i of L, let $f(i)$ be the string of length 2^i which starts with the string i and has 0's in all other positions. The language $f(L) = \{f(i): i \in L\}$

is in NP (since L is in NDEXP), and thus there is a polynomial-time transformation that transforms each string $f(i)$ to an instance of SDTEAM such that the instance of SDTEAM is a "yes" instance if and only if $f(i) \in f(L)$. By "padding" these instances of SDTEAM to make them of size a power of two, and filling in the gaps with "null" instances, we obtain the sequence of the lemma. □

To show the theorem, consider a sequence of instances as constructed in the lemma. For each such instance I_i, we construct a continuous function $c_i: [0, 1]^4 \to [0, 1]$, as in Lemma 5. Consider now a *scaled, shifted* version of c_i, call it c_i', with support $[1-2^{-i}, 1-2^{-(i+1)}]^2 \times [0, 1]^2$ (see Fig. 2), defined in this range as

$$c_i'(y_1, y_2, u_1, u_2) = \frac{1}{2^{i+1}} c_i(2^{i+1}(y_1 - 1 + 2^{-i}), (2^{i+1}(y_2 - 1 + 2^{-i}), u_1, u_2)).$$

c_i is zero outside this domain. Finally, define the function c to be

$$c(y_1, y_2, u_1, u_2) = \sum_{i=1}^{\infty} c_i'(y_1, y_2, u_1, u_2).$$

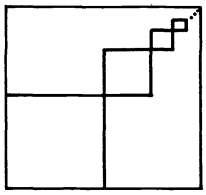


FIG. 2

This function is Lipschitz continuous with constant 1 (due to the scaling), as required by the theorem, and it is easy to see that it satisfies condition (i) (compare with (iii) of Lemma 5). To show (ii), notice that the optimum value J^* corresponding to c can be expressed in terms of the optima J_i^{*d} of the instances I_i that comprise it, as follows:

$$J^* = \sum_{i=1}^{\infty} J_i^{*d} \frac{1}{20 \cdot 2^{4i}} \frac{1}{2^{i+1}} \frac{1}{2^{2(i+1)}}.$$

The first term of the addend is the cost of the original discrete instance, known to be either 0 or 1 in SDTEAM. The second term was introduced by the construction of Lemma 5. The third is due to the scaling, whereas the last term represents the area of the support of instance c_i', as defined above. Thus, $J^* = \frac{1}{160} \sum_{i=1}^{\infty} J_i^{*d} 2^{-7i}$. From the form of the sum, it is evident that, if we could compute J^* within accuracy ε in time polynomial in $1/\varepsilon$, then we could compute the optimum cost of I_i in time polynomial in the size of I_i, contrary to Lemma 7. □

Theorem 3 has a weak converse. It can be shown that, if an instance of LCTEAM as described in Theorem 3 exists, then EXP $\neq$ NEXP. The argument goes as follows: If such a hard instance exists, then its *discretizations* (that is, the sequence of discrete problems resulting by subdividing the unit interval in 2^i equal intervals, and by defining the cost function on this grid by a restriction of the continuous cost function) are not all solvable by polynomial-time algorithms. Thus, we have a hard exponentially sparse sequence of *optimization* problems of the DTEAM type (in which we are asked to determine the optimum cost). Since each optimization problem in this sequence can be reduced to an exponential number of a *recognition* problems (asking whether the

cost of an instance is below some bound), say, by binary search, [PS], we obtain a hard polynomially sparse sequence of instances of this problem, known to be NP-complete. The existence of such hard sequences is known (see [HSI], or the argument above) to be equivalent to EXP $\neq$ NEXP.

A uniformly hard instance. If P $\neq$ NP we can show something stronger than the nonexistence of uniformly polynomial algorithms proved in Theorem 2. In particular, we can show that there is a uniformly hard instance of LCTEAM, which "fools" all polynomial-time oracle algorithms. Our construction has to use diagonalization arguments which are not polynomially constructive, and as a result the instance constructed is not one that can be computed efficiently. Since a complete proof of this result would require the introduction of machinery in a scale disproportional to the information added, we only present an outline of the proof.

THEOREM 4. *There is a uniformly hard instance of* LCTEAM *if and only if* P $\neq$ NP.

Sketch. One direction follows from Theorem 2. For the *if* direction, we first need to define a discrete analog of an oracle algorithm. One way to do this is to consider *sequence algorithms*, that is, algorithms which operate on infinite sequences of instances of a problem. Such an algorithm accepts as its input an infinite tape with the sequence, together with an integer i, and it returns the answer ("yes" or "no") of the ith instance of the sequence. To capture the charges due to precision of the queries and the answers of oracle machines, we require that the algorithm is charged $\lceil \log k \rceil$ to determine the value of the kth bit of its input tape.

We first show that, if P $\neq$ NP, there is a sequence of instances of the SDTEAM problem, of size exponentially increasing, which cannot be solved by any sequence algorithm. The construction is carried out by enumerating all polynomial-time sequence algorithms, and using the ith non-"null" instance in the sequence to rule out the ith sequence algorithm as a potential solver of the present instance (i.e., sequence). Since P $\neq$ NP, an instance on which the ith sequence algorithm does the wrong thing, and which is of size larger than some given bound, must exist. To avoid the possibility in which the ith algorithm takes "advice" from the previous or subsequent instances in solving the current instance, we interject a doubly exponential number of "null" instances between two such consecutive instances.

We finally construct an instance of the LCTEAM problem from the given sequence of SDTEAM problems, exactly as in the proof of Theorem 3. If this instance could be solved by some oracle algorithm, it can be argued that the sequence constructed in the previous paragraph can be solved by a sequence algorithm, which is impossible by its construction. □

5. Discussion. We have shown that the team decision problem with a Lipschitz continuous cost function and uniform probability distribution holds the same place in the continuous world that NP-complete problems do in the discrete world: it possesses an approximate algorithm which is polynomial in the desired accuracy if and only if P = NP. A similar result can be also proved if Lipschitz continuity is replaced by some other, possibly stronger, smoothness requirement such as once or twice differentiability, etc. Only the construction in Lemma 5 would have to be a little more elaborate. A similar result is also possible for Witsenhausen's counterexample in stochastic control if the assumption of normality of the underlying random variables is relaxed. Since the team decision problem is a basic component of (generally harder) problems in decentralized stochastic control, such problems (at least in the absence of any more special structure) are qualitatively different from the vast majority of traditional problems in continuous mathematics and classical control. Such problems,

including nonlinear optimization, filtering and control, as well as partial differential equations, possess algorithms which are polynomial in the desired accuracy, when some smoothness conditions are satisfied, and are solvable from a realistic point of view.

The proofs of our results are based on the fact that the discrete version of the team problem is NP-complete. In this sense, we demonstrate that NP-completeness results can be exploited to make inferences about the computational complexity of continuous problems. It should be noted, however, that the various notions of intractability used call for conjectures of varying strength from Complexity Theory, all of them implying $P \neq NP$.

The proofs of Theorems 2, 3, and 4 determine a methodology that can be applied to obtain similar negative complexity results concerning other continuous problems as well. Abstracting the main elements of the proofs, we see that the following properties of LCTEAM were heavily used:

(i) The discrete version of the problem of interest should be NP-complete.

(ii) We should be able to take an instance of the discrete problem and construct an instance of the continuous problem as in Lemma 5, while respecting certain smoothness requirements.

(iii) The above construction should be simple enough, so that the corresponding oracle calls can be efficiently simulated by a Turing machine.

(iv) Finally, it should be possible to take a sequence of increasingly large discrete instances and imbed them into a single one, while keeping the dimension of the continuous problem constant.

Finally, let us comment on the relation and some differences of our framework with other theories of complexity for continuous problems. The complexity of any algorithm solving a continuous problem can be roughly divided into two kinds of activities: oracle calls to obtain information about the instance to be solved and computations based on the values returned by the oracle. A lot of past research [TW], [TWW], [NY] has obtained lower bounds on the overall complexity by deriving lower bounds on the number of oracle calls necessary to obtain enough information so that an ε-approximate solution is possible. This is a valid approach for the types of problems emphasized in that research (mainly mathematical programming and numerical integration of partial differential equations) and has produced many interesting results; the main reason is that in such problems the amount of any further computation necessary can be bounded by a polynomial (and some times linear) function of the number of oracle calls. The team problem, however, is different: while $O(1/\varepsilon^4)$ oracle calls provide sufficient information for an ε-approximate solution, we have shown that further computations require time which is exponential in ε (unless $P = NP$). In other words, the structure of the team problem forces us to emphasize its computational complexity rather than its informational requirements.

Much closer to our approach are the very interesting recent results in [Ko]. In that paper, it is shown that there are ordinary differential equations which are given in terms of easily computable functions, but which cannot be integrated efficiently, unless P = PSPACE. In this sense, Ko's results are quite similar in spirit to Theorem 3. One of the differences is that Ko's notion of efficiency requires that algorithms operate in time polynomial in the *logarithm* of $1/\varepsilon$.

REFERENCES

[AHU] A. V. AHO, J. E. HOPCROFT AND J. D. ULLMAN, *The Design and Analysis of Computer Algorithms*, Addison-Wesley, Reading, MA, 1974.

[GJ] M. R. GAREY AND D. S. JOHNSON, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W. H. Freeman, San Francisco, CA, 1979.

[GJW] M. R. GAREY, D. S. JOHNSON AND H. S. WITSENHAUSEN, *The complexity of the generalized Lloyd-Max problem*, IEEE Trans. Inform. Theory, IT-28 (1982), pp. 255-256.

[HC] Y. C. HO AND T. S. CHANG, *Another look at the nonclassical information problem*, IEEE Trans. Automat. Control, AC-25 (1980), pp. 537-540.

[HSI] J. HARTMANIS, V. SEWELSON AND N. IMMERMAN, *Sparse Sets in* NP−P: EXPTIME *vs.* NEXPTIME, Proc. 1983 STOC Conference, 1983, pp. 382-391.

[HU] J. E. HOPCROFT AND J. D. ULLMAN, *Introduction to Automata Theory, Languages and Computation*, Addison-Wesley, Reading, MA, 1979.

[Ko] K.-I. KO, *On the computational complexity of ordinary differential equations*, Inform. and Control, 58 (1984), pp. 157-194.

[LA] W. S. LEVINE AND M. ATHANS, *On the determination of optimal output feedback gains for linear multivariable systems*, IEEE Trans. Automat. Control, AC-15 (1970), pp. 44-48.

[MR] J. MARSCHAK AND R. RADNER, *The Economic Theory of Teams*, Yale Univ. Press, New Haven, CT, 1972.

[NY] A. S. NEMIROVSKY AND D. B. YUDIN, *Problem Complexity and Method Efficiency in Optimization*, John Wiley, New York, 1983.

[Pa] C. H. PAPADIMITRIOU, *Games against nature*, Proc. 1983 IEEE Conference on Foundations of Computer Science; J. Comput. System Sci. (1986), to appear.

[PS] C. H. PAPADIMITRIOU AND K. STEIGLITZ, *Combinatorial Optimization: Algorithms and Complexity*, Prentice-Hall, Englewood Cliffs, NJ, 1982.

[PT] C. H. PAPADIMITRIOU AND J. N. TSITSIKLIS, *On the complexity of designing distributed protocols*, Inform. and Control, 53 (1982), pp. 211-218.

[Ra] R. RADNER, *Team decision problems*, Ann. Math. Statist., 33 (1962), pp. 857-881.

[TW] J. F. TRAUB AND H. WOZNIAKOWSKI, *A General Theory of Optimal Algorithms*, Academic Press, New York, 1980.

[TWW] J. F. TRAUB, G. W. WASILKOWSKI AND H. WOZNIAKOWSKI, *Information, Uncertainty, Complexity*, Addison-Wesley, Reading, MA, 1983.

[TA] J. N. TSITSIKLIS AND M. ATHANS, *On the complexity of decentralized decision making and detection problems*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 42-50.

[Ts] J. N. TSITSIKLIS, *Problems in decentralized decision making and computation*, Ph.D. Thesis, Dept. Electrical Engineering and Computer Science, Massachusetts Inst. Technology, Cambridge, MA, 1984.

[WV] J. C. WALRAND AND P. VARAIYA, *Optimal causal coding-decoding problems*, IEEE Trans. Inform. Theory, IT-29 (1983), pp. 814-820.

[Wi] H. S. WITSENHAUSEN, *A counterexample in stochastic optimum control*, this Journal, 6 (1968), pp. 138-147.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 4, July 1986

005

STOCHASTIC MINIMIZATION WITH CONSTANT STEP-SIZE: ASYMPTOTIC LAWS*

GEORG CH. PFLUG†

Abstract. The behavior of the stationary distribution of a Markovian process of the form

$$(*) \qquad X_{n+1}^a = \pi_S(X_n^a - aY_n^a)$$

as a tends to zero is studied. The process (*) is a constant step-size (constant gain) constrained stochastic approximation process, with π_S being the projection onto the convex set of constraints S. If the gains a are held constant, the process $\{X_n^a\}$ does not converge to the point of solution but the stationary distribution of $\{X_n^a\}$ converges to the unit mass at the solution, if $a \to 0$. Properly normalized the stationary distribution converges to a nondegenerated limit. This limit depends on the smoothness of the feasible set S in the neighborhood of the solution (in particular on the dimension of the largest linear subspace contained in the tangential cone).

Key words. stochastic optimization, stochastic approximation, constrained problems, constant gain, asymptotic distribution

1. Introduction and general results. The aim of this paper is to study the probabilistic behavior of stochastic recursions of the form

$$(1) \qquad X_{n+1}^a = \pi_S(X_n^a - aY_n^a), \qquad X_n^a \in \mathbb{R}^m, \quad n \geqq 1,$$

where π_S denotes the projection on the closed convex set S.

We assume that X_n^a, Y_n^a are random vectors in $\mathbb{R}^m$ and that the conditional distribution of Y_n^a given the history $X_1^a, \cdots, X_n^a$ depends only on X_n^a (and is in particular not explicitly dependent either on the iteration step n or on the real parameter $a > 0$). Thus $\{X_n^a\}$ is, for each fixed a, a *Markovian process.* In the interpretation given below, a is the *step-size parameter.* In this paper we shall be especially interested in the stationary distribution of (1) for small a.

Recursive processes of the form (1) are widely used to approximate the solution of the *stochastic program*

$$(2) \qquad \int_\Omega H(x, \omega)\, d\nu(\omega) = \min!, \qquad x \in S$$

where ν is a probability measure on some measurable space (Ω, A) and $H(\cdot, \cdot)$ is a known jointly measurable function $\mathbb{R}^m \times \Omega \to \mathbb{R}^1$. The set S is a closed, convex set of constraints in $\mathbb{R}^m$. The application $x \mapsto H(x, \omega)$ maps $\mathbb{R}^m$ into $L^1(\Omega, A, \nu)$. If $F(x) := \int H(x, \omega)\, d\nu(\omega)$ is known explicitly then the problem (2) can be viewed as a deterministic constrained minimization problem. In most cases of applications however $F(\cdot)$ is not known in an explicit manner. There are, in principle, three possible strategies for solution in this case:

(i) Try to calculate $F(\cdot)$ by numeric integration and proceed as in the deterministic case.

(ii) Try to approximate ν by a measure ν_0 such that $\int_\Omega H(x, \omega)\, d\nu_0(\omega)$ is close to $\int_\Omega H(x, \omega)\, d\nu(\omega)$ such that $\int_\Omega H(x, \omega)\, d\nu_0(\omega)$ is easily computed.

(iii) Use stochastic approximation.

* Received by the editors May 15, 1984, and in revised form April 3, 1985.

† Mathematical Institute, J. Liebig University, Arndtstrasse 2, Giessen D-6300, Federal Republic of Germany.

The first two approaches lead to deterministic procedures. Only the stochastic approximation method is stochastic in nature. This method will be considered here. To explain the idea behind this we have to introduce the notion of an L^1-derivative.

DEFINITION. A m-dimensional vector $\mathbf{H}_x(x^*, \omega)$ of L^1-functions is called the L^1-derivative of the function $x \mapsto H(x, \omega)$ at the point x^*, if

$$\int |H(x, \omega) - H(x^*, \omega) - (x - x^*)' \mathbf{H}_x(x^*, \omega)| \, d\nu(\omega) = o(\|x - x^*\|)$$

as x tends to x^*.

Suppose that the application $x \mapsto H(x, \omega)$ has for every x a L^1-derivative $\mathbf{H}_x(x, \cdot) = (H_{x,1}(x, \cdot), \cdots, H_{x,m}(x, \cdot))$. Then the random vector $\mathbf{H}_x$ is an unbiased estimate for the gradient(grad) of F, since

$$\int \mathbf{H}_x(x, \omega) \, d\nu(\omega) = \text{grad } F(x) =: f(x)$$

by componentwise integration.

The construction of a random sequence $\{X_n\}$ approximating the solution x_0 of the stochastic program (2) is done in a recursive way. Suppose X_1 is any starting value and assume that X_n is already constructed. Let Y_n be a random vector which is independent of $X_1, \cdots, X_{n-1}$ but not on X_n, such that its conditional distribution given $X_n = x_n$ coincides with the image measure of ν under $\mathbf{H}_x(x_n, \cdot)$, i.e.

$$P\{Y_n \in A | X_n = x_n\} := \nu\{\omega | \mathbf{H}_x(x_n, \omega) \in A\}.$$

By this construction

$$E(Y_n | X_1, \cdots, X_n) = E(Y_n | X_n) = f(X_n) = \text{grad } F(X_n).$$

The unknown solution point of (2) can be approximated by a stochastic version of the gradient method, using the *stochastic gradients* Y_n, namely

$$X_{n+1} = \pi_s(X_n - a_n Y_n). \tag{3}$$

Here a_n are appropriately chosen step-sizes. Many authors have considered such recursive procedures ([1], [3], [4] among others). It was shown that under some conditions—including $a_n \geqq 0$, $\sum a_n = \infty$, $\sum a_n^2 < \infty - X_n \to x_0$ a.s. where x_0 is the unique minimal point of (2).

For the practical implementation of algorithms of the type (3) asymptotic laws (for $a_n \to 0$) are of limited importance. The reason is that for the first iterations the absolute values of the step-size constants a_n are much more important than their speed of convergence to zero. For larger n however the constants $a_n = a/n$, which is the usual choice, converge very slowly to zero. Therefore we might as well take a_n to be constant, but small.

For this reason we shall consider in this paper the algorithm (1), i.e. the situation for fixed but small step size a. For the unconstrained process, a similar approach was used by Kushner and Hai Huang [5]. As we shall show, X_n^a converges in law to a *stationary sequence* fulfilling (1) irrespectively of the starting value. Let X^a be distributed according to this stationary distribution. We shall be especially interested in the limiting distribution of

$$\tilde{X}_n^a =: \frac{X^a - x_0}{g(a)} \tag{4}$$

as a tends to zero. $g(a)$ is a properly chosen normalization and x_0 is the point of

solution of the problem (2), which is assumed to be unique. Thus x_0 satisfies

$$x_0 \in S \qquad -f(x_0) \perp C_0 \quad \text{(i.e. } y' \cdot f(x_0) \geqq 0 \text{ for all } y \in C_0\text{)}$$

where C_0 is the *tangent cone* of S at x_0, i.e. C_0 is the closure of the set $\{\lambda(x - x_0) | x \in S, \lambda \geqq 0\}$.

We have assumed that the process X_n^a is a time-homogeneous Markov chain. Denote by $P^a(\cdot, \cdot)$ its transition probabilities, i.e.

$$P^a(x, A) = P\{X_{n+1}^a \in A | X_n^a = x\}$$

where $x \in S$ and $A \in \mathfrak{B}_S = \{\text{Borel sets}\} \cap S$. The transition probabilities can be chosen as Markov kernels, i.e.

(i) $x \mapsto P^a(x, A)$ is measurable for every $A \in \mathfrak{B}_S$,
(ii) $A \mapsto P^a(x, A)$ is a probability measure for every $x \in S$.

Markov kernels induce operations on the set of probability measures on $\mathfrak{B}_S$ by

$$P^a\mu\colon \quad A \mapsto \int P^a(x, A)\, d\mu(x)$$

and on the set of all bounded $\mathfrak{B}_S$ measurable functions by

$$P^a f\colon \quad x \mapsto \int f(y) P^a(x, dy).$$

The n-fold convolution of P^a is denoted by $P_n^a(\cdot, \cdot)$ and is again a Markov kernel.

We shall show that under some not very restrictive assumptions the chain X_n^a is recurrent (in the sense of Harris) and aperiodic. Hence there is a unique invariant measure μ^a and

$$\| P_n^a \gamma - \mu^a \| \to 0$$

for every "starting measure" γ, where $\| \ \|$ denotes the variational norm. μ^a is the distribution of the stationary random vector X^a. Moreover we shall be interested in the limit distribution of the normalized random vectors $\tilde{X}^a$ as $a \to 0$ (see (4)).

The analytic theory of Markov chains is used extensively in this paper. (A general reference is the book of Revuz [8].) In contrast to the unconstrained case it is not always possible to give a diffusion approximation for the process in the constrained case. (See § 3.) Therefore a discrete time Markovian process is used to describe also the asymptotic behavior.

We begin with stating the assumptions. We rewrite the algorithm (1) in the form

$$X_{n+1}^a = \pi_S(X_n^a - af(X_n^a) - aZ(X_n^a, \omega)) \tag{5}$$

where $Z(x, \omega) := Z(x) := \mathbf{H}_x(x, \omega) - f(x)$ are independent zero mean variables. Sometimes we write simply Z_n instead of $Z(X_n)$. When there is no ambiguity, we occasionally drop the superscript a, writing X_n instead of X_n^a. The inner product in $\mathbb{R}^m$ is denoted by $\langle \cdot, \cdot \rangle$. A' denotes the transpose of A and $\xrightarrow{w}$ means weak convergence. The starting value X_1 of the recursion may be an arbitrary random variable with finite second moment. We write P_{X_1} for the conditional probability given the starting value X_1. P_x means the law of the process started at $X_1 = x$. The corresponding expectation is denoted by E_x.

Assumptions (A).

(i) For every $\varepsilon > 0$ there is a $\delta_\varepsilon > 0$ such that

$$\inf_{\{x \in S \mid \|x - x_0\| \geqq \varepsilon\}} \frac{\langle x - x_0, f(x)\rangle}{\|x - x_0\|^2} \geqq \delta_\varepsilon.$$

(ii) $\|f(x)\| \leqq A + B\|x - x_0\|$ for some constants A, B.

(iii) $E(\|Z(x)\|^2) \leqq A_1 + B_1\|x + x_0\|^2$ for some constants A_1, B_1.

(iv) $x \mapsto f(x)$ is continuous for $x \in S$.

(v) $x \mapsto Z(x, \cdot)$ is continuous as a mapping

$$\mathbb{R}^m \to [L_2(\Omega, A, \nu)]^m.$$

(vi) The Lebesgue measure on $\mathbb{R}^m$ is absolutely continuous with respect to the disribution of $Z(x)$, for all $x \in S$.

Remark on the assumptions. The Assumption A(i) implies that $V(x) := \|x - x_0\|^2$ is a Lyapunov-type function. Likewise any function with similar properties could be used instead. Assumptions A(iv) and A(v) imply the Feller property of the transition probabilities. Indeed, if g is continuous and bounded on S then due to the continuity of the projection operator $E(g(X_2)|X_1 = x) = E(g(\pi_S(x - af(x) - aZ(x, \cdot))))$ is continuous in x. The condition A(vi) ensures the irreducibility of the process (Lemma 2) and is equivalent to: If $A \subseteq S$ has positive Lebesgue measure then $P(x, A) > 0$ for all x. This condition seems to be rather stringent, but in fact can always be easily fulfilled by adding to each random observation Y_n a small noise with nonvanishing density in $\mathbb{R}^m$ (e.g. a normally distributed noise).

LEMMA 1. *Let* $B_\varepsilon := \{x \in S \mid \|x - x_0\| \leqq \varepsilon\}$, *and let* τ_ε *be the first hitting time of* B_ε. *Then for every* $\varepsilon > 0$ *there is an* $a(\varepsilon) >$ *such that for* $0 < a \leqq a(\varepsilon)$

$$P_{X_1}\{\tau_\varepsilon > k\} \leqq \min\left(\varepsilon^{-2}(1 - a\beta_\varepsilon)^{k-1}\|X_1 - x_0\|^2, 1\right) \tag{6}$$

where $\beta_\varepsilon < 2\delta_\varepsilon$.

Proof. We denote by $\mathcal{F}_n$ the σ-algebra generated by $X_1, \cdots, X_n$. It follows from the assumptions that there are constants A_2, B_2 such that $\|f(x)\|^2 + E(\|Z(x)\|^2) \leqq A_2 + B_2\|x - x_0\|^2$. On the set $\{\tau_\varepsilon > n\} \in \mathcal{F}_n$ we have

$$\begin{aligned}
E(\|X_{n+1} - x_0\|^2|\mathcal{F}_n) &= E(\|\pi_S(X_n - af(X_n) - aZ_n) - x_0\|^2|\mathcal{F}_n) \\
&\leqq E(\|X_n - af(X_n) - aZ_n - x_0\|^2|\mathcal{F}_n) \\
&= \|X_n - x_0\|^2 + a^2\|f(X_n)\|^2 + a^2E(\|Z_n\|^2|\mathcal{F}_n) - 2a\langle f(X_n), X_n - x_0\rangle \\
&\leqq \|X_n - x_0\|^2 + a^2A_2 + a^2B_2\|X_n - x_0\|^2 - 2a\delta_\varepsilon\|X_n - x_0\|^2 \\
&\leqq \|X_n - x_0\|^2(1 - a\beta_\varepsilon)
\end{aligned}$$

for $\beta_\varepsilon < 2\delta_\varepsilon$ and sufficiently small a. It follows that $\|X_{n\wedge\tau_\varepsilon} + x_0\|^2(1 - a\beta_\varepsilon)^{-(n\wedge\tau_\varepsilon)}$ is a nonnegative supermartingale. Since for such supermartingales V_n

$$P_{V_1}\{V_n \geqq b\} \leqq \min\left(\frac{V_1}{b}, 1\right)$$

(cf. Neveu [6, Prop. II-2-7]) we conclude that

$$\begin{aligned}
P_{X_1}\{\tau_\varepsilon > k\} &\leqq P_{X_1}\{\|X_{k\wedge\tau_\varepsilon} - x_0\|^2 > \varepsilon^2\} \\
&\leqq P_{X_1}\{\|X_{k\wedge\tau_\varepsilon} - x_0\|^2(1 - \alpha\beta_\varepsilon)^{-(k\wedge\tau_\varepsilon)} \geqq \varepsilon^2(1 - a\beta_\varepsilon)^{-k}\} \\
&\leqq \min\left(\varepsilon^{-2}(1 - a\beta_\varepsilon)^{k-1}\|X_1 - x_0\|^2, 1\right).
\end{aligned}$$

LEMMA 2. *Let λ_S be the restriction of the Lebesgue measure λ to the set S. Then $\{X_n\}$ is λ_S-irreducible and aperiodic.*

Proof. For every fixed $x \in S$ we can decompose the transition probability $P(x, \cdot)$ in the absolute continuous part w.r.t. Lebesgue measure and the orthogonal part. According to the Assumption A(vi) the absolute continuous part has a nonvanishing density. Hence if $\lambda_S(A) > 0$ then $P(x, A) > 0$ whence the irreducibility and the aperiodicity follow.

LEMMA 3. *The process $\{X_n\}$ is recurrent in the sense of Harris and ergodic.*

Proof. By [8, Thms. 3.2.6, 3.2.7] we have to show that the potential kernel

$$G(x, A) = E\left(\sum_{n=1}^{\infty} 1_A(X_n) \,\middle|\, X_1 = x\right)$$

is not a proper kernel. We define an increasing sequence of stopping times by

$$\begin{aligned} \tau_1 &= \inf\{n > 0 | X_n \in B_\varepsilon\}, \\ &\vdots \\ \tau_{m+1} &= \inf\{n > \tau_m | X_n \in B_\varepsilon\}. \end{aligned}$$

We claim that $P(\tau_m < \infty) = 1$ for all m. Assume that there is a first $m \in \mathbb{N}$ such that $P(\tau_m = \infty) > 0$. Because of the strong Markov property and Lemma 1

$$P\{\tau_m - \tau_{m-1} > k\} \leqq E(\min(\varepsilon^{-2}(1 - \alpha\beta_\varepsilon)^{k-1} \| X_{\tau_{m-1}+1} - x_0 \|^2, 1))$$

which contradicts $P\{\tau_m = \infty\} > 0$. Let $g(\cdot) \not\equiv 0$ be a continuous nonnegative function with support in S. Due to the Feller property of the chain

$$h(x) := E(g(X_2) | X_1 = x)$$

is continuous in x. Because of Assumption A(vi) and the continuity of g, $h(x) > 0$. Hence $\inf_{x \in B_\varepsilon} h(x) \geqq \eta > 0$ since B_ε is compact. Therefore

$$E_x(g(X_{\tau_k+1})) \geqq \eta \quad \text{for all } x \in S$$

and

$$E_x\left(\sum_{n=1}^{\infty} g(X_n)\right) \geqq E_x\left(\sum_{k=1}^{\infty} g(X_{\tau_k+1})\right) = \infty.$$

Thus by considering a continuous $g \not\equiv 0$, $0 \leqq g \leqq 1_A$ we see that

$$G(x, A) := E\left(\sum_{n=1}^{\infty} 1_A(X_n) \,\middle|\, X_1 = x\right)$$

equals ∞ for every open set $A \subseteq S$ and this implies that the kernel is not proper.

The Markov chain is therefore recurrent in the sense of Harris. There is a unique (up to multiplicative constant) σ-finite invariant measure μ. It remains to show that μ is finite, which is equivalent to the ergodicity of the process.

Since on $\{\| X_n - x_0 \| \leqq \varepsilon\} \in \mathcal{F}_n$

$$E(\| X_{n+1} - x_0 \|^2 | \mathcal{F}_n) \leqq \| X_n - x_0 \|^2 + a^2(A_2 + B_2\varepsilon^2)$$

there are constants ξ_ε, $\eta_\varepsilon > 0$ such that

$$E(\| X_{n+1} - x_0 \|^2 | \mathcal{F}_n) \leqq \| X_n - x_0 \|^2 + a^2\xi_\varepsilon 1_{B_\varepsilon}(X_n) - a\eta_\varepsilon 1_{\bar{B}_\varepsilon}(X_n). \tag{7}$$

We consider the chain with $X_1 \equiv x_0$. Then (7) implies that $v_n := E(\| X_n - x_0 \|^2)$ is finite for every n. Moreover

$$0 \leqq v_{n+1} \leqq v_n + a^2\xi_\varepsilon P_{n-1}(x_0, B_\varepsilon) - a\eta_\varepsilon P_{n-1}(x_0, \bar{B}_\varepsilon). \tag{8}$$

If the chain were null recurrent, then $P_n(x_0, B_\varepsilon) \to 0$ (see [8, Theorem VI, 2.11]). In this case there exists a n_0 such that for $n \geqq n_0 P_n(x_0, B_\varepsilon) < \eta_\varepsilon / a\xi_\varepsilon + \eta_\varepsilon)$ which leads to a contradiction to (8). Hence the invariant measure μ^a can be chosen as a probability measure, and the Lemma is shown.

Integrating (7) with respect to μ^a we get

$$a\eta_\varepsilon \mu^a(\bar{B}_\varepsilon) \leqq a^2 \xi_\varepsilon \mu^a(B_\varepsilon).$$

Thus, for $a \to 0$ $\mu^a(\bar{B}_\varepsilon) \to 0$. This implies that the invariant measure μ^a of the process (1) satisfies

$$\mu_a \xrightarrow{w} \delta_{x_0} \quad \text{as } a \to 0$$

where δ_{x_0} denotes the Dirac measure at the point x_0. Nontrivial limit laws can be obtained by a proper normalization of the process. We shall consider first the smooth unconstrained case.

2. The smooth unconstrained case. In addition to the Assumptions (A) we impose the following conditions.

Assumptions (B).

(i) There are no constraints, i.e. $S = \mathbb{R}^m$.

(ii) f is differentiable at x_0 with Jacobian matrix A, i.e.

$$f(x) = A \cdot (x - x_0) + R(x - x_0)$$

where $\|R(y)\| = o(\|y\|)$ as $y \to 0$.

(iii) $\delta_\varepsilon \geqq \delta > 0$. (For the definition of δ_ε see Assumption A(i).)

THEOREM 1. (Kushner and Hai Huang). *Let the Assumptions* (A) *and* (B) *be fulfilled. Let* $\tilde{\mu}^a$ *be the distribution of* $a^{-1/2}(X^a - x_0)$ *where* X^a *is distributed according to* μ^a, *the stationary distribution of* (1). *Then*

$$\tilde{\mu}^a \xrightarrow{w} N(0, V) \quad \text{as } a \to 0,$$

a normal distribution with covariance matrix V fulfilling the equation

$$AV + VA' = \Sigma \tag{9}$$

where Σ *is the covariance matrix of* $Z(x_0, \omega)$.

Remark. By the vec-operation which transforms matrices to vectors by putting the columns one below the other the equation (9) can be made explicit for V:

$$\operatorname{vec} V = (I_m \otimes A + A' \otimes I_m)^{-1} \cdot \operatorname{vec} \Sigma$$

where $\otimes$ denotes the Kronecker product and I_m is the $m \times m$ unit matrix. (C.f. Graham [2].) There is still another representation of the solution, namely

$$V = \int_0^\infty \exp(uA) \cdot \Sigma \exp(uA')\, du.$$

This matrix-integral formula was found by Walk [9] in a slightly different context.

Proof of the Theorem. We show first that the measures μ^a are uniformly tight. From the proof of Lemma 1 we know that

$$E(\|X_{n+1} - x_0\|^2) \leqq E(\|X_n - x_0\|^2)(1 + a^2 B_2 - 2a\delta) + a^2 A_2.$$

We set $b_n := E(\|X_n - x_0\|^2)$. From

$$b_{n+1} \leqq b_n(1 + a^2 B_2 - 2a\delta) + a^2 A_2$$

it follows that for $a < \delta B_2^{-1}$

$$\limsup b_n \leqq a^2 A_2 (2a\delta - a^2 B_2)^{-1} \tag{10}$$

which shows that the variance of the stationary distribution is finite and $O(a)$.

Let $U_n := a^{-1/2}(X_n^a - x_0)$ where X_n^a is a stationary sequence of the process (1). We have to show that the distribution of U_n tends to a normal distribution as $a \to 0$. The U_n fulfill the recursion

$$U_{n+1} = U_n - a \cdot A \cdot U_n + \sqrt{a}\, R(\sqrt{a}\, U_n) + \sqrt{a} \cdot Z_n$$

where $Z_n = Z(\sqrt{a}\, U_n + x_0, \omega)$.

Let φ be a three times differentiable function on $\mathbb{R}^m$ with compact support. By an application of Taylor's theorem it may be shown that there is a constant C_1 such that φ satisfies for every u, x, y

$$\begin{aligned} &|\varphi(u+x+y) - \varphi(u) - \nabla\varphi(u)(x+y) - \tfrac{1}{2}(x+y)'\nabla^2\varphi(u)(x+y)| \\ &\quad \leqq C_1(\|x\|^3 + \|y\| \cdot \|x\|^2 + \|y\|^2). \end{aligned}$$

Let K be a sufficiently large constant. We use the above formula for

$$\begin{aligned} u &= U_n, \\ x &= a \cdot A \cdot U_n + \sqrt{a} \cdot R(\sqrt{a}\, U_n) + \sqrt{a} \cdot Z_n 1_{\{\|Z_n\| \leqq K\}}, \\ y &= \sqrt{a}\, Z_n 1_{\{\|Z_n\| > K\}}, \end{aligned}$$

and get by taking the conditional expectation w.r.t. $\mathcal{F}_n = \sigma(X_1, \cdots, X_n) = \sigma(U_1, \cdots, U_n)$:

$$\begin{aligned} E(\varphi(U_{n+1})|\mathcal{F}_n) &= \varphi(U_n) - \nabla\varphi(U_n)(a \cdot A \cdot U_n + \sqrt{a}\, R(\sqrt{a}\, U_n)) \\ &\quad + \tfrac{1}{2}(a \cdot A \cdot U_n + \sqrt{a}\, R(\sqrt{a}\, U_n)'\nabla^2\varphi(U_n)(aAU_n + \sqrt{a}\, R(\sqrt{a}\, U_n)) \\ &\quad + \tfrac{1}{2}a \operatorname{tr}(\nabla^2\varphi(U_n) \cdot \Sigma_K) + O(\|(a \cdot A \cdot U_n + \sqrt{a}\, R(\sqrt{a}\, U_n)\|^3) + O(a^{3/2}) \\ &\quad + O(\|a \cdot A \cdot U_n + \sqrt{a}\, R(\sqrt{a}\, R(a\sqrt{a}\, U_n)\|^2 \sqrt{a} \cdot E(\|Z_n\| \cdot 1_{\{\|Z_n\| > K\}})) \\ &\quad + O(aE(\|Z_n\|^2 \cdot 1_{\{\|Z_n\| > K\}}), \end{aligned} \tag{11}$$

where $\Sigma_K = E(Z_n Z_n' 1_{\{\|Z_n\| \leqq K\}} | \mathcal{F}_n)$. Let $\varepsilon > 0$. By taking a sufficiently small we may achieve that $\|R(\sqrt{a} \cdot u)\| \leqq \varepsilon \cdot \sqrt{a} \cdot \|u\|$ for all u belonging to the (bounded) support of φ. Since U_n is stationary $E(\varphi(U_{n+1})) = E(\varphi(U_n))$ and we get by taking the expectation on both sides of (11)

$$|-E(\nabla\varphi(U_n) \cdot a \cdot A \cdot U_n) + \tfrac{1}{2}aE(\operatorname{tr}(\nabla^2\varphi(U_n) \cdot \Sigma_K))| \leqq \varepsilon \cdot O(a) + O(a^{3/2}).$$

Dividing by a, we get

$$|E(\nabla\varphi(U_n)A \cdot U_n) - \tfrac{1}{2}E(\operatorname{tr}(\nabla^2\varphi(U_n) \cdot \Sigma_K))| \leqq \varepsilon \cdot 0(1) + o(1)$$

as $a \to 0$. Since μ^a tends to δ_{x_0} as $a \to 0$ and K was arbitrary it follows that Σ_K can be made arbitrarily close to $\Sigma := E(Z(x_0, \cdot)Z'(x_0, \cdot))$ by choosing K large and a small. Because of the uniform tightness, $\tilde{\mu}^a$ has cluster points as a tends to zero. Let U be distributed according to such a cluster point $\tilde{\mu}^0$. Then

$$|E(\nabla\varphi(U) \cdot A \cdot U) - \tfrac{1}{2}E(\operatorname{tr}(\nabla^2\varphi(U)\Sigma))| \leqq \varepsilon \cdot O(1)$$

for every ε and for every test function φ. Hence

$$\int [\nabla\varphi(u) \cdot A \cdot u - \tfrac{1}{2}\operatorname{tr}(\nabla^2\varphi(u) \cdot \Sigma)]\, d\tilde{\mu}^0(u) = 0 \tag{12}$$

for every φ. Let $M(A, \Sigma)$ be the set of probability measures satisfying (12) for all test functions φ. We have to show that $M(A, \Sigma)$ contains only the $N(0, V)$-distribution.

First, $N(0, V)$ is a member of $M(A, \Sigma)$ which can be seen by direct calculation. By simple algebra one also can show that $\tilde{\mu}_1 \in M(A, \Sigma_1)$ and $\tilde{\mu}_2 \in M(A, \Sigma_2)$ imply that the convolution $\tilde{\mu}_1 * \tilde{\mu}_2$ is an element of $M(A, \Sigma_1 + \Sigma_2)$. Thus convoluting any element $\tilde{\mu}$ of $M(A, \Sigma)$ with a $N(0, \varepsilon\Sigma)$ distribution we get an element of $M(A, (1+\varepsilon)\Sigma)$ which has an infinitely often differentiable density and is arbitrarily close to $\tilde{\mu}$. Therefore it suffices to show that the only element of $M(A, \Sigma)$ with C^∞-density is the normal $N(0, V)$ distribution.

Let $\tilde{\mu} \in M(A, \Sigma)$ have a C^∞-density $f(u)$. Then (12) can be written as

$$\int [\nabla\varphi(u) \cdot A \cdot u - \tfrac{1}{2} \operatorname{tr} (\nabla^2 \varphi(u)\Sigma)] f(u)\, du = 0.$$

By partial integration on $\mathbb{R}^m$ this is equivalent to

$$\int \varphi(u) [\operatorname{tr} (J_{A \cdot u \cdot f(u)}) + \tfrac{1}{2} \operatorname{tr} (\nabla^2 f(u) \cdot \Sigma)]\, du = 0$$

where $J_{Auf(u)}$ is the Jacobian of $u \mapsto A \cdot u \cdot f(u)$. Hence $f(u)$ must fulfill the differential equation $\operatorname{tr} (J_{Auf(u)}) + \frac{1}{2} \operatorname{tr} (\nabla^2 f(u)\Sigma) \equiv 0$ which has the unique solution

$$f(u) = \text{const} \cdot \exp\left(-\frac{u' V^{-1} u}{2}\right)$$

with $AV + VA' = \Sigma$. Thus $\tilde{\mu}^0$ must be the $N(0, V)$ distribution.

3. The constrained case—degenerate solution. Let C_0 be the tangent cone of S at x_0. It may happen that the cone C_0 contains no proper linear subspace. (Then evidently x_0 lies on the boundary of S.) This situation is investigated in detail in this section. We need some further assumptions.

Assumptions (C).

(i) $$\inf_{x \in S} \frac{\langle x - x_0, f(x) \rangle}{\|x - x_0\| \, \|f(x)\|} \geqq \delta > 0,$$

(ii) $$\|f(x)\| \geqq c > 0 \text{ for } x \in S.$$

Remark that Assumption C(i) implies that C_0 contains no proper linear subspace. Before we state the main result, we shall prove a lemma which is of interest for itself.

LEMMA 4. *Let $P^a(x, A)$, $a \geqq 0$ be a family of Markov kernels. Suppose that*

(i) *$P^a(x, \cdot)$ converges weakly to $P^0(x, \cdot)$ i.e. in bounded Lipschitz metric as $a \to 0$, uniformly on compact sets of x,*

(ii) *$P^0(x, \cdot)$ is Feller, i.e. $x \mapsto P^0(x, \cdot)$ is continuous in the weak topology.*

Then $\mu^a \xrightarrow{w} \mu^0$ implies that $P^a\mu^a \xrightarrow{w} P^0\mu^0$. In particular all limits of sequences of invariant laws for $P^a(\cdot, \cdot)$ are invariant laws for $P^0(\cdot, \cdot)$.

Proof. We have to show that for every bounded continuous function g

$$\int g(y) P^a(x, dy)\, d\mu^a(x) \to \int g(y) P^0(x, dy)\, d\mu^0(x) \tag{13}$$

as $a \to 0$.

Since $P^0(x, \cdot)$ is Feller, $x \mapsto \int g(y) P^0(x, dy)$ is bounded and continuous and hence

$$\int g(y) P^0(x, dy)\, d\mu^a(x) \to \int g(y) P^0(x, dy)\, d\mu^0(x). \tag{14}$$

Due to Assumption (i)

$$\int g(y)P^a(x, dy) \to \int g(y)P^0(x, dy)$$

uniformly on compact sets. If $\mu^a \xrightarrow{w} \mu^0$ then $\{\mu^a, a \geqq 0\}$ is uniformly tight. Hence

$$\int g(y)P^a(x, dy)\, d\mu^a(x) - \int g(y)P^0(x, dy)\, d\mu^a \to 0 \tag{15}$$

by uniformity and tightness. Putting together (14) and (15) we get (13).

THEOREM 2. *Let Assumptions* (A) *and* (C) *be fulfilled. Let* $\tilde{\mu}^a$ *be the distribution of* $a^{-1}(X^a - x_0)$, *where* X^a *is distributed according to* μ^a, *the stationary distribution of* (1). *Then*

$$\tilde{\mu}^a \xrightarrow{w} \tilde{\mu}^0$$

where $\tilde{\mu}^0$ *is the stationary distribution of the process*

$$X_{n+1} = \pi_{C_0}(X_n - f(x_0) - Z(x_0, \cdot)).$$

Thus the limiting law may be found by solving an integral equation.

Proof. Let $W_n^a = a^{-1}(X_n^a - x_0)$, X_n^a being stationary. Let $C_a = \{a^{-1}(x - x_0) | x \in S\}$. Clearly C_a is a closed convex set, $C_a \subseteq C_0$ and $C_a \uparrow C_0$ as $a \to 0$. Moreover $\pi_{C_a}(x) = a^{-1}(\pi_S(ax + x_0) - x_0)$. Hence W_n^a fulfills the relation

$$W_{n+1}^a = \pi_{C_a}(W_n^a - f(x_0 + aW_n^a) - Z(x_0 + aW_n^a, \cdot)). \tag{16}$$

Let $P^a(x, A)$ be the Markov kernel belonging to (16). We have to show that the assumptions of Lemma 4 are fulfilled. By Assumption A(v) $Z(x_0 + au, \cdot)$ converges in L^2 to $Z(x_0, \cdot)$ uniformly for each bounded set of u's. Consequently, by eventually passing to a subsequence, $\pi_{C_a}(u - f(x_0 + au) - Z(x_0 + au, \cdot))$ converges a.e. to $\pi_{C_0}(u - f(x_0) - Z(x_0, \cdot))$ uniformly on bounded sets, which implies the Assumption (i) of Lemma 4. The continuity of $u \mapsto \pi_{C_0}(u - f(x_0) + Z(x_0, \cdot))$ entails the Feller property of $P^0(x, \cdot)$.

The assertion of Theorem 2 is shown if we prove the tightness of the set $\{\mu^a, a \geqq 0\}$. By C(i) and C(ii) using A(i) with $\varepsilon = 1$ we may find constants $\beta < \min(2\delta \cdot c, \delta_1)$, A_3 and B_3 such that

$$\begin{aligned}
E(\|W_{n+1}^a\|^2) &\leqq E(\|W_n^a\|^2) - E(\langle W_n^a, f(x_0 + aW_n^a)\rangle) + E(\|Z(x_0 + aW_n^a, \cdot)\|^2) \\
&\leqq E(\|W_n^a\|^2) - \beta E(\|W_n^a\| \cdot 1_{\{\|W_n^a\| < 1/a\}}) - 2\delta_1 a E(\|W_n^a\|^2 1_{\{\|W_n^a\| \geqq 1/a\}}) \\
&\quad + A_3 + a^2 B_3 E(\|W_n^a\|^2 1_{\{\|W_n^a\| \geqq 1/a\}}) \\
&\leqq E(\|W_n^a\|^2) - \beta E(\|W_n^a\|) - \delta_1 a E(\|W_n^a\|^2 1_{\{\|W_n^a\| \geqq 1/a\}}) \\
&\quad + A_3 + a^2 B_3 E(\|W_n^a\|^2 1_{\{\|W_n^a\| \geqq 1/a\}}).
\end{aligned}$$

If W_n^a is distributed according to the stationary distribution and a is sufficiently small then it follows that

$$E(\|W_n^a\|) \leqq A_3 \beta^{-1}.$$

This implies tightness.

4. The constrained, nondegenerate situation. In this section we assume that the minimal point x_0 lies on the boundary of S and the tangent cone C_0 contains a p-dimensional subspace ($1 \leqq p \leqq m - 1$). Under some suitable assumptions, the limiting law is then, as we shall prove, concentrated on this subspace.

To begin with we consider the following special case. Suppose that the set of constraints S is of the form

$$S = \bar{S} \times \mathbb{R}^p \tag{17}$$

where $\bar{S}$ is a q-dimensional convex set ($p+q=m$) containing no proper subspace. Then the projection π_S is of the special form

$$\pi_S(x) = \begin{cases} \pi_{\bar{S}}(\bar{x}) \\ \bar{\bar{x}} \end{cases}$$

where $\binom{\bar{x}}{\bar{\bar{x}}}$ is the decomposition of the vector x in a q- (resp. p-) dimensional part. Suppose that $\bar{C}_0$ is tangent cone to $\bar{S}$ at $\bar{x}_0$. According to the above partition the recursion may be written in the form

$$\begin{aligned} \bar{X}^a_{n+1} &= \pi_{\bar{S}}(\bar{X}^a_n - a\bar{f}(X^a_n) - a\bar{Z}_n), \\ \bar{\bar{X}}^a_{n+1} &= \bar{\bar{X}}^a_n - a\bar{\bar{f}}(X^a_n) - a\bar{\bar{Z}}_n. \end{aligned} \tag{18}$$

For this system we make the following assumptions, which are a combination of Assumptions (B) and (C).

Assumptions (D).

(i) $\|f(x)\| \leqq c > 0$ for $x \in S$.

(ii)
$$\begin{pmatrix} \bar{x}_0 \\ \bar{\bar{x}} \end{pmatrix} \mapsto \bar{\bar{f}}\begin{pmatrix} \bar{x}_0 \\ \bar{\bar{x}} \end{pmatrix}$$

is differentiable at $\bar{\bar{x}}_0$ with the Jacobian matrix A, i.e.

$$\bar{\bar{f}}\begin{pmatrix} \bar{x}_0 \\ \bar{\bar{x}} \end{pmatrix} = A \cdot (\bar{\bar{x}} - \bar{\bar{x}}_0) + R(\bar{\bar{x}} - \bar{\bar{x}}_0)$$

where $\|R(\bar{\bar{y}})\| = o(\|\bar{\bar{y}}\|)$ as $\bar{\bar{y}} \to 0$.

(iii) $\langle x - x_0, f(x)\rangle \geqq \delta(\|\bar{x} - \bar{x}_0\| + \|\bar{\bar{x}} - \bar{\bar{x}}_0\|^2)$ for a $\delta > 0$.

(iv) f fulfills the Lipschitz condition

$$\left| f\begin{pmatrix} x \\ y \end{pmatrix} - f\begin{pmatrix} x \\ z \end{pmatrix} \right| = o(\|y - z\|^{1/2})$$

uniformly in x, y, z.

THEOREM 3. *Let Assumptions* (A) *and* (D) *be fulfilled. Let* $\tilde{\mu}^a$ *be the distribution of* $a^{-1/2}(X^a - x_0)$ *where* X^a *is distributed according to* μ^a, *the stationary distribution of* (16).

Then

$$\tilde{\mu}^a \xrightarrow{w} N(0, \bar{\bar{V}}) \quad \text{as } a \to 0,$$

a normal distribution (*concentrated on* $\mathbb{R}^p$) *with covariance matrix*

$$V = \begin{pmatrix} 0 & 0 \\ 0 & \bar{\bar{V}} \end{pmatrix}$$

satisfying

$$A\bar{\bar{V}} + \bar{\bar{V}}A' = \bar{\bar{\Sigma}}$$

where $\bar{\bar{\Sigma}}$ *is the covariance matrix of* $\bar{\bar{Z}}(x_0, \cdot)$.

Proof. Let $U^a_n := a^{-1/2}(X^a_n - x_0)$. U^a_n may be partitioned into the two parts $\bar{U}^a_n$ and $\bar{\bar{U}}^a_n$. Exactly as in Theorems 1 and 2 it can be shown using D(iii) that the distribution of $\{a^{-1/2}\bar{U}^a_n, \bar{\bar{U}}^a_n\}$ is uniformly tight. Thus $\bar{U}^a_n$ converges in distribution to zero.

Moreover with high probability $\bar{U}_n^a$ is smaller than $a^{1/2}K$ for a constant K and by D(iv) $a^{-1/2}|f(x_0+\sqrt{a}\,U_n^a)-f(x_0+\sqrt{a}\,\bar{U}_n^a)|\to 0$ in probability. The recursion for $\bar{U}_n^a$ is

$$\bar{U}_{n+1}^a=\bar{U}_n^a+\sqrt{a}\,\bar{f}(x_0+\sqrt{a}\,U_n^a)+\sqrt{a}\,\bar{\bar{Z}}(x_0+\sqrt{a}\,U_n^a).$$

On a set of arbitrary high probability

$$\bar{U}_{n+1}^a=\bar{U}_n^a+\sqrt{a}\,\bar{f}(\bar{x}_0+\sqrt{a}\,\bar{U}_n^a)+o(a)+\sqrt{a}\,\bar{\bar{Z}}(x_0+\sqrt{a}\,\bar{U}_n^a)$$

and the rest of the proof is identical to that of Theorem 2. The more general case of a convex set S with tangent cone C_0 at x_0 containing a p-dimensional subspace can be reduced to the above situation by a nonlinear local parameter transformation. We remark that once the global conditions for convergence are met the asymptotic behavior depends exclusively on local properties.

Instead of giving long and complicated formulas we illustrate this case by two examples.

Example 1. Suppose that S is the m-dimensional unit ball and $x_0=(1,0,\cdots,0)$. We may introduce polar coordinates $(r,\varphi_1,\cdots,\varphi_{m-1})$ in $\mathbb{R}^m$ such that $x_0=(1,0,\cdots,0)$ also in this coordinate system. Then

$$S=\bar{S}\times(\mathbb{R}^{m-1}\cap B)$$

where $\bar{S}=\{x\in\mathbb{R}_1\,|\,|x|\leqq 1\}$ and B is the set of possible values for $\varphi_1,\cdots,\varphi_{m-1}$; i.e. an open set which contains the point zero. Also the gradient function may be written in polar coordinates form $f(r,\varphi_1,\cdots,\varphi_{m-1})$. If this function is differentiable at x_0 and the other conditions are met then there is a limiting normal distribution, which is concentrated on a $(m-1)$-dimensional subspace, which corresponds to the hyperplane tangent to S at x_0.

It is worth noticing that the asymptotic distribution strongly depends on the "curvature" of the set S in a neighborhood of x_0. In particular not only the differentiability of f, but also the differentiability of the transformation into polar coordinates is needed in the above example to establish asymptotic normality. If the convex set S cannot be transformed in a differentiably way into the form (17) then asymptotic normality is not more valid. Let us consider the following example.

Example 2. Suppose that $S=\{\binom{x}{y}\in\mathbb{R}^2\,|\,y\leqq|x|^\alpha\}$, $1<\alpha<2$ and $f\binom{x}{y}\equiv\binom{0}{1}$. Then $x_0=\binom{0}{0}$. The tangent cone is $C_0=\{\binom{x}{y}|y\geqq 0\}$ and contains a 1-dimensional subspace. There is a reparametrization such that S is of the form (17) in the new coordinates (ξ,η) namely

$$\xi=\operatorname{sgn}x(\sqrt{u^*+|u^*|^{2\alpha}}),$$
$$\eta=\operatorname{sgn}(y-|x|^\alpha)\sqrt{(x-u^*)^2+(y-|u^*|^\alpha)^2}$$

where u^* is the solution of $(x-u)^2+(y-|u|^\alpha)^2=\min!$ This solution is unique in a neighborhood of $\binom{0}{0}$ if $x\neq 0$. If $x=0$ set $\xi=0$ and $\eta=y$. S may be expressed in these new coordinates

$$S=\left\{\begin{pmatrix}\xi\\ \eta\end{pmatrix}\in\mathbb{R}^2\,\middle|\,\eta\geqq 0\right\}$$

but the mapping $(x,y)\mapsto(\xi,\eta)$ is not differentiable in both directions in a neighborhood of $\binom{0}{0}$, since $\eta(x)=0(x^{2/(2\alpha-1)})$.

In order to cover also such cases it is necessary to investigate unconstrained systems for which

$$f(x-x_0)=\|x-x_0\|^{\gamma-1}A\cdot(x-x_0)+R(x-x_0)$$

with $\|R(y)\| = o(\|y\|^{\gamma})$, $\gamma \leqq 1$. The smooth case is included by setting $\gamma = 1$. It may be shown that in such cases the suitable normalization in (4) is $g(a) = a^{1/(1+\gamma)}$ and the asymptotic distribution is of the Weibull type. But this is beyond the scope of this paper. A related result for the one-dimensional case was shown by the author in [7].

REFERENCES

[1] YU. ERMOLIEV, *Methods of Stochastic Programming*, Monographs in Optimization and Operations Research, Nauka, Moscow, 1976. (In Russian.)

[2] A. GRAHAM, *Kronecker Products and Matrix Calculus with Applications*, Ellis Horwood, Chichester, 1981.

[3] J. P. HIRIART-URRUTY, Thèse, Annales de l'Université de Clermont No. 58, 12ième fascicule, 1976.

[4] H. J. KUSHNER AND D. S. CLARK, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Applied Mathematical Sciences, Vol. 26, Springer-Verlag, New York, 1978.

[5] H. J. KUSHNER AND HAI HUANG, *Asymptotic properties of stochastic approximation with constant coefficients*, this Journal, 19 (1981), pp. 87–105.

[6] J. NEVEU, *Discrete Parameter Martingales*, North-Holland, Amsterdam, 1975.

[7] G. CH. PFLUG, *The Robbins-Monro procedure in nonstandard situations*, Math. Institute, Univ. of Gießen (submitted for publication).

[8] D. REVUZ, *Markov Chains*, North-Holland, Amsterdam, 1975.

[9] H. WALK, *An Invariance principle for the Robbins-Monro process in a Hilbert space.*, Z. Wahrsch. Verw. Geb., 39 (1977), pp. 135–150.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 4, July 1986

006

AN EXACT FORMULA FOR A LINEAR QUADRATIC ADAPTIVE STOCHASTIC OPTIMAL CONTROL LAW*

RAYMOND RISHEL†

Abstract. For a Linear Quadratic Adaptive Stochastic Optimal Control Problem a formula for the optimal control is obtained by applying variational methods to an equivalent problem obtained from the Girsanov transformation. The formula does depend on a quantity defined through a martingale representation of the conditional remaining cost.

Key words. adaptive control, dual control, optimal adaptive stochastic control

AMS(MOS) classifications. 93C40, 93E20

Introduction. Linear quadratic adaptive stochastic problems are important in engineering and have a long history. Some very early papers on adaptive control are [7] where they are called dual control problems to emphasize the dual aspects of learning about the system and controlling it at the same time. Examples of more recent work are [1], [12], [10], [2]. These problems can be considered as partially observed stochastic control problems. However, probably because of the lack of theory they have not generally been approached from this point of view. Instead methods such as those in [1], [12], [10], [2] are used. There have been fundamental recent advances in the theory of partially observed stochastic control, for instance [8], [9], [3], [4]. The purpose of this paper is to apply methods such as these to a linear quadratic adaptive stochastic control problem to investigate their applicability. In particular, although the methods of [4] do not apply exactly, this paper uses methods which follow the spirit of [4] to obtain an explicit formula for the optimal control law for this linear quadratic stochastic adaptive control problem. The methods are also somewhat analogous to those of [6] although again these do not apply directly.

Unfortunately the formula for the optimal control involves a quantity which must be determined from a martingale representation theorem, and thus is not obtained in a constructive manner. However, it is hoped that knowing the explicit form of the optimal control law will lead to new methods for computing it or approximating it.

The linear quadratic adaptive stochastic optimal control problem considered is the minimization of

$$E\left[\int_0^T [x(t)'Mx(t)+u(t)'Nu(t)]\,dt\right] \tag{1}$$

where $x(t)$ is a solution of the controlled stochastic differential equation

$$dx=[A(z)x+B(z)u]\,dt+dW \tag{2}$$

with initial condition

$$x(0)=x_0 \tag{3}$$

in which the matrices $A(z)$ and $B(z)$ depend on a vector z of unknown parameters with given prior probability density $P(z)$.

* Received by the editors May 9, 1984, and in revised form April 11, 1985.
† Department of Mathematics, University of Kentucky, Lexington, Kentucky 40506.

We shall show that necessary conditions that $u(t)$ be an optimal control based only on past measurements of x are that it have the form

$$u(t) = -\frac{N^{-1}}{2} \int B(z)'H(t, z)p(t, z)\, dz \tag{4}$$

where $p(t, z)$ is the conditional probability density of z given the past measurements of x up to time t and where $H(t, z)$ and $\theta(t, z)$ are uniquely determined as solutions of

$$d\theta(t, z) = -[x(t)'Mx(t) + u(t)'Nu(t)]\, dt + H(t, z)'\, dW \tag{5}$$

with terminal condition

$$\theta(T) \equiv 0. \tag{6}$$

An equivalent characterization of $H(t, z)$ is that it is uniquely determined from the representation

$$\begin{aligned} &\int_0^T [x(t)'Mx(t) + u(t)'Nu(t)]\, dt \\ &\quad = E\left\{\int_0^T [x(t)'Mx(t) + u(t)'Nu(t)]\, dt | x_0, z\right\} + \int_0^T H(t, z)'\, dW \end{aligned} \tag{7}$$

of the optimal criteria as a stochastic integral with respect to the Weiner process W.

The derivation begins by obtaining the linear quadratic control problem from an equivalent control problem through use of the Girsanov transformation; then variations are taken in the equivalent problem to obtain the optimality conditions.

1. Formulation. Consider formulating the adaptive control problem (1), (2), (3) through a change of probability measure. To do this, proceed as follows. Let $A(y)$ and $B(y)$ denote respectively $n \times n$ and $n \times m$ dimensional matrix valued Borel measurable functions of a q-dimensional vector variable y. Let $(\Omega, \mathcal{F}, P)$ denote a probability space Ω with σ-field of subsets $\mathcal{F}$ and probability measure P on Ω. Let $W(t)$ be an n-dimensional Weiner process defined on $(\Omega, \mathcal{F})$. Let x_0 and z denote n-dimensional and q-dimensional random vectors such that x_0, z, and $W(t)$ are mutually independent. Let z have probability density function $p(z)$. Define $x(t)$ by

$$x(t) = x_0 + W(t). \tag{8}$$

Define the σ-fields F_t and G_t by, F_t is the right continuous regularization of

$$\sigma[x(s);\ 0 \leqq s \leqq t] \tag{9}$$

and G_t is the right continuous regularization of

$$\sigma[z, x(s);\ 0 \leqq s \leqq t]. \tag{10}$$

Define an admissible control $u(t)$ to be an m-dimensional vector valued stochastic process satisfying I and II.

I. $u(t)$ is F_t adapted.

II. $E[\exp(\frac{1}{2}\int_0^T \|A(z)x(t) + B(z)u(t)\|^2\, dt)] < \infty$.

Define $q^u(t, z)$ by

$$\begin{aligned} q^u(t, z) \triangleq \exp\Bigg[&\int_0^t [A(z)x(s) + B(z)u(s)]'\, dW(s) \\ &-\frac{1}{2}\int_0^t \|A(z)x(s) + B(z)u(s)\|^2\, ds\Bigg]. \end{aligned} \tag{11}$$

It follows from II that

$$P\left(\int_0^T \|A(z)x(t)+B(z)u(t)\|^2\,dt<\infty\right)=1$$

and thus $q^u(t, z)$ is well defined.

Assumption II implies, [11, Thm. 6.1], that $q^u(t, z)$ is a G_t martingale and that

$$E[q^u(T, z)]=1. \tag{12}$$

If p^u is the probability measure whose Radon-Nikodym derivative with respect to P is $q^u(T, z)$, that is if

$$P^u=q^u(T, z)P, \tag{13}$$

then (12) asserts P^u is a probability measure. The Girsanov theorem, [11, Thm. 6.3], implies if $W^u(t)$ is defined by

$$W^u(t)=W(t)-\int_0^t [A(z)x(s)+B(z)u(s)]\,ds, \tag{14}$$

that $W^u(t)$ is a (P^u, G_t) Weiner process. Combining (8) and (14) gives

$$x(t)=x_0+\int_0^t [A(z)x(s)+B(z)u(s)]\,ds+W^u(t). \tag{15}$$

Denote taking expectations with respect to P^u by E^u and with respect to P by E. Thus

$$\begin{aligned} &E^u\left\{\int_0^T [x(t)'Mx(t)+u(t)'Nu(t)]\,dt\right\} \\ &\quad=E\left\{q^u(T, z)\int_0^T [x(t)'Mx(t)+u(t)'Nu(t)]\,dt\right\}. \end{aligned} \tag{16}$$

Thus the problem of minimizing the left-hand side of (16) subject to (15) holding over the class of admissible controls satisfying I and II is equivalent to the problem of minimizing the right side of (16), subject to (8) and (11), over the class of admissible controls satisfying I and II. Notice that Assumption I asserts the control $u(t)$ is F_t-measurable for each t, so that $u(t)$ depends only on the past of $x(t)$ and we can consider z as a vector of parameters unknown to the controller.

Define $J(u)$ by

$$J(u)=E\left\{q^u(T, z)\int_0^T [x(t)'Mx(t)+u(t)'Nu(t)]\,dt\right\}. \tag{17}$$

Motivated by these considerations we shall consider the problem of determining the optimal control for the problem of minimizing $J(u)$ subject to (8) and (11) over the class of controls satisfying I and II.

2. Variations. If $u(t)$ is an optimal control and $v(t)$ is a stochastic process which is bounded and F_t is adapted, then it can be shown that $u(t)+\varepsilon v(t)$ satisfies I and II for $-1\leqq\varepsilon\leqq 1$. Thus if

$$\delta J(u, v)\triangleq\frac{d}{dt}J(u+\varepsilon v)\Big|_{\varepsilon=0} \tag{18}$$

exists, then

$$\delta J(u, v)=0 \tag{19}$$

is a necessary condition for optimality.

LEMMA 1. *$\delta J(u, v)$ exists and is given by*

$$\delta J(u, v) = E\left\{ q^u(T, z)\left[\int_0^T (B(z)v(t))'\, dW^u \int_0^T (x(t)'Mx(t)+u(t)'Nu(t))\, dt + \int_0^T 2v(t)'Nu(t)\, dt\right]\right\}. \tag{20}$$

Proof. Since $v(t)$ is bounded it can be shown that

$$\left|q^{u+\varepsilon v}(T, z)\int_0^T [x(t)'Mx(t)+(u(t)+\varepsilon v(t))'N(u(t)+\varepsilon v(t))]\, dt - q^u(T, z)\int_0^T [x(t)'Mx(t)+u(t)'Nu(t)]\, dt\right| \leqq \varepsilon b, \tag{21}$$

where

$$E[b] < \infty. \tag{22}$$

This and Lebesgue's dominated convergence theorem justify computing (18) by differentiating with respect to ε under the expectation sign in

$$J(u+\varepsilon v) = E\left\{ q^{u+\varepsilon v}(T, z)\int_0^T [x(t)'Mx(t)+(u(t)+\varepsilon v(t))'N(u(t)+\varepsilon v(t))]\, dt\right\}. \tag{23}$$

Since

$$q^{u+\varepsilon v}(T, z) = \exp\left[\int_0^T [A(z)x(s)+B(z)(u(s)+\varepsilon v(s))]\, dW - \frac{1}{2}\int_0^T \|A(z)x(s)+B(z)(u(s)+\varepsilon v(s))\|^2\, ds\right], \tag{24}$$

$$\begin{aligned}\frac{d}{d\varepsilon} q^{u+\varepsilon v}(T, z)\Big|_{\varepsilon=0} &= q^u(T, z)\left[\int_0^T [B(z)v(s)]'\, dW - \int_0^T [B(z)v(s)]'[A(z)x(s)+B(z)u(s)]\, ds\right] \\ &= q^u(T, z)\int_0^T [B(z)v(s)]'\, dW^u.\end{aligned} \tag{25}$$

Thus (20) follows by differentiating under the expectation sign in (23) using (25) and the product rule for differentiation.

Remark. Let $\beta[0, T]$ and $\beta(E^q)$ denote the Borel fields on the respective spaces $[0, T]$ and E^q. Since $G_t = \sigma(z) \times F_t$ standard measure theoretic arguments imply for any G_t adapted process $\phi(t)$ there will be a function $\tilde{\phi}(t, y, \omega)$ which is $\beta[0, T] \times B(E^q) \times \mathcal{F}$ measurable and F_t measurable for each (t, y) such that

$$\phi(t)(\omega) = \tilde{\phi}(t, z(\omega), \omega). \tag{26}$$

Motivated by this we shall use the notation $\phi(t, z)$ to denote a G_t adapted process and interpret this to mean that

$$\phi(t, z)(\omega) = \tilde{\phi}(t, z(\omega), \omega). \tag{27}$$

THEOREM 1. *If R is a G_T measurable random variable for which*

$$E^u[R^2] < \infty, \tag{28}$$

there is a unique n-dimensional vector-valued G_t adapted process $\phi(t, z)$ for which

$$E^u\left[\int_0^t \|\phi(t, z)\|^2\, dt\right] < \infty \tag{29}$$

and

$$E^u[R|G_t] = E^u[R|z, x_0] + \int_0^t \phi(s, z)'\, dW^u. \tag{30}$$

Proof. By the representation theorem for square integrable martingales, [11, Thm. 5.4], the square integrable martingale $E^u[R|G_t]$ has the representation

$$E^u[R|G_t] = \int_0^t \phi(s, z)'\, dW^u + \mathcal{M}(t) \tag{31}$$

where $\mathcal{M}(t)$ is a martingale orthogonal to the stochastic integrals and $\phi(t, z)$ is a G_t adapted process for which

$$E^u\left[\int_0^T \|\phi(t, z)\|^2\, dt\right] < \infty. \tag{32}$$

It will follow that

$$\mathcal{M}(t) = \mathcal{M}(0) \tag{33}$$

if it can be shown that

$$E\left[(\mathcal{M}(t) - \mathcal{M}(0)) f(z) g(x_0) \prod_{j=1}^{n} F_j(W_{t_j}^u)\right] = 0 \tag{34}$$

for any bounded Borel measurable functions $f(y)$, $g(x)$, $F_j(x)$ and any t_j, $0 \leqq t_1 \leqq \cdots \leqq t_n \leqq t$. This can be done by repeating the proof of [11, Thm. 5.5] with minor changes. The crucial property needed is that if $t > s$ then the increment $W^u(t) - W^u(s)$ is independent of G_s. The theorem follows from (33) since

$$E^u[R|G_0] = E^u[R|z, x_0] = \mathcal{M}(0). \tag{35}$$

The uniqueness of $\phi(t, z)$ will follow exactly as the uniqueness argument in [11, Thm. 5.7, p. 170].

THEOREM 2. *If*

$$E^u\left[\left(\int_0^T [x(t)'Mx(t) + u(t)'Nu(t)]\, dt\right)^2\right] < \infty \tag{36}$$

and $H(t, z)$ is the G_t adapted process in the representation

$$\begin{aligned}\int_0^T [x(t)'Mx(t) + u(t)'Nu(t)]\, dt \\ = E^u\left[\int_0^T [x(t)'Mx(t) + u(t)'Nu(t)]\, dt | z, x_0\right] + \int_0^T H(t, z)'\, dW^u\end{aligned} \tag{37}$$

for which

$$E\left[\int_0^T \|H(t, z)\|^2\, dt\right] < \infty, \tag{38}$$

then $\delta J(u, v)$ is given by

$$\delta J(u, v) = E\left\{q^u(T, z) \int_0^T v(t)'[B(z)'H(t, z) - 2Nu(t)]\, dt\right\}. \tag{39}$$

Proof. Using (37), rules for expected values of products of stochastic integrals, and the orthogonality of stochastic integrals with respect to W^u and G_0 measurable functions gives

$$E\left\{q^u(T,z)\int_0^T [B(z)v(t)]'\,dW^u \int_0^T (x(t)'Mx(t)+u(t)'Nu(t))\,dt\right\}$$
$$= E\left\{q^u(T,z)\int_0^T v(t)'B(z)'H(t,z)\,dt\right\}. \tag{40}$$

Thus the theorem follows from (20) and (40).

Since we must have $\delta J(u,v)=0$ for every bounded F_t adapted process $v(t)$, a standard argument using (39) implies

$$E\{q^u(T,z)[B(z)'H(t,z)-2Nu(t)]|F_t\}=0 \tag{41}$$

for almost every t. Thus we have from (41) that a necessary condition for $u(t)$ to be an optimal control is that

$$u(t)=-\frac{1}{2}N^{-1}\frac{E[q^u(T,z)B(z)'H(t,z)|F_t]}{E[q^u(T,z)|F_t]} \tag{42}$$

holds for almost every t. Since $q^u(t,z)$ is a G_t martingale using conditioning first with respect to G_t and then with respect to F_t and using the law of iterated conditional expectations gives

$$u(t)=-\frac{1}{2}N^{-1}\frac{E[q^u(t,z)B(z)'H(t,z)|F_t]}{E[q^u(t,z)|F_t]}. \tag{43}$$

Since $G_t=\sigma(z)\times F_t$ and $\sigma(zr$ and F_t are independent σ-fields under P, taking the conditional expectation of a G_t measurable random variable given F_t under the measure P can be expressed by integrating the random variable with respect to the prior probability density $p(z)$ of z. Thus (43) can be rewritten as

$$u(t)=-\frac{1}{2}N^{-1}\frac{\int q^u(t,z)B(z)'H(t,z)p(z)\,dz}{\int q^u(t,z)p(z)\,dz}. \tag{44}$$

Similarly the conditional density of z given F_t under the measure P^u is given by

$$p^u(t,z)\triangleq\frac{q^u(t,z)p(z)}{\int q^u(t,z)p(z)\,dz}. \tag{45}$$

Thus (44) can also be expressed as

$$u(t)=-\frac{1}{2}N^{-1}\int p^u(t,z)B(z)'H(t,z)\,dz$$

where $p^u(t,z)$ is given by (45).

The quantity $H(t,z)$ can also be determined from an adjoint equation. To see this define $\theta(t,z)$ by

$$\theta(t,z)=-\int_0^t [x(s)'Mx(s)+u(s)'Nu(s)]\,ds$$
$$+E\left[\int_0^T [x(s)'Mx(s)+u(s)'Nu(s)]\,ds|x_0,z\right]+\int_0^t H(s,z)'\,dW^u. \tag{46}$$

Then from (37)

$$\theta(T, z) \equiv 0, \tag{47}$$

and (46) implies

$$d\theta(t, z) = -[x(t)'Mx(t) + u(t)'Nu(t)]\, dt + H(t, z)'\, dW^u \tag{48}$$

which are the desired adjoint equation and terminal boundary condition.

Solutions of (48) and (47) for which

$$E\left[\int_0^T \|H(t, z)\|^2\, dt\right] < \infty \tag{49}$$

holds are unique. To see this let $\tilde{\theta}(t, z)$ and $\tilde{H}(t, z)$ satisfy (48), (47) and (49). Then integrating (48) from 0 to T gives

$$\int_0^T [x(t)'Mx(t) + u(t)'Nu(t)]\, dt = \tilde{\theta}(0, z) + \int_0^T \tilde{H}(t, z)'\, dW^u. \tag{50}$$

Since the increments of W^u are independent of G_0 and (49) holds for $\tilde{H}(t, z)$, the conditional expectation of the stochastic integral given G_0 is zero and (50) implies

$$E\left[\int_0^T x(t)'Mx(t) + u(t)'Nu(t)]\, dt \,\middle|\, z, x_0\right] = \tilde{\theta}(0, z). \tag{51}$$

The uniqueness of the representation (37), (50) and (51) implies

$$\tilde{H}(t, z) = H(t, z). \tag{52}$$

Then $\tilde{\theta}(t, z) = \theta(t, z)$ follows by integrating (48) from 0 to t, using (51) and (52).

Collecting these results into a theorem, we have

THEOREM 3. *If $u(t)$ is an optimal control for the problem of minimizing*

$$E^u\left[\int_0^T [x(t)'Mx(t) + u(t)'Nu(t)]\, dt\right] \tag{53}$$

subject to $x(t)$ being a solution of

$$dx = [A(z)x + B(z)u]\, dt + dW^u, \qquad x(0) = x_0, \tag{54}$$

and

$$E^u\left[\left(\int_0^T [x(t)'Mx(t) + u(t)'Nu(t)]\, dt\right)^2\right] < \infty \tag{55}$$

is satisfied for the optimal control, then $u(t)$ must have the form

$$u(t) = -\frac{1}{2} N^{-1} \int p^u(t, z) B(z)' H(t, z)\, dz \tag{56}$$

where $\theta(t, z)$ and $H(t, z)$ are solutions of

$$d\theta(t, z) = -[x(t)'Mx(t) + u(t)'Nu(t)]\, dt + H(t, z)'\, dW^u, \qquad \theta(T, z) = 0 \tag{57}$$

where

$$p^u(t, z) = \frac{q^u(t, z)p(z)}{\int q^u(t, z)p(z)\, ds} \tag{58}$$

and

$$q^u(t,z)=\exp\left[\int_0^t [A(z)x(s)+B(z)u(s)]'\,dW^u - \frac{1}{2}\int_0^t \|A(z)x(s)+B(z)u(s)\|^2\,ds\right]. \tag{59}$$

3. Conclusions. Theorem 3 gives a stochastic two-point boundary value problem involving equations (54) and (56)-(59) with (54) and (57) being the primary equations of the two-point boundary value problem. *If* this could be solved to determine $H(t,z)$, then (56) would be an explicit representation of the optimal control.

It is natural to ask, "Is there a Riccati equation through which a solution of (54) and (56)-(59) can be expressed?" The existence of a Riccati equation solution in a corresponding linear quadratic control problem in [5] and the results derived there make this appear hopeful. A number of guesses can be made for the form of the solution; however, whether there is a solution of Riccati form is still an open question.

It is also natural to question whether an algorithm for approximating the optimal control could be based on equations (54) and (56)-(59) of Theorem 3. This is also an open question.

REFERENCES

[1] K. J. ÅSTRÖM AND B. WITTENMARK, *On self-tuning regulators*, Automatica, 9 (1973), p. 183.

[2] A. BECKER, P. R. KUMAR AND C. Z. WEI, *Adaptive control with the stochastic approximation algorithm: geometry and convergence*, Univ. Maryland, Baltimore County Report 83-8, 1983.

[3] V. E. BENEŠ AND I. KARATZAS, *On the relations of Zakai's and Mortensen's equations*, this Journal, to appear.

[4] A. BENSOUSSAN, *Maximum principle and dynamic programming approaches of the optimal control of partially observed diffusions*, INRIA Report.

[5] J.-M. BISMUT, *Linear quadratic optimal stochastic control with random coefficients*, this Journal, 14 (1976), pp. 419-444.

[6] R. J. ELLIOTT, *The optimal control of a stochastic system*, this Journal, 15 (1977), pp. 756-778.

[7] A. A. FELDBAUM, *Dual control theory* I-IV, Automation and Remote Control, Vol. 21 (1960), pp. 874-1033 and Vol. 22 (1961), pp. 1-109.

[8] W. H. FLEMING, *Nonlinear semigroup for controlled partially observed diffusions*, this Journal, 20 (1982), pp. 286-301.

[9] W. H. FLEMING AND E. PARDOUX, *Optimal control for partially observed diffusions*, this Journal, 20 (1982), pp. 261-285.

[10] G. GOODWIN, P. RAMADGE AND P. CAINES, *Discrete time stochastic control*, this Journal, 19 (1981), pp. 829-853.

[11] R. S. LIPTSER AND A. N. SHIRYAYEV, *Statistics of Random Processes* I, *General Theory*, Springer-Verlag, 1977.

[12] L. LJUNG, *Analysis of recursive stochastic algorithms*, IEEE Trans. Automat. Control, AC-22 (1977), pp. 551-575.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 4, July 1986

007

SUFFICIENT OPTIMALITY CONDITIONS FOR STRATIFIED CONTROL PROBLEMS*

STEFAN MIRICĂ†

Abstract. Sufficient optimality conditions of dynamic programming type avoiding the axioms of Boltyanskii's "regular synthesis" for control problems that have weakly stratified Hamiltonians are proved. An improved version of Boltyanskii's "fundamental lemma" is applied to the "value function" defined as the minimum of the cost functional along the solutions of a Hamiltonian inclusion which plays the role of a "system of characteristics" for the Hamilton-Jacobi-Bellman equation of dynamic programming.

Key words. optimal control problems, dynamic programming, sufficient optimality conditions, stratified sets and mappings, Hamiltonian systems

AMS(MOS) subject classification. 49C20

1. Introduction. The aim of this paper is to prove verifiable sufficient optimality conditions of dynamic programming type that avoid the axioms of Boltyanskii's regular synthesis ([1], [5]-[7], [17], [24]) for optimal control problems whose Hamiltonians are stratified functions in a very weak sense. As one may easily verify, a large number of classes of optimal control problems in the literature ([1], [7], [10], etc.) admit a stratified structure (as, for instance, the problems defined by subanalytic sets and mappings, [5], [6], [12], [22], [23]) so the approach based on the theory of stratified sets and mappings is by no means very restrictive.

The proofs of the main results in § 4 rely essentially on the preliminary results in § 2 concerning the behaviour of regular and absolutely continuous mappings with respect to stratified sets and mappings.

In § 3 we consider autonomous optimal control problems of Mayer type that have weakly stratified Hamiltonians and we introduce a generalized Hamiltonian system which is strongly connected with Pontryagin's maximum principle as well as with the theory of Hamiltonian vector fields on symplectic manifolds (e.g. [11]) and plays the role of a "system of characteristics" for the first order partial differential equation of dynamic programming often called the "Hamilton-Jacobi-Bellman equation". Under some additional hypotheses we prove that the solutions of the Hamiltonian inclusion (which we call *characteristics*) are extremals of the optimal control problem, i.e. their projections on the phase space are admissible trajectories satisfying maximum principle.

In § 4 we prove first a stronger version of Boltyanskii's "fundamental lemma" ([1], [5]-[7], [17], etc.); then we prove several theorems giving sufficient optimality conditions in terms of the "value function" defined as the minimum of the cost functional along the characteristics, separately, in the case of regulated (in particular, piecewise continuous) admissible controls and in the case of measurable bounded controls. We note that in the case of regulated admissible controls the Hamiltonian and the value function should be weakly C^1-stratified and continuous and the characteristics should be regular with respect to the stratification of the Hamiltonian while in the case of measurable bounded controls the Hamiltonian and the value function are required to be weakly C^1-stratified and locally Lipschitzian but the characteristics need only be absolutely continuous. The last theorem replaces a certain assumption on the value function with the easier verifiable property that the characteristics may be embedded in a family of "stratified flows".

* Received by the editors May 3, 1984, and in revised form January 24, 1985.
† Faculty of Mathematics, University of Bucharest, Academiei 14, 70109 Bucharest, Romania.

In § 5 we consider the well-known problem of "moon landing" ([7], [10], [16], etc.) in order to illustrate the way the above mentioned results may be used and their possible advantages over the usual approach involving necessary optimality conditions, uniqueness of the extremals and existence of optimal controls ([1], [7], [10], etc.). It is apparent that when applicable, the results in this paper may greatly simplify the arguments proving that the result of the computations (more or less the same in any approach) is indeed the optimal feedback control. Moreover, the intricate process of finding the extremals of the problem by the simultaneous operations of maximizing the Pontryagin function and integrating the adjoint system is replaced in this approach by the more systematized (if not easier) operation of finding the solutions of a differential inclusion which often turns out to be a piecewise smooth differential system (for which, it is conceivable that one may combine theoretical arguments with computer results).

The autonomous optimal control problems of Mayer type were preferred for the sake of simplicity of notations but it is obvious that similar results may be proved for general (nonautonomous) Bolza or Lagrange problems (which can always be written as autonomous Mayer problems, [7], [10]). Moreover, since one looks for optimal trajectories rather than for optimal controls, our approach is particularly suited for control problems defined by differential inclusions ([2], [8]). As far as the sufficient optimality conditions are concerned, the results in this paper may also be extended to optimal control problems in which the control space depends on the phase space variable.

It should be noted that the results in this paper considerably extend and improve the results in [19], at the same time correcting some errors in the statements and the proofs of Lemmas 2.2, 2.3 and 2.5 in [19] (which, however, do not affect the validity of the main results) which should be replaced by Lemmas 2.5 and 2.6 in this paper.

2. The behaviour of regular and absolutely continuous mappings with respect to stratified sets and functions. We shall use the following notations: Z—the set of integers, N—the set of positive integers, R^n—the real n-dimensional Euclidean space, $\langle \cdot, \cdot \rangle$ the scalar product in such a space; we identify a vector $x \in R^n$ with the corresponding linear functional $y \mapsto \langle x, y \rangle$ in $L(R^n, R)$. If X is a set, then $\mathcal{P}(X)$ ($\mathcal{P}_0(X)$) denotes the family of its (respectively, nonempty) subsets; if $X \subset R^n$, then $\mathrm{Cl}(X)$ denotes its closure in the usual topology; I, $[a, b]$, $[a, b) \subset R$ denote intervals of the real line; $[\![a, b]$ denotes one of the intervals $[a, b]$, $(a, b]$.

DEFINITION 2.1. A nonempty subset $X \subset R^n$ is said to be *weakly C^k-stratified* for some $k \in \{1, 2, \cdots \infty, \omega\}$ if it admits a locally finite partition $\mathcal{S}$ (i.e. any compact subset of X intersects only a finite number of members of $\mathcal{S}$) into connected regular submanifolds of R^n called strata.

We recall that X is said to be *C^k-stratified* ([5], [6], [12], [23], etc.) if $\mathcal{S}$ has the following additional property: for any $S_1, S_2 \in \mathcal{S}$ for which $S_1 \neq S_2$, $S_1 \cap \mathrm{Cl}(S_2) \neq \varnothing$ one has: $S_1 \subset \mathrm{Cl}(S_2)$ and $\dim(S_1) < \dim(S_2)$; we note that this property as well as the so called "Whitney property" of a stratification [5] are not needed in the proofs in this paper so we shall use only weakly stratified sets.

A weak stratification $\mathcal{S}$ of X is said to be *compatible* with a family $\mathcal{A} \subset \mathcal{P}(R^n)$ if any $A \in \mathcal{A}$ is either disjoint of any stratum of $\mathcal{S}$ or is a union of strata.

If $S \subset R^n$ is a submanifold of class C^k and $x \in S$, then T_xS denotes the *tangent space of S at x* and it will always be considered as a subspace of $R^n \cong T_xR^n$. We recall (e.g. [15]) that $\bar{x} \in T_xS$ iff there exists a C^k-mapping $c(\cdot) : (-a, a) \to S$, $a > 0$ such that: $c(0) = x$ and $c'(0) = \bar{x}$.

If $\mathcal{S}$ is a weak stratification of X, then one may define at each point $x \in X$ a *tangent space of X that corresponds to the* (weak) *stratification* $\mathcal{S}$ as follows: $T_x X = T_x S$ if $x \in S \in \mathcal{S}$; if the stratification $\mathcal{S}_1$ is compatible with the stratification $\mathcal{S}_2$ of the same subset $X \subset R^n$ and if for $x \in X$, $T_x^i X$ denotes the tangent space with respect to $\mathcal{S}_i$, $i = 1, 2$, then obviously: $T_x^1 X \subset T_x^2 X$.

DEFINITION 2.2. A mapping $f(\cdot): X \subset R^n \to R^m$ is said to be *weakly C^k-stratified* if there exists a weak C^k-stratification $\mathcal{S}_f$ of X such that for any $S \in \mathcal{S}_f$ the restriction $f_S(\cdot) = f(\cdot)|S$ is of class C^k; $f(\cdot)$ is said to be a *weak C^k-stratified submersion* if it has the following additional property; for any $S \in \mathcal{S}_f$ $f(S) \subset R^m$ is a regular submanifold and the restriction $f_S(\cdot): S \to f(S)$ is a submersion of class C^k (i.e. $\operatorname{rank}(Df_S(x)) = \dim(f(S))$ for any $x \in S$); $f(\cdot)$ is said to be stratified by $\mathcal{S}_f$.

We recall that $f(\cdot): X \subset R^n \to Y \subset R^m$ is said to be *C^k-stratified* ([5], [6], [22], etc.) if there exist a C^k-stratification $\mathcal{S}_X$ of X and a C^k-stratification $\mathcal{S}_Y$ of Y such that for any $S \in \mathcal{S}_X$ one has: $f(S) \in \mathcal{S}_Y$ and $f_S(\cdot): S \to f(S)$ is a submersion of class C^k.

We note that while any mapping $f(\cdot): X \subset R^n \to R^m$, of class C^1 on an open subset $X \subset R^n$ is weakly C^1-stratified, there obviously exist C^∞ mappings that are not C^1-stratified in the sense above. In this paper, with the only exception of the first component of the "stratified flow" in Definition 4.5 (which should be a weak C^1-stratified submersion) the functions involved need be only weakly C^1-stratified in the sense of Definition 2.2.

If $f(\cdot): X \subset R^n \to R^m$ is weakly C^1-stratified by the weak C^1-stratification $\mathcal{S}_f$ of X, then we define the *derivative of $f(\cdot)$ with respect to $\mathcal{S}_f$* as follows: $Df(x) = Df_S(x) \in L(T_x S, R^m)$ if $x \in S \in \mathcal{S}_f$. We note that if $\dim(S) = n$ (hence $S \subset R^n$ is open) then for any $x \in S$, $Df(x) \in L(R^n, R^m)$ is the usual (Fréchet) derivative of $f(\cdot)$ at x but if $\dim(S) < n$ then $f(\cdot)$ may be even discontinuous at the points in S; if $\dim(S) = 0$, hence $S = \{x\}$ is a singleton, then we take $T_x S = \{0\}$ and $Df(x) = 0$.

DEFINITION 2.3. An absolutely continuous mapping $x(\cdot): I \subset R \to R^n$ is said to be *regular* if its derivative, $x'(\cdot)$, is regulated ([3], § II.1) i.e. it has one-sided limits at each point in I and therefore it has a countable set of discontinuities, all of the first kind (at such points $x'(\cdot)$ may not even exist).

The mapping $x(\cdot)$ is said to be regular with respect to the weak stratification $\mathcal{S}$ of $X \subset R^n$ if there exists a countable partition $\{I_k; k \in N\}$ of I into subintervals such that for any $k \in N$ there exists $S_k \in \mathcal{S}$ such that $x(I_k) \subset S_k$, the restriction mapp $x_k(\cdot) = x(\cdot)|I_k$ is of class C^1 and its derivative has one-sided limits at the endpoints of the interval I_k.

We denote by $J_{x(\cdot)}$ the set of endpoints of the intervals I_k, $k \in N$, and we call them *switching points* of $x(\cdot)$; if $J_{x(\cdot)}$ is finite then $x(\cdot)$ is said to be *finitely regular* or *piecewise smooth* and its derivative is said to be *piecewise continuous.*

In what follows we shall use the following particular case of [4, Chap. XI, Corollary 4.2] (which is also a generalization of [3, § I.2, Prop. 2]):

LEMMA 2.4. *Let $f(\cdot): [a, b] \subset R \to R$ be continuous and such that there exists a countable subset $J_f \subset [a, b]$ with the following property: for any $t \in [a, b) \setminus J_f$ there exists a sequence $\{t_k\} \subset [a, b]$ such that $t_k \searrow t$* (i.e. $t_k > t$ for any $k \in N$) *and:*

$$\lim_{k \to \infty} \frac{f(t_k) - f(t)}{t_k - t} \geqq 0. \tag{2.1}$$

Then $f(b) \geqq f(a)$.

LEMMA 2.5. *Let $X \subset R^n$ be weakly C^1-stratified by $\mathcal{S}$, let $x(\cdot): [a, b] \subset R \to X$ be absolutely continuous and let $J_{x(\cdot)}$ be the set of points $t \in [a, b]$ at which either $x(\cdot)$ is*

not differentiable or t is isolated to the right in $x^{-1}(S)$ (i.e. there exists $r>0$ such that $(t, t+r)\cap x^{-1}(S)=\emptyset$) if $x(t)\in S\in\mathcal{S}$.

Then $J_{x(\cdot)}$ has zero Lebesgue measure and:

$$x'(t)\in T_{x(t)}X \quad \text{for any } t\in[a, b]\backslash J_{x(\cdot)}. \tag{2.2}$$

Moreover, if $x(\cdot)$ is regular, then $J_{x(\cdot)}\subset[a, b]$ is countable.

Proof. From its definition it follows that the set $J_{x(\cdot)}$ is the union $J_d\cup J_r$ where J_d is the null set of points $t\in[a, b]$ at which $x(\cdot)$ is not differentiable and J_r is the set of points $t\in[a, b]$ that are isolated to the right in $x^{-1}(S)\subset[a, b]$ if $x(t)\in S\in\mathcal{S}$. According to a known result (e.g. Problem $K(d)$ in [14, Chap. I]) the set of points that are isolated to the right in $x^{-1}(S)$ is countable for any $S\in\mathcal{S}$ and, on the other hand, since $\mathcal{S}$ is locally finite, the compact subset $x([a, b])\subset X$ intersects only a finite number of strata. It follows that the set J_r is countable; hence $J_{x(\cdot)}$ is a null set and if $x(\cdot)$ is regular, then $J_{x(\cdot)}=J_d\cup J_r$ is countable as J_d is also countable.

Further on, if $t\in[a, b]\backslash J_{x(\cdot)}$ then $x'(t)=\lim_{s\to t}(x(s)-x(t))/(s-t)$ exists and if $x(t)\in S\in\mathcal{S}$ then (since $t\notin J_r$) there exists a sequence $t_k\searrow t$ such that $x(t_k)\in S$ for any $k\in N$ and therefore $x'(t)=\lim_{k\to\infty}(x(t_k)-x(t))/(t_k-t)\in T_{x(t)}S=T_{x(t)}X$ as one may easily see taking a local coordinate chart with the submanifold property at $x(t)\in S$ (e.g. [15]).

LEMMA 2.6. *Let $f(\cdot): X\subset R^n\to R$ be continuous and weakly C^1-stratified, let $x(\cdot):[a, b]\subset R\to X$ be absolutely continuous and let $J_{x(\cdot)}$ be the set (defined in Lemma 2.5) of points in $[a, b]$ at which (2.2) does not hold.*

(i) *If $x(\cdot)$ is regular and the derivatives of $f(\cdot)$ and $x(\cdot)$ verify:*

$$Df(x(t))\cdot x'(t)\geqq 0 \quad \text{for any } t\in[a, b]\backslash J_{x(\cdot)}, \tag{2.3}$$

then $f(x(b))\geqq f(x(a))$.

(ii) *If $x(\cdot)$ is regular and satisfies:*

$$Df(x(t))\cdot x'(t)=0 \quad \text{for any } t\in[a, b]\backslash J_{x(\cdot)}, \tag{2.4}$$

then $t\mapsto f(x(t))$ is a constant function.

(iii) *If $f(\cdot)$ is, in addition, locally Lipschitzian (i.e. it is Lipschitzian on every compact subset of X) then the statements* (i) *and* (ii) *above hold for any absolutely continuous mapping $x(\cdot)$.*

Proof. Let $h(\cdot):[a, b]\to R$ be defined by: $h(t)=f(x(t))$, $t\in[a, b]$.

To prove (i), we note that if $x(\cdot)$ is regular, $t\in[a, b]\backslash J_{x(\cdot)}$ and $x(t)\in S\in\mathcal{S}_f$ then $x'(t)$ exists and $t\in x^{-1}(S)$ is not isolated to the right; hence there exists a sequence $t_k\searrow t$, $t_k\in x^{-1}(S)$ and therefore, since $f_S(\cdot)$ and $x(\cdot)$ are differentiable at $x(t)$ and, respectively, at t, it follows that

$$\lim_{k\to\infty}\frac{h(t_k)-h(t)}{t_k-t}=\lim_{k\to\infty}\frac{f(x(t_k))-f(x(t))}{t_k-t}=Df(x(t))\cdot x'(t)\geqq 0$$

since the local representative [15] $f_\alpha(\cdot): U\subset R^m\to R$ of $f_S(\cdot)$ at $x(t)\in S$ with respect to a local coordinate chart $(U, \alpha(\cdot))$, being Fréchet differentiable, has the property: $\lim_{u\to v, s\to 0}(f_\alpha(y+su)-f_\alpha(y))/s=Df_\alpha(y)\cdot v$ at any $y\in U$, $v\in R^m$, $m=\dim(S)$.

From (2.3) it follows now that $h(\cdot)=f(\cdot)ox(\cdot)$ satisfies the hypotheses in Lemma 2.4; hence $f(x(b))=h(b)\geqq h(a)=f(x(a))$.

The statement (ii) follows from (i) applied to the functions $f(\cdot)$ and $-f(\cdot)$ on every interval $[a, t]$, $t\in(a, b]$.

To prove (iii), we note that if $f(\cdot)$ is locally Lipschitzian and $x(\cdot)$ is absolutely continuous, then $h(\cdot)=f(\cdot)ox(\cdot)$ is absolutely continuous; hence at any point $t\in$

$[a, b]\backslash J_x(\cdot)$ at which $h(\cdot)$ is differentiable one has: $h'(t) = Df(x(t)) \cdot x'(t)$; since an absolutely continuous mapping is the indefinite integral of its derivative, (2.3) implies that $h(\cdot)$ is nondecreasing and (2.4) implies that $h(\cdot)$ is a constant function.

3. Generalized Hamiltonian systems for stratified optimal control problems. The optimal control problem we are considering is defined by the nonempty subsets: $X \subset R^n$ (*phase space*), $X_F \subset X$ (*target* or *terminal set*), $U \subset R^m$ (*control space*) and by the mappings: $g(\cdot): X_F \to R$ (*terminal payoff*) and $f(\cdot, \cdot): X \times U \to R^n$ (*controlled vector field*).

An optimal control problem is also defined by the set of admissible controls. Since in our approach Lemma 2.6 is essential and, on the other hand, it seems that the statements (i) and (ii) in this Lemma do not hold if $x(\cdot)$ is merely absolutely continuous (instead of being regular) unless $f(\cdot)$ is locally Lipschitzian, we shall consider separately the problem P_r in which the admissible controls are regulated mappings (the same type of results being valid for the problem P_{cp} in which the admissible controls are piecewise continuous) and the problem P_m in which the admissible controls are measurable and bounded.

For every $x_0 \in X\backslash X_F$ the set $\mathcal{U}_r(x_0)$ (respectively $\mathcal{U}_m(x_0)$) of *admissible controls with respect to x_0 for the problem P_r* (respectively, P_m) consists of all regulated (respectively, measurable bounded) mappings $u(\cdot): [0, t_F(u(\cdot))] \to U$ such that the (Carathéodory) solution $x(\cdot; x_0, u(\cdot))$ of the initial value problem:

$$x' = f(x, u(t)), \qquad x(0) = x_0 \tag{3.1}$$

is defined on $[0, t_F(u(\cdot))]$ and satisfies:

$$x_F(u(\cdot)) = x(t_F(\cdot); x_0, u(\cdot)) \in X_F, \tag{3.2}$$

$$x(t; x_0, u(\cdot)) \in X\backslash X_F \quad \text{for any } t \in [0, t_F(u(\cdot))). \tag{3.3}$$

The "terminal payoff," $g(\cdot)$, defines a "performance" $P(u(\cdot))$ for any admissible control $u(\cdot)$ as follows:

$$P(u(\cdot)) = g(x_F(u(\cdot)) \tag{3.4}$$

and an admissible control $\tilde{u}(\cdot) \in \mathcal{U}_r(x_0)$ (respectively, $\tilde{u}(\cdot) \in \mathcal{U}_m(x_0)$) is said to be *optimal with respect to the point $x_0 \in X\backslash X_F$ for the problem P_r* (respectively, P_m) if it satisfies:

$$P(\tilde{u}(\cdot)) \leqq P(u(\cdot)) \quad \text{for any } u(\cdot) \in \mathcal{U}_r(x_0) (u(\cdot) \in \mathcal{U}_m(x_0)). \tag{3.5}$$

An *optimal feedback control* for the problem P_r (respectively, P_m) is a mapping $v(\cdot): X_0 \subset X \to U$ such that for any $x_0 \in X_0\backslash X_F$ the initial value problem:

$$x' = f(x, v(x)), \qquad x(0) = x_0 \tag{3.6}$$

has a solution $\tilde{x}(\cdot; x_0)$ which is an *optimal trajectory* with respect to x_0 for the problem P_r (respectively, P_m) i.e. the mapping $\tilde{u}(t; x_0) = v(\tilde{x}(t; x_0))$ is optimal in $\mathcal{U}_r(x_0)$ (resp. in $\mathcal{U}_m(x_0)$).

If the control $\tilde{u}(\cdot)$ is optimal (in the sense of (3.5)), only among the admissible controls $u(\cdot)$ for which the corresponding trajectories $x(\cdot; x_0, u(\cdot))$ remain on a given subset $X_0 \subset X$ (i.e. $x(t; x_0, u(\cdot)) \in X_0\backslash X_F$ for any $t \in [0, t_F(u(\cdot))]$) we say that $\tilde{u}(\cdot)$ *solves the problem $P_r|X_0$* (respectively, $P_m|X_0$) at $x_0 \in X_0\backslash X_F$.

The so called "Pontryagin function" or "pseudo-Hamiltonian" [8] defined by:

$$\mathcal{H}(x, p, u) = \langle p, f(x, u)\rangle, \qquad p \in R^n\backslash\{0\} = R_0^n, \qquad (x, u) \in X \times U \tag{3.7}$$

as well as the Hamiltonian ("true Hamiltonian", [8]) defined by:

$$H(x,p)=\max\{\mathcal{H}(x,p,u);\ u\in U\},\qquad (x,p)\in A\subset X\times R_0^n \tag{3.8}$$

(where A denotes the set of all points $(x,p)\in X\times R_0^n$ for which $\mathcal{H}(x,p,\cdot)$ has a maximum on U) will play an essential role in what follows. We shall use also the corresponding marginal multifunction defined by:

$$\hat{U}(x,p)=\{u\in U;\ \mathcal{H}(x,p,u)=H(x,p)\},\qquad (x,p)\in A. \tag{3.9}$$

DEFINITION 3.1. The optimal control system $\Sigma=(X, X_F, U, f, g)$ is said to be *weakly stratified* if its data and the Hamiltonian have the following properties:

(i) $X\subset R^n$ is connected and $f(\cdot,\cdot)$ is continuous with respect to both variables and of class C^1 with respect to the first variable (i.e. if X is open, then the derivative $(x,u)\to D_1f(x,u)$ is continuous and if X is not open then $f(\cdot,\cdot)$ has an extension with this property on $X_1\times U$ where $X_1\supset X$ is open).

(ii) The terminal payoff $g(\cdot): X_F\to R$ is continuous and weakly C^1-stratified by a (weak) stratification $\mathcal{S}_g$ of X_F, of dimension $n_F<n$ (i.e. $\dim(S)\leqq n_F$ for any $S\in\mathcal{S}_g$).

(iii) The Hamiltonian, $H(\cdot,\cdot)$, is continuous and weakly C^1-stratified by a (weak) stratification $\mathcal{S}_H=\{A_j, j\in J\}$ of A.

We note that property (i) in Definition 3.1 implies the uniqueness of the solution $x(\cdot;x_0,u(\cdot))$ of (3.1) for any admissible control $u(\cdot)$ and also the fact that $x(\cdot;x_0,u(\cdot))$ is regular if $u(\cdot)$ is regulated. In fact, the uniqueness is implied by the weaker property of $f(\cdot,\cdot)$ being locally Lipschitzian with respect to the first variable but the class C^1 was assumed in view of property (iii). The existence of local solutions of (3.1) is given by Peano's theorem at any interior point $x_0\in X\setminus X_F$ and by Nagumo-type theorems [25], [26] requiring an additional tangency property of the controlled vector field, $f(\cdot,\cdot)$, at the boundary points of the subset $X\subset R^n$.

It should be noted also that the existence of admissible controls need not be assumed in this setting.

DEFINITION 3.2. Let $\Sigma=(X, X_F, U, f, g)$ be a weakly stratified control system and let $H(\cdot,\cdot)$ be its Hamiltonian.

The set-valued mapping $\xi_H(\cdot,\cdot)$ defined by

$$\xi_H(x,p)=\{(f(x,u),-pD_1f(x,u))\in T_{(x,p)}A;\ u\in\hat{U}(x,p)\},\qquad (x,p)\in A \tag{3.10}$$

is said to be the *controlled Hamiltonian orientor field* of $H(\cdot,\cdot)$ and the differential inclusion

$$(x',p')\in\xi_H(x,p) \tag{3.11}$$

is said to be the Hamiltonian inclusion of the control system Σ.

The set-valued mapping $\xi_H^{\#}(\cdot,\cdot)$ defined by:

$$\begin{aligned}\xi_H^{\#}(x,p)=\{&(\bar{x},\bar{p})\in T_{(x,p)}A;\ \langle\bar{x},\bar{q}\rangle-\langle\bar{p},\bar{y}\rangle\\ &=DH(x,p)\cdot(\bar{y},\bar{q})\text{ for any }(\bar{y},\bar{q})\in T_{(x,p)}A\},\qquad (x,p)\in A\end{aligned} \tag{3.12}$$

is said to be the natural Hamiltonian orientor field of $H(\cdot,\cdot)$.

Remark 3.3. We note that in the case a stratum $A_j\in\mathcal{S}_H$ is a *natural symplectic submanifold of* R^{2n} (i.e. the restriction $\Omega|A_j$ of the natural symplectic form Ω on R^{2n} is a symplectic form on A_j) the restriction $\xi_H^{\#}(\cdot,\cdot)|A_j$ coincides with the *Hamiltonian vector field* $(dH_j)^{\#}$ on A_j defined by $H_j(\cdot,\cdot)=H(\cdot,\cdot)|A_j$ [11], [20]. From (3.12) it follows [20] that at any point $(x,p)\in A$ the set $\xi_H^{\#}(x,p)$ is either a singleton (at "symplectic" points) or a linear manifold (at "singular" points) or the empty set (at "transversal" points).

The justification of Definition 3.2 as well as important properties of the solutions of (3.11) will follow from:

PROPOSITION 3.4. *If $\Sigma=(X, X_F, U, f, g)$ is a weakly stratified optimal control system then the derivatives of $H(\cdot,\cdot)$ and $\mathcal{H}_u(\cdot,\cdot)=\mathcal{H}(\cdot,\cdot,u)$, $u\in U$, are related as follows*:

$$DH(x,p)\cdot(\bar{x},\bar{p})=D\mathcal{H}_u(x,p)\cdot(\bar{x},\bar{p})=\langle pD_1f(x,u),\bar{x}\rangle+\langle f(x,u),\bar{p}\rangle \tag{3.13}$$
$$\text{for any } (x,p)\in A,\qquad (\bar{x},\bar{p})\in T_{(x,p)}A,\ u\in\hat{U}(x,p).$$

Proof. Let $(x,p)\in A_j\in\mathcal{S}_H$, $(\bar{x},\bar{p})\in T_{(x,p)}A$, $u\in\hat{U}(x,p)$ and let $c(\cdot):(-a,a)\to A_j$, $a>0$, be of class C^1 such that: $c(0)=(x,p)$, $c'(0)=(\bar{x},\bar{p})$; from the definition of the derivative of a mapping defined on a differentiable manifold (e.g. [15]) it follows: $DH(x,p)\cdot(\bar{x},\bar{p})=\lim_{t\to 0}(H(c(t))-H(c(0))/t$ and similarly for the derivative of $\mathcal{H}_u(\cdot,\cdot)$. The second part of (3.13) follows directly from (3.7). To prove the first part of (3.13), we note that from (3.8) it follows: $H(c(t))\geqq\mathcal{H}_u(c(t))$ for any $t\in(-a,a)$ and $H(c(0))=\mathcal{H}_u(c(0))$ and therefore for $t\searrow 0$ $(t>0)$ we have: $(H(c(t))-H(c(0)))/t\geqq(\mathcal{H}_u(c(t))-\mathcal{H}_u(c(0)))/t\to D\mathcal{H}_u(x,p)\cdot(\bar{x},\bar{p})$ and hence $DH(x,p)\cdot(\bar{x},\bar{p})\geqq D\mathcal{H}_u(x,p),(\bar{x},\bar{p})$; reasoning in the same way for $t\nearrow 0$, we obtain the reversed inequality; hence (3.13).

COROLLARY 3.5. *The multifunctions $\xi_H(\cdot,\cdot)$ and $\xi_H^{\#}(\cdot,\cdot)$ are related by*:

$$\xi_H(x,p)\subset\xi_H^{\#}(x,p)\quad \text{for any } (x,p)\in A. \tag{3.14}$$

Moreover, if the multifunction $\tilde{U}(\cdot,\cdot)$ is defined by:

$$\tilde{U}(x,p)=\{u\in\hat{U}(x,p);\ (f(x,u),-pD_1f(x,u))\in\xi_H(x,p)\},\qquad (x,p)\in A, \tag{3.15}$$

then $\xi_H(\cdot,\cdot)$ is given by:

$$\xi_H(x,p)=\{(f(x,u),-pD_1f(x,u));\ u\in\tilde{U}(x,p)\},\qquad (x,p)\in A \tag{3.16}$$

and the following property holds:

$$DH(x,p)\cdot(\bar{x},\bar{p})=0$$
$$\text{for any } (\bar{x},\bar{p})\in\xi_H(x,p),\qquad (x,p)\in A^0=\{(x,p)\in A;\ \tilde{U}(x,p)\neq\varnothing\}. \tag{3.17}$$

Proof. If $(\bar{x},\bar{p})\in\xi_H(x,p)$ and $(\bar{y},\bar{q})\in T_{(x,p)}A$ then from (3.10) it follows that there exists $u\in\hat{U}(x,p)$ such that $(\bar{x},\bar{p})=(f(x,u),-pD_1f(x,u))$; hence from (3.13) it follows: $\langle\bar{x},\bar{q}\rangle-\langle\bar{p},\bar{q}\rangle=\langle f(x,u),\bar{q}\rangle+\langle pD_1f(x,u),\bar{y}\rangle=D\mathcal{H}_u(x,p)\cdot(\bar{y},\bar{q})=DH(x,p)\cdot(\bar{y},\bar{q})$ i.e. $(x,p)\in\xi_H^{\#}(x,p)$ and (3.14) is proved.

If $u\in\tilde{U}(x,p)\subset\hat{U}(x,p)$, then from (3.15) and (3.12) it follows that $(\bar{x},\bar{p})=(f(x,u),-pD_1f(x,p))\in T_{(x,p)}A$; hence $(\bar{x},\bar{p})\in\xi_H(x,p)$ and (3.16) is proved.

Finally, if $(x,p)\in A^0$ (hence $\xi_H(x,p)\neq\varnothing$) and $(\bar{x},\bar{p})\in\xi_H(x,p)$, then from (3.14) it follows that $(\bar{x},\bar{p})\in\xi_H^{\#}(x,p)$ and from (3.12) it follows that $DH(x,p)\cdot(\bar{y},\bar{q})=\langle\bar{x},\bar{q}\rangle-\langle\bar{p},\bar{y}\rangle$ for any $(\bar{y},\bar{q})\in T_{(x,p)}A$; hence for $(\bar{y},\bar{q})=(\bar{x},\bar{p})$ we get (3.17).

The following sets of "terminal values" for the solutions of the Hamiltonian inclusion (3.11) are introduced in accordance with the *transversality conditions* in Pontryagin's maximum principle:

$$X^*_{F,n}=\{(x,p)\in A\cap(X_F\times R^n);\ H(x,p)=0,\langle p,\bar{x}\rangle=-Dg(x)\cdot\bar{x}(\forall)\bar{x}\in T_xX_F\}, \tag{3.18}$$

$$X^*_{F,a}=\{(x,p)\in A\cap(X_F\times R^n);\ H(x,p)=0,\langle p,\bar{x}\rangle=0(\forall)\bar{x}\in T_xX_F\}, \tag{3.19}$$

$$X^*_F=X^*_{F,n}\cup X^*_{F,a}. \tag{3.20}$$

The set $X^*_{F,n}$ corresponds to the "normal extremals" (in the sense of mathematical programming, [10]) of the optimal control system and the set $X^*_{F,a}$ corresponds to the "abnormal" ones. If $X^*_F=X^*_{F,n}$, then the optimal control system Σ is said to be *normal.*

DEFINITION 3.6. A mapping $x^*(\cdot)=(x(\cdot), p(\cdot)):[\![\tau, 0]\!]\to A$ which is regular with respect to the stratification $\mathcal{S}_H$ of A, satisfies:

$$x^*(0)=(x(0), p(0))\in X_F^*, \tag{3.21}$$

$$x(t)\in X\setminus X_F \quad \text{for any } t\in[\![\tau, 0) \tag{3.22}$$

and satisfies (3.11) everywhere except at its switching points is said to be a *regular characteristic* of the system Σ.

An absolutely continuous mapping $x^*(\cdot)=(x(\cdot), p(\cdot))$ satisfying (3.21) and (3.22) and satisfying (3.11) almost everywhere is said to be a *characteristic* of Σ.

We denote by $\mathcal{C}_r$ the set of all regular characteristics and by $\mathcal{C}_m$ the set of all characteristics of Σ. We denote also by $\mathcal{C}_r^n$ (respectively, $\mathcal{C}_m^n$) the set of regular characteristics that instead of (3.21) satisfy:

$$x^*(0)=(x(0), p(0))\in X_{F,n}^*, \tag{3.23}$$

and we call them *normal characteristics.*

PROPOSITION 3.7. *If $\Sigma=(X, X_F, U, f, g)$ is a weakly stratified optimal control system, then any regular characteristic $x^*(\cdot)=(x(\cdot), p(\cdot))\in\mathcal{C}_r$ satisfies*:

$$H(x(t), p(t))=0 \quad \textit{for any } t\in[\![\tau, 0], \tag{3.24}$$

$$\langle p(t), x'(t)\rangle=0 \quad \textit{for any } t\in[\![\tau, 0]\!]\setminus J_{x^*(\cdot)} \tag{3.25}$$

where $J_{x^*(\cdot)}$ denotes the set of the switching points of $x^*(\cdot)$.

If, in addition, $H(\cdot,\cdot)$ is locally Lipschitzian, then every characteristic $x^*(\cdot)\in\mathcal{C}_m$ satisfies (3.24) *and* (3.25) *where $J_{x^*(\cdot)}$ is the null set in Lemma* 2.5 *at which the relation*:

$$(x'(t), p'(t))\in T_{(x(t),p(t))}A \tag{3.26}$$

does not hold.

Proof. In both cases, (3.24) is an immediate consequence of (3.11), (3.17) and Lemma 2.6.

To prove (3.25), we note that from (3.10), (3.11) and (3.15) it follows that for any $t\in[\![\tau, 0]\!]\setminus J_{x^*(\cdot)}$ there exists $u(t)\in\tilde{U}(x(t), p(t))$ such that:

$$(x'(t), p'(t))=(f(x(t), u(t)), -p(t)D_1f(x(t), u(t))); \tag{3.27}$$

hence from (3.7), (3.8) and (3.24) it follows: $\langle p(t), x'(t)\rangle=\langle p(t), f(x(t), u(t))\rangle=\mathcal{H}(x(t), p(t), u(t))=H(x(t), p(t))=0$.

Remark 3.8. As it is apparent from the general idea of the approach in this paper, the projection, $x(\cdot)$, on the phase space of a characteristic, $x^*(\cdot)=(x(\cdot), p(\cdot))$, should be an admissible trajectory and we may, of course, introduce this requirement in the Definition 3.6 of the characteristics. In this case a regular characteristic need not be "regular with respect to the stratification $\mathcal{S}_H$" as it was asked in Definition 3.6. We prefer, however, to give some sufficient conditions implying that the projections on the phase space of the characteristics in Definition 3.6 are admissible trajectories.

We consider first the mapping $\xi(\cdot,\cdot,\cdot):X\times R^n\times U\to R^n$ defined by:

$$\xi(x, p, u)=(f(x, u), -pD_1f(x, u)), \qquad (x, p, u)\in X\times R^n\times U \tag{3.28}$$

and for any stratum $A_j\in\mathcal{S}_H$ for which $A_j\cap A^0\neq\varnothing$ (recall that $(x, p)\in A^0$ iff $\xi_H(x, p)\neq\varnothing$) we define:

$$V_j(x, p, v)=\begin{cases}\{u\in\tilde{U}(x, p);\ \xi(x, p, u)=v\} & \text{if } (x, p)\in A_j\cap A^0,\quad v\in R^{2n}\\ \limsup V_j(y, q, w) & \text{if } (x, p)\in A\cap Cl(A_j\cap A^0)\setminus(A_j\cap A^0)\\ (y, q, w)\to(x, p, v),\ (y, q, w)\in(A_j\cap A^0)\times R^{2n}\end{cases} \tag{3.29}$$

where $\limsup_{z\to z_0} V(z)=\{u;\ (\exists)z_k\to z_0,\ u_k\in V(z_k) \text{ such that } u_k\to u\}$.

DEFINITION 3.9. The optimal control system $\Sigma = (X, X_F, U, f, g)$ is said to be *regular* if it is weakly stratified and for any stratum $A_j \in \mathcal{S}_H$ for which $A_j \cap A^0 \neq \emptyset$, the set-valued mapping $V_j(\cdot,\cdot,\cdot)$ defined by (3.29) is compact-valued and is Lipschitzian with respect to the Hausdorff distance on any compact subset of its effective domain $\{(x, p, v);\ V_j(x, p, v) \neq \emptyset\}$.

The system Σ is said to be *Lipschitzian* if it is weakly stratified, the Hamiltonian $H(\cdot,\cdot)$ is locally Lipschitzian and the multifunction $\tilde{U}(\cdot,\cdot)$ defined by (3.15) is compact-valued and has closed graph.

We prove now that if Σ is either regular or Lipschitzian, then the projection on the phase space of any characteristic is an admissible trajectory:

PROPOSITION 3.10. *If Σ is a regular optimal control system, then for any regular characteristic $x^*(\cdot) = (x(\cdot), p(\cdot)) : [\tau, 0] \to A$ the projection, $\tilde{x}(\cdot)$, on the phase space, defined by*:

$$\tilde{x}(t) = x(t+\tau), \qquad t \in [0, -\tau] \tag{3.30}$$

is an admissible trajectory with respect to $x(\tau) \in X$ for the problem P_r.

If Σ is Lipschitzian, then the projection $\tilde{x}(\cdot)$ of any characteristic $x^*(\cdot) \in \mathcal{C}_m$ is an admissible trajectory with respect to $x(\tau) \in X$ for the problem P_m.

Proof. Let Σ be regular, $x^*(\cdot) = (x(\cdot), p(\cdot)) \in \mathcal{C}_r$ and let $\{I_k;\ k \in N\}$ be a partition of $[\tau, 0]$ into subintervals such that for any $k \in N$ there exists $A_k \in \mathcal{S}_H$ such that $x^*(I_k) \subset A_k$, the restriction $x_k^*(\cdot) = x^*(\cdot)|I_k$ is of class C^1 and its derivative has one-sided limits at the endpoints of I_k (Definition 2.3). From Definition 3.9 it follows that for any $k \in N$ the set-valued mapping $t \to V_k(x(t), p(t), (x'(t), p'(t)))$ is compact-valued and Lipschitzian on the compact interval $Cl(I_k)$; hence according to [13, Thm. 2] it has a continuous selection $u_k(\cdot)$; from (3.28) and (3.29) it follows that $u(\cdot) : [\tau, 0] \to U$ defined by: $u(t) = u_k(t)$ if $t \in I_k$, $k \in N$ and the first component, $x(\cdot)$, of $x^*(\cdot)$, verify (3.27) at any point $t \in [\tau, 0] \setminus J_{x(\cdot)}$; hence from (3.21) and (3.22) it follows that $\tilde{u}(\cdot)$ defined by: $\tilde{u}(t) = u(t+\tau)$ for $t \in [0, -\tau]$ is a regulated admissible control whose admissible trajectory is $\tilde{x}(\cdot)$ defined by (3.30).

If Σ is Lipschitzian and $x^*(\cdot) = (x(\cdot), p(\cdot)) \in \mathcal{C}_m$, then from the implicit function theorem for orientor fields (e.g. [7, § 8.2]) it follows that the multifunction $t \mapsto \tilde{U}(x(t), p(t))$ has a measurable selection $u(\cdot)$ satisfying (3.27) a.e.; hence $\tilde{u}(t) = u(t+\tau)$, and $t \in [0, -\tau]$ is a measurable bounded admissible control for the problem P_m whose trajectory is $\tilde{x}(\cdot)$ defined by (3.30).

Remark 3.11. The rather complicated property of a regular control system in Definition 3.9 was assumed in the absence of suitable theorems concerning the existence of continuous selections $u_k(t) \in \tilde{U}(x(t), p(t))$, $t \in I_k$ satisfying (3.27) (i.e. Filippov's lemmas for continuous implicit functions). This property may obviously be replaced by more restrictive but easier verifiable conditions like the following one: *for any $(x, p) \in A^0$* (for which $\xi_H(x, p) \neq \emptyset$) *$\xi_H(x, p)$ is a singleton and for stratum $A_j \in \mathcal{S}_H$ for which $A_j \cap A^0 \neq \emptyset$, the restriction $\tilde{U}(\cdot,\cdot)|(A_j \cap A^0)$ admits an extension to $Cl(A_j \cap A^0)$ that is either single-valued and continuous or compact-valued and locally Lipschitzian.*

4. Sufficient optimality conditions. As an easy corollary of Lemma 2.6 in § 2 we obtain the following improved version of Boltyanskii's "fundamental lemma" ([1, Chap. III, § 11], [5], etc.):

LEMMA 4.1. *Let $\Sigma = (X, X_F, U, f, g)$ be an optimal control system with properties* (i) *and* (ii) *in Definition* 3.1 *and let $W(\cdot) : X \to R$ be continuous and weakly C^1-stratified*

such that:

(4.1) $$W(x) = g(x) \quad \textit{for any } x \in X_F,$$

(4.2) $$DW(x) \cdot f(x, u) \geqq 0 \quad \textit{for any } x \in X \quad \textit{and any } u \in U \textit{ for which } f(x, u) \in T_x X.$$

Then for any $x_0 \in X \setminus X_F$ *and any admissible control* $u(\cdot) \in \mathcal{U}_r(x_0)$ *one has*:

(4.3) $$W(x_0) \leqq P(u(\cdot))$$

and therefore any admissible control $\tilde{u}(\cdot) \in \mathcal{U}_r(x_0)$ *for which*:

(4.4) $$W(x_0) = P(\tilde{u}(\cdot))$$

is optimal with respect to x_0 *for the problem* P_r.

If, in addition, $W(\cdot)|(X \setminus X_F)$ *is locally Lipschitzian, then the statements above hold for any measurable bounded admissible control.*

Proof. If $u(\cdot) \in \mathcal{U}_r(x_0)$ and $x(\cdot) = x(\cdot; x_0, u(\cdot))$ is the corresponding admissible trajectory then $x(\cdot)$ is regular and according to Lemma 2.5 there exists a countable subset $J_{x(\cdot)} \subset [0, t_F(u(\cdot))]$ such that $x'(t) = f(x(t), u(t)) \in T_{x(t)}X$ for any $t \in [0, t_F(u(\cdot))] \setminus J_{x(\cdot)}$ and therefore, since (4.2) implies: $DW(x(t)) \cdot x'(t) \geqq 0$, (4.3) follows from (4.1) and Lemma 2.6.

If $W(\cdot)|(X \setminus X_F)$ is locally Lipschitzian, $x_0 \in X \setminus X_F$ and $u(\cdot) \in \mathcal{U}_m(x_0)$ then from Lemma 2.5 it follows that there exists a null set $J_{x(\cdot)} \subset [0, t_F(u(\cdot))]$ such that $x'(t) = f(x(t), u(t)) \in T_{x(t)}X$ for any $t \in [0, t_F(u(\cdot))] \setminus J_{x(\cdot)}$ and (4.2) implies: $DW(x(t)) \cdot x'(t) \geqq 0$ for any $t \in [0, t_F(u(\cdot))] \setminus J_{x(\cdot)}$; since for any $t_1 \in (0, t_F(u(\cdot)))$, $x([0, t_1]) \subset X \setminus X_F$, the function $t \mapsto W(x(t))$ is absolutely continuous on $[0, t_1]$ and (4.3) follows in the same way as in the proof of the second part of Lemma 2.6.

Remark 4.2. We note that in Boltyanskii's fundamental lemma [1] and its other versions ([5], [6], etc.) the phase space $X \subset R^n$ is open, the terminal set X_F is a singleton, condition (4.2) is verified on $X \setminus M$ where $M \subset R^n$ is a "piecewise smooth" set of dimension less than n [1] or a Whitney-stratified set [5], [6] and the admissible controls are allowed to be only piecewise continuous (and the proof is very hard!). Apparently [Lemma 4.1, Condition 4.2] is more restrictive than the corresponding one in Boltyanskii's lemma since it must be verified also on lower dimensional strata of X but, on the other hand, Boltyanskii's lemma is used only in connection with the very complicated object called "regular synthesis" ([1], [5]–[7], [17], [24], etc.) for which, as it is shown in [17], condition (4.2) is necessarily verified on all the "cells" of the synthesis.

In what follows we use Lemma 4.1 in connection with the characteristics of the optimal control problem, hence before their projections on the phase space are "synthesized" into a precisely described global picture called "regular synthesis".

It should be noted, however, that the approach in this paper based on Lemmas 2.4–2.6 and 4.1 leaves out important classes of optimal control problems that have discontinuous value functions [9], [22].

We consider first the problem P_r for the optimal control system $\Sigma = (X, X_F, U, f, g)$ which is assumed to be regular (Definition 3.9) and we introduce the following notations: for any $\eta = (x_F, p_F) \in X_F^*$ we denote by $\mathcal{C}_r(\eta)$ the *set of all noncontinuable to the left regular characteristics* $x_k^*(\cdot; \eta) = (x_k(\cdot; \eta), p_k(\cdot; \eta)) : [\![\tau_k(\eta), 0] \to A$, $k \in \mathcal{H}_r(\eta)$, satisfying:

(4.5) $$x_k^*(0; \cdot) = \eta \in X_F^*,$$

(4.6) $$x_k(t) \in X \setminus X_F \quad \text{for any } t \in [\![\tau_k(\eta), 0).$$

As it is apparent from the definition of $\mathscr{C}_r(\eta)$ we allow the Hamiltonian system (3.11) to have more than one regular solution through the same points $(0, \eta)$ though this is not the case for significant classes of optimal control problems in the literature.

We define also the following sets and functions:

$$\chi_F(\eta) = y_F(t, \eta) = x_F, \quad \eta_F(t, \eta) = \eta \quad \text{if } \eta = (x_F, p_F) \in X_F^*, \quad t \in R, \tag{4.7}$$

$$Y_r^k(\eta) = [\![\tau_k(\eta), 0]\!] \times \{\eta\}, \qquad Y_r = \cup\{Y_r^k(\eta);\ \eta \in X_F^*,\ k \in \mathscr{K}_r(\eta)\} \tag{4.8}$$

$$X_r = \{x_k(t; \eta);\ (t, \eta) \in Y_r,\ k \in \mathscr{K}_r(\eta)\}, \tag{4.9}$$

$$W_r(x) = \min\{g(\chi_F(\eta));\ (t, \eta) \in Y_r,\ k \in \mathscr{K}_r(\eta),\ x_k(t; \eta) = x\}, \qquad x \in \tilde{X}_r \tag{4.10}$$

and we assume that the set $\tilde{X}_r \subset X_r$ on which $W_r(\cdot)$ is defined verifies:

$$\tilde{X}_r \setminus X_F \neq \varnothing. \tag{4.11}$$

The corresponding marginal multifunctions of (4.10) are given by:

$$\tilde{Y}_r(x) = \{(t, \eta) \in Y_r;\ g(\chi_F(\eta)) = W_r(x),\ x_k(t; \eta) = x,\ k \in \mathscr{K}_r(\eta)\}, \qquad x \in \tilde{X}_r, \tag{4.12}$$

$$\tilde{K}_r(x) = \{k \in \mathscr{K}_r(\eta);\ (t, \eta) \in \tilde{Y}_r(x),\ x_k(t; \eta) = x\}, \qquad x \in \tilde{X}_r, \tag{4.13}$$

THEOREM 4.3. *Let $\Sigma = (X, X_F, U, f, g)$ be a regular optimal control system and let $W_r(\cdot): \tilde{X}_r \to R$ defined by* (4.10) *be continuous and weakly C^1-stratified and such that for any $x_0 \in \tilde{X}_r \setminus X_F$ there exist $(t_0, \eta) \in \tilde{Y}_r(x_0)$, $k \in \tilde{K}_r(x_0)$ that verify*:

$$DW_r(x_0) \cdot \bar{x} = -\langle p_k(t_0; \eta), \bar{x}\rangle \quad \text{for any } \bar{x} \in T_{x_0}\tilde{X}_r. \tag{4.14}$$

Then for any such elements, the mapping $\tilde{x}(\cdot; x_0)$ defined by:

$$\tilde{x}(t; x_0) = x_k(t + t_0; \eta) \quad \text{for any } t \in [0, -t_0] \tag{4.15}$$

is an optimal trajectory with respect to x_0 for the problem $P_r|\tilde{X}_r$.

If, in addition $H(\cdot, \cdot)$ and $W_r(\cdot)|(\tilde{X}_r \setminus X_F)$ are locally Lipschitzian, then $\tilde{x}(\cdot; x_0)$ is optimal also for the problem $P_m|\tilde{X}_r$.

Proof. From Proposition 3.10 it follows that $\tilde{x}(\cdot; x_0)$ defined by (4.15) is a regular admissible trajectory realized by a (regulated) admissible control whose performance is given by: $P(\tilde{u}(\cdot)) = g(\chi_F(\eta)) = W_r(x_0)$. From (4.6) it follows that $W_r(\cdot)$ verifies (4.1) and from (4.14) it follows that $W_r(\cdot)$ verifies (4.2) since if $\bar{x} = f(x_0, u) \in T_{x_0}\tilde{X}_r$ then $\langle p_k(t_0; \eta), f(x_0, u)\rangle = \mathscr{H}(x_k(t_0; \eta), p_k(t_0; \eta), u) \leqq H(x_k(t_0; \eta), p_k(t_0; \eta)) = 0$ (Proposition 3.7) and therefore Theorem 4.3 follows from Lemma 4.1.

To state an analogous theorem for the problem P_m in which the admissible controls are measurable and bounded, we define $\mathscr{C}_m(\cdot)$, $\mathscr{K}_m(\cdot)$, Y_m, X_m, $\tilde{X}_m$, $W_m(\cdot)$, $\tilde{Y}_m(\cdot)$, $\tilde{K}_m(\cdot)$ as in (4.5)–(4.13) replacing the term "regular characteristics" by the term "characteristics". The proof of the next theorem is entirely similar to the proof of Theorem 4.3 so we omit it.

THEOREM 4.4. *Let $\Sigma = (X, X_F, U, f, g)$ be a Lipschitzian optimal control system (Definition* 3.9) *and let $W_m(\cdot): \tilde{X}_m \to R$ defined by* (4.10) *be continuous, weakly C^1-stratified and such that its restriction to $\tilde{X}_m \setminus X_F$ is locally Lipschitzian. If for any $x_0 \in \tilde{X}_m \setminus X_F$ there exist $(t_0, \eta) \in \tilde{Y}_m(x_0)$ and $k \in \tilde{K}_m(x_0)$ such that* (4.14) *holds, then $\tilde{x}(\cdot; x_0)$ defined by* (4.15) *is an optimal trajectory with respect to x_0 for the problem $P_m|\tilde{X}_m$.*

If, in addition, $\tilde{x}(\cdot; x_0)$ is realized by a regulated admissible control, then it is optimal also for the problem $P_r|\tilde{X}_m$.

Remark 4.5. If for any $x \in \tilde{X}_r$ we define: $V_r(x) = \cup\{\tilde{U}(x_k^*(t; \eta));\ (t, \eta) \in \tilde{Y}_r(x),\ k \in \tilde{K}_r(x)\}$, then obviously any mapping $v(\cdot): \tilde{X}_r \to U$ satisfying: $v(x) \in V_r(x)$ for any

$x \in \tilde{X}_r$ is an optimal feedback control for the problem $P_r|\tilde{X}_r$; a similar statement holds for the problem $P_m|\tilde{X}_m$.

We note that condition (4.14) together with the inequality: $DW_r(x(t)) \cdot f(x(t), u(t)) \geqq 0$ along any admissible pair $(u(\cdot), x(\cdot))$ of the problem $P_r|\tilde{X}_r$ obtained in the proof of Theorem 4.3 may be interpreted as a generalization of the well-known Hamilton-Jacobi-Bellman equation of dynamic programming:

$$\min \{DW_r(x) \cdot f(x, u); u \in U, f(x, u) \in T_x\tilde{X}_r\} = 0 \quad \text{for any } x \in \tilde{X}_r$$

for which the Hamiltonian inclusion (3.11) plays the role of a "system of characteristics".

We note also that Theorems 4.3 and 4.4 remain valid if in (4.5)-(4.13) one takes any family (not necessarily all) of (regular) characteristics but in this case the sets $\tilde{X}_r$ and $\tilde{X}_m$ may be too small and the optimal solutions of the problems $P_r|\tilde{X}_r$, $P_m|\tilde{X}_m$ may not be optimal solutions for the original problems, P_r and, respectively, P_m, since there may exist admissible trajectories that do not remain in $\tilde{X}_r$ ($\tilde{X}_m$).

One of the main difficulties in applying Theorems 4.3 and 4.4 to concrete problems is the verification of condition (4.14). The results to follow show that we can avoid condition (4.14) if we restrict ourselves to "normal" problems and assume that the characteristics may be embedded in a "stratified Hamiltonian flow" defined as follows:

DEFINITION 4.6. A continuous mapping $x^*(\cdot, \cdot) = (x(\cdot, \cdot), p(\cdot, \cdot)): Y \subset (-\infty, 0] \times X^*_{F,n} \to A$ is said to be a *stratified Hamiltonian flow* of (3.11) if it has the following properties:

(i) There exist the upper semicontinuous functions $\tau(\cdot)$, $\tau_i(\cdot): X^*_{F,n} \to [-\infty, 0]$, $i \in Z$, that are weakly C^1-stratified by a stratification $\mathcal{S}^*_F$ of $X^*_{F,n}$ and satisfy:

$$\tau(\eta) \leqq \cdots \leqq \tau_{i-1}(\eta) \leqq \tau_i(\eta) \leqq \cdots \leqq 0 \quad \text{for any } \eta \in X^*_{F,n}, \tag{4.16}$$

$$\lim_{i \to +\infty} \tau_i(\eta) = 0, \quad \lim_{i \to -\infty} \tau_i(\eta) = \tau(\eta) \quad \text{for any } \eta \in X^*_{F,n}, \tag{4.17}$$

$$Y = \{(t, \eta); \eta \in X^*_{F,n}, t \in [\![\tau(\eta), 0]\!]\} \tag{4.18}$$

and for any $\eta \in X^*_{F,n}$ the mapping $x^*(\cdot; \eta): [\![\tau(\eta), 0]\!] \to A$ is a regular characteristic of (3.11) satisfying (4.5)-(4.6), with the switching points $\{\tau_i(\eta); i \in Z\}$.

(ii) If Z_0 denotes the set of those $i \in Z$ for which the set:

$$Y_i = \{(t, \eta) \in Y; \tau_{i-1}(\eta) < t < \tau_i(\eta)\} \tag{4.19}$$

is not empty, then for any $i \in Z_0$ the "time-derivative" defined by:

$$x'(t; \eta) = D_1 x(t; \eta), \qquad (t, \eta) \in Y_i \tag{4.20}$$

has one-sided limits at $\tau_{i-1}(\eta)$, $\tau_i(\eta)$, and the restrictions $x^*_i(\cdot, \cdot)$, $x_i(\cdot, \cdot)$, $x'_i(\cdot, \cdot)$ of $x^*(\cdot, \cdot)$, $x(\cdot, \cdot)$, $x'(\cdot, \cdot)$, respectively, at $\bar{Y}_i$ defined by

$$\bar{Y}_i = \{(t, \eta) \in Y; \tau_{i-1}(\eta) < t < \tau_i(\eta)\} \tag{4.21}$$

are weakly C^1-stratified by a stratification $\mathcal{S}_i$ of $\bar{Y}_i$ such that $x_i(\cdot, \cdot)$ is a weakly C^1-stratified submersion and for any $S \in \mathcal{S}_i$ there exists $A_S \in \mathcal{S}_H$ such that $x^*(S) \subset A_S$.

(iii) For any $i \in Z_0$ the derivatives of $x_i(\cdot, \cdot)$ and $x'_i(\cdot, \cdot)$ verify:

$$Dx'_i(t; \eta) = D_1 Dx_i(t; \eta) \quad \text{for any } (t, \eta) \in \bar{Y}_i, \tag{4.22}$$

$$\begin{aligned} &Dx_i(\tau_i(\eta); \eta) \cdot (\bar{t}, \bar{\eta}) = (\bar{t} - D\tau_i(\eta) \cdot \bar{\eta}) x'_i(t; \eta) + \bar{x}_F \\ &\text{if } \bar{\eta} = (\bar{x}_F, \bar{p}_F) \subset T_\eta X^*_{F,n} \text{ and } \bar{t} \in R. \end{aligned} \tag{4.23}$$

(iv) For any $i \in Z_0$ and $S \in \mathcal{S}_i$ there exist $B_S^i \in \mathcal{S}_F^*$, $C_S^i \in \mathcal{S}_g$ such that for any $(t, (x_F, p_F)) \in S$ one has: $\eta = (x_F, p_F) \in B_S^i$, $x_F \in C_S^i$.

Remark 4.7. Since in the case $\tau_{-i}(\eta) = \tau(\eta) < 0 = \tau_0(\ ) = \tau_i(\eta)$ for any $i \in N$, $\eta \in R^{2n}$, (4.22) and (4.23) are the usual properties of the flow of a smooth vector field [15], the stratified Hamiltonian flow in Definition 4.6 is a generalization of such an object. In fact, the obvious way to obtain a stratified Hamiltonian flow is to "concatenate" a family, $\{x_i^*(\cdot;\cdot);\ i \in Z\}$, of smooth flows "via" a set, $\{\tau_i(\cdot);\ i \in Z\}$, of functions with property (i) in Definition 4.6 as follows [21]: $x^*(t;\eta) = x_i^*(t - \tau_i(\eta);\ x^*(\tau_i(\eta), \eta))$ for any $t \in [\tau_{i-1}(\eta), \tau_i(\eta)]$, $\eta \in X_{F,n}^*$.

We note that according to (4.16) and (4.17) we allow 0 and $\tau(\eta)$ to be cluster points of the set $\{\tau_i(\eta);\ i \in Z\}$ of the switching points of the regular characteristic $x^*(\cdot;\eta)$; obviously, the stratified Hamiltonian flow in Definition 4.6 may be generalized so as to contain even more pathological cases.

LEMMA 4.8. *If Σ is a weakly stratified optimal control system and $x^*(\cdot,\cdot) = (x(\cdot,\cdot), p(\cdot,\cdot))$ is a stratified Hamiltonian flow, then for any $i \in Z_0$, $(t,\eta) \in Y_i$ (Definition 4.6) and any $(\bar{t}, \bar{\eta}) \in T_{(t,\eta)}\bar{Y}_i$ one has:*

$$\langle p_i(t,\eta), Dx_i(t,\eta)\cdot(\bar{t},\bar{\eta})\rangle = -Dg(y_F(t,\eta))\cdot\bar{x}_F \quad \text{if } \bar{\eta} = (\bar{x}_F, \bar{p}_F). \tag{4.24}$$

Proof. According to (3.25) in Proposition 3.7 we have: $\langle p_i(t,\eta), x_i'(t,\eta)\rangle = 0$ for any $(t,\eta) \in \bar{Y}_i$ and therefore, since this is a stratified constant function, its derivative must vanish everywhere: using (4.22), (3.11) and (3.12), we obtain: $x_i'(t,\eta)\cdot Dp_i(t,\eta) + p_i(t,\eta)D_1Dx_i(t,\eta) = 0$ and therefore:

$$\begin{aligned} d/dt(p_i(t,\eta)Dx_i(t,\eta)) &= p_i'(t,\eta)Dx_i(t,\eta) + p_i(t,\eta)D_1Dx_i(t,\eta) \\ &= p_i'(t,\eta)Dx_i(t,\eta) - x_i'(t,\eta)Dp_i(t,\eta) \\ &= -DH(x_i^*(t,\eta))\cdot Dx_i^*(t,\eta) \\ &= -D(H(\cdot,\cdot)\circ x_i^*(\cdot,\cdot))(t,\eta) = 0 \end{aligned}$$

since from Proposition 3.7 and Definition 4.6 it follows that the function $(t,\eta) \mapsto H(x^*(t,\eta)) = 0$ is stratified and constant on $\bar{Y}_i$. From this formula we infer that for any $(t,\eta) \in \bar{Y}_i$, the mapping $s \mapsto p_i(s,\eta)Dx_i(s,\eta)$ is constant on the interval $[\tau_{i-1}(\eta), \tau_i(\eta)]$; hence $p_i(t,\eta)Dx_i(t,\eta) = p_i(\tau_i(\eta),\eta)Dx_i(\tau_i(\eta),\eta)$ for any $(t,\eta) \in \bar{Y}_i$ and therefore from (4.23) and (3.25) (Proposition 3.7) it follows that for any $(\bar{t},\bar{\eta}) \in T_{(t,\eta)}\bar{Y}_i$, $\bar{\eta} = (\bar{x}_F, \bar{p}_F)$ one has: $\langle p_i(t,\eta), Dx_i(t,\eta)\cdot(\bar{t},\bar{\eta})\rangle = \langle p_i(\tau_i(\eta),\eta), \bar{x}_F\rangle$. Since $p(\cdot,\eta)$ is continuous, from (4.17) and (3.18) it follows (4.24) and the Lemma is proved.

THEOREM 4.9. *Let $\Sigma = (X, X_F, U, f, g)$ be a regular optimal control system, let $x_k^*(\cdot,\cdot) = (x_k(\cdot,\cdot), p_k(\cdot,\cdot)) : Y_k \subset (-\infty, 0] \times X_{F,n}^* \to A$ $k \in \mathcal{K}$, be stratified Hamiltonian flows of (3.11) and let Y_r, X_r, $\tilde{X}_r$, $W_r(\cdot)$, $\tilde{Y}_r(\cdot)$, $\tilde{K}_r(\cdot)$ be defined as in (4.7)–(4.13) such that $W_r(\cdot) : \tilde{X}_r \to R$ is continuous and weakly C^1-stratified by $\mathcal{S}_r$ that is compatible with every family $\{x_{ki}(S);\ S \in \mathcal{S}_{ki}\}$, $k \in \mathcal{K}$, $i \in Z_0^k$ in Definition 4.6(ii).*

Then for any $x_0 \in \tilde{X}_r \setminus X_F$ and any $(t,\eta) \in \tilde{Y}_r(x_0)$, $k \in \tilde{K}_r(x_0)$, the mapping $\tilde{x}(\cdot; x_0)$ defined by (4.15) is an optimal trajectory with respect to x_0 for the problem $P_r|\tilde{X}_r$.

If Σ is a Lipschitzian optimal control system and $W_r(\cdot)|(\tilde{X}_r \setminus X_F)$ is locally Lipschitzian, then $\tilde{x}(\cdot; x_0)$ is optimal with respect to x_0 for the problem $P_m|\tilde{X}_r$.

Proof. The statements follow from Theorems 4.3 and 4.4, respectively, if we prove that for any $x \in \tilde{X}_r \setminus X_F$, $(t,\eta) \in \tilde{Y}_r(x)$, $k \in \tilde{K}_r(x)$, (4.14) is verified.

Let us consider $x \in S \in \mathcal{S}_r$ and $\bar{x} \in T_xS = T_x\tilde{X}_r$; since $x_k(t,\eta) = x$, from Definition 4.6 it follows that there exists $i \in Z_0^k$ (i.e. such that the set Y_{ki} defined by (4.19) is not empty) such that $(t,\eta) \in \bar{Y}_{ki}$. Since by Definition 4.6 restriction $x_{ki}(\cdot,\cdot) = x_k(\cdot,\cdot)|\bar{Y}_{ki}$ is a weakly stratified submersion (Definition 2.2), there exists $P \in \mathcal{S}_{ki}$ such that $x_{ki}(P) = S$

and $x_{ki}(\cdot,\cdot)|P$ is locally a projection [15]; hence for any $\bar{x}\in T_xS$ there exists $(\bar{t},\bar{\eta})\in T_{(t,\eta)}P = T_{(t,\eta)}\tilde{Y}_{ki}$ such that:

$$Dx_{ki}(t,\eta)\cdot(\bar{t},\bar{\eta}) = \bar{x}. \tag{4.25}$$

We shall prove next that in this case, if $y_F(\cdot,\cdot)$ is the projection defined in (4.7) then:

$$DW_r(x)\cdot\bar{x} = Dg(y_F(t,\eta))\cdot\bar{x}_F \quad \text{if } \bar{\eta}=(\bar{x}_F,\bar{p}_F). \tag{4.26}$$

Let $\gamma(\cdot):(-a,a)\to P$, $a>0$, be of class C^1 such that: $\gamma(0)=(t,\eta)$ and $\gamma'(0)=(\bar{t},\bar{\eta})$ and let $c(\cdot)$ be defined by: $c(s)=x_{ki}(\gamma(s))$, $s\in(-a,a)$; since $x_{ki}(\cdot,\cdot)|P$ is of class C^1 it follows that $c(\cdot)$ is of class C^1 and from (4.25) it follows: $c(0)=x$, $c'(0)=\bar{x}$ and hence the derivative of $W_r(\cdot)$ is given by: $DW_r(x)\cdot\bar{x}=\lim_{s\to 0}(W_r(c(s))-W_r(x))/s$. Further on, since $(t,\eta)\in\tilde{Y}_r(x)$, from (4.12) and (4.10) it follows that $W_r(c(s))\leqq g(y_F(\gamma(s)))$ for any $s\in(-a,a)$ and $W_r(x)=g(y_F(\gamma(0)))$ and since from the definition of $y_F(\cdot,\cdot)$ ($y_F(t,\eta)=x_F$ if $\eta=(x_F,p_F)$, $t\in R$) and the compatibility condition in Definition 4.6(iv), it follows $\lim_{s\to 0}(\gamma(s)-\gamma(0))/s=(\bar{t},\bar{\eta})$ and $\lim_{s\to 0}(y_F(\gamma(s))-y_F(\gamma(0)))/s=\bar{x}_F$ if $\bar{\eta}=(\bar{x}_F,\bar{p}_F)$, if $s\in(0,a)$ we have:

$$(W_r(c(s))-W_r(x))/s\leqq(g(y_F(\gamma(s)))-g(y_F(t,\eta)))/s\to Dg(y_F(t,\eta))\cdot\bar{x}_F.$$

Hence $DW_r(x)\cdot\bar{x}\leqq Dg(y_F(t,\eta))\cdot\bar{x}_F$. Reasoning in the same way for $s\in(-a,0)$, we obtain the reversed inequality hence (4.26).

To end the proof, it suffices now to note that from (4.24) in Lemma 4.8 applied to the stratified Hamiltonian flow $x_k^*(\cdot,\cdot)$ it follows: $Dg(y_F(t,\eta))\cdot\bar{x}_F=-\langle p_{ki}(t,\eta), Dx_{ki}(t,\eta)\cdot(\bar{t},\bar{\eta})\rangle$ and therefore from (4.25) and (4.26) it follows (4.14) and Theorem 4.9 is proved.

5. An example. The problem of a spacecraft attempting to make a soft landing on the moon using a minimum amount of fuel leads to the following optimal control problem ([7], [10], [16], etc.): *minimize* $(-x_3(t_F(u(\cdot))))$ *subject to*: $x_1'=x_2$, $x_2'=-g_0+u(t)/x_3$, $x_3'=-ku(t)$, $x_1(0)=x_1^0>0$, $x_2(0)=x_2^0\in R$, $x_3(0)=x_3^0\in I=[M,\alpha/g_0)$, $x_1(t_F(u(\cdot)))=x_2(t_F(u(\cdot)))=0$, $x_3(t_F(u(\cdot)))\in I$, $u(t)\in[0,\alpha]$ and $x_1(t)>0$, $x_3(t)\in I$ for any $t\in[0,t_F(u(\cdot)))$, where g_0, M, k, $\alpha>0$ are given constants satisfying the condition: $M<\alpha/g_0$.

Obviously this is a Mayer optimal control problem of the form (3.1)–(3.5) defined by the following elements:

$$X=X_0\cup X_F,\quad X_0=(0,\infty)\times R\times I,\quad X_F=\{0\}\times\{0\}\times I,\quad U=[0,\alpha], \tag{5.1}$$

$$f(x,u)=(x_2,-g_0+u/x_3,-ku),\quad g(0,0,x_3)=-x_3,\quad x_F=(0,0,x_3)\in X_F. \tag{5.2}$$

In view of the "physical" interpretations of the data of the problem it seems reasonable to look for optimal controls in the class of piecewise continuous admissible controls (i.e. to solve the problem P_{cp}) but we shall find piecewise continuous controls that are optimal in the class of measurable bounded admissible controls (i.e. for the problem P_m) solving thus also the problems P_r and P_{cp}.

The Pontryagin function and the Hamiltonian are given by:

$$\mathcal{H}(x,p,u)=p_1x_2-g_0p_2+uh(x,p),\qquad h(x,p)=p_2/x_3-kp_3, \tag{5.3}$$

$$\begin{aligned} &H(x,p)=p_1x_2-g_0p_2+\nu(h(x,p)),\\ &\nu(s)=s \quad \text{if } s>0,\qquad \nu(s)=0 \quad \text{if } s\leqq 0, \end{aligned} \tag{5.4}$$

$$\begin{aligned} &\hat{U}(x,p)=\{\alpha\} \quad \text{if } h(x,p)>0,\qquad \hat{U}(x,p)=\{0\} \quad \text{if } h(x,p)<0,\\ &\hat{U}(x,p)=U \quad \text{if } h(x,p)=0. \end{aligned} \tag{5.5}$$

It is easy to see that the data and the Hamiltonian have all the properties in Definition 3.9 so the optimal control system $\Sigma = (X, X_F, U, f, g)$ is at the same time regular and Lipschitzian. In fact the disjoint components, X_0 and X_F of X are C^ω-stratified by $\{X^1, X^2\}$ and $\{X^3, X^4\}$, respectively, where:

$$\begin{aligned} &X^1 = (0,\infty)\times R\times (M, \alpha/g_0), \qquad X^2 = (0,\infty)\times R\times\{M\},\\ &X^3 = \{0\}\times\{0\}\times(M, \alpha/g_0), \qquad X^4 = \{(0,0,M)\}. \end{aligned} \tag{5.6}$$

The set $A = X\times(R^3\backslash\{0\})$ on which $H(\cdot,\cdot)$ is defined is partitioned by the sets:

$$\begin{aligned} &A_+ = \{(x,p)\in A;\ h(x,p)>0\}, \qquad A_- = \{(x,p)\in A;\ h(x,p)<0\},\\ &A_0 = \{(x,p)\in A;\ h(x,p)=0\} \end{aligned} \tag{5.7}$$

each of them being C^ω-stratified by: $A_+^i = A_+\cap(X^i\times R^3)$, $A_-^i = A_-\cap(X^i\times R^3)$, $A_0^i = A_0\cap(X^i\times R^3)$, $i=1,2,3,4$, respectively.

Remark 5.1. For the verification of the above statements as well as for the computations to follow it is enough to recall that from the implicit functions theorem it follows that if $U\subset R^n$ is open, $F(\cdot): U\to R^m$ $(m<n)$ is of class C^k, $k\in\{1,2,\cdots\infty,\omega\}$, and the derivative $DF(x)\in L(R^n, R^m)$ is surjective for any $x\in U$ then the set $S=\{x\in U;\ F(x)=0\}$, if not empty, is a submanifold of class C^k, of dimension $n-m$, whose tangent space at each point is given by: $T_xS=\{\bar{x}\in R^n;\ DF(x)\cdot\bar{x}=0\}$. From the definition of a differentiable submanifold (e.g. [15]) it follows also that if $F(\cdot): U\to R^m$ is of class C^k, then its graph, $G_{F(\cdot)}=\{(x, F(x));\ x\in U\}$ is a submanifold of class C^k, of dimension n, of $U\times R^m$, whose tangent space is given by: $T_{(x,F(x))}G_{F(\cdot)} = \{(\bar{x}, DF(x)\cdot\bar{x});\ \bar{x}\in R^n = T_xU\}$.

To describe the controlled Hamiltonian orientor field defined by (3.10), we consider first the mapping $\xi(\cdot,\cdot,\cdot)$ defined by (3.28):

$$\xi(x,p,u) = (f(x,u), P(x,p,u)) = ((x_2, -g_0+u/x_3, -ku), (0, -p_1, up_2/(x_3)^2)) \tag{5.8}$$

and note that from (3.10), (5.5) and (5.7) it follows that $\xi_H(x,p)=\xi(x,p,\alpha)$ if $(x,p)\in A_+$ and $\xi(x,p,\alpha)\in T_{(x,p)}A_+$, $\xi_H(x,p)=\xi(x,p,0)$ if $(x,p)\in A_-$ and $\xi(x,p,0)\in T_{(x,p)}A_-$, $\xi_H(x,p)=\{\xi(x,p,u);\ u\in U,\ \xi(x,p,u)\in T_{(x,p)}A_0\}$, if $(x,p)\in A_0$ and $\xi_H(x,p)=\varnothing$ otherwise.

An easy computation using Remark 5.1 shows that we can refine the above-mentioned stratification of A_0 as follows: $A_0^i = A_{0+}^i\cup A_{0-}^i\cup A_{00}^i$ where $A_{0+}^i = \{(x,p)\in A_0^i;\ p_1>0\}$, $A_{0-}^i=\{(x,p)\in A_0^i;\ p_1<0\}$, $A_{00}^i=\{(x,p)\in A_0^i;\ p_1=0\}$, $i=1,2,3,4$, such that the Hamiltonian orientor field is given by:

$$\xi_H(x,p) = \begin{cases} \{\xi(x,p,\alpha)\} & \text{if } (x,p)\in A_+^1,\\ \{\xi(x,p,0)\} & \text{if } (x,p)\in A_-^1\cup A_-^2,\\ \xi(x,p,U) & \text{if } (x,p)\in A_{00}^1\cup A_{00}^2,\\ \varnothing & \text{otherwise.}\end{cases} \tag{5.9}$$

The set $X_{F,n}^*$ of the terminal values for normal characteristics (defined by (3.18)) contains the set $X_{F,a}^*$ (hence the problem is normal) and is given by: $X_{F,n}^* = X_F^* = \cup\{X_{+i}^*\cup X_{-i}^*\cup X_{0+}^*\cup X_{0-}^*;\ i=3,4\}$, where: $X_{+i}^* = \{(x,p)\in A_+^i,\ p_2 = q(x_3,p_3),\ p_3\in P_3^i\}$, $X_{-i}^* = \{(x,p)\in A_-^i;\ p_2=0,\ p_3\in P_3^i\}$, $X_{0+}^* = \{(x,p)\in A_{0+}^4;\ p_2=p_3=0\}$, $X_{0-}^* = \{(x,p)\in A_{0-}^4;\ p_2=p_3=0\}$, $i=3,4$, $P_3^3=\{1\}$, $P_3^4=(0,\infty)$, $q(x_3,p_3)=\alpha k x_3/(\alpha-g_0x_3)$.

Since the Hamiltonian cannot vanish on the sets A_{00}^i, $i=1,2,3,4$, from Proposition 3.7 and (5.9) it follows that for any $\eta=(x_F, p_F)\in X_F^*$ and any characteristic $x^*(\cdot,\eta)$ one has:

$$x^*(t,\eta)\in A_+^1\cup A_-^1\cup A_-^2 \quad \text{a.e. on } [\![\tau(\eta), 0]\!]. \tag{5.10}$$

From (5.7) it follows that the range of any characteristic is determined by the sign of the function $\chi(\cdot, \eta)$ defined by:

$$\chi(t, \eta) = h(x^*(t, \eta)), \qquad t \in [\![\tau(\eta), 0] \tag{5.11}$$

where $h(\cdot, \cdot)$ is defined by (5.3). Since $x^*(\cdot, \eta)$ is absolutely continuous, from (5.3) it follows that $\chi(\cdot, \eta)$ is absolutely continuous and from (5.9) it follows that the derivative of $\chi(\cdot, \eta)$ is given by: $\chi'(t, \eta) = -p_1^F / x_3(t, \eta)$ a.e. on $[\![\tau(\eta), 0]$, $\eta = (x_F, p_F) \in X_F^*$ and therefore, since $x_3(t, \eta) \in I = [M, \alpha/g_0)$, $\chi(\cdot, \eta)$ is a monotone function. It follows that any characteristic of the problem has at most one switching point and is finitely regular with respect to the C^ω-stratification $\mathcal{S}_H = \{A_+^i, A_-^i, A_{0+}^i, A_{0-}^i, A_{00}^i, i = 1, 2, 3, 4\}$.

We shall prove that all the characteristics are "integral curves" of a certain stratified Hamiltonian flow in the sense of Definition 4.6 and Remark 4.7.

From (5.8) and (5.9) it follows that no characteristic may terminate at a point $\eta \in X_{-3}^* \cup X_{-4}^* \cup X_{0+}^*$ since there exists $r > 0$ such that $\chi(t, \eta) < 0$ for any $t \in (-r, 0]$ but, on the other hand, the integral curve of $\xi(\cdot, \cdot, 0)$ through $(0, \eta)$ does not remain in A on the interval $(-r, 0)$; hence we take $\tau(\eta) = 0$ in this case.

From (5.8) and (5.9) it follows also that for any $\eta \in X_{+3}^* \cup X_{+4}^* \cup X_{0-}^*$ there exists a unique characteristic $x^*(\cdot, \eta)$ satisfying (3.21)-(3.22) and $\tau_1(\eta) < 0$ such that

$$x^*(t, \eta) = x_+^*(t, \eta) = (x^+(t, x^F), p^+(t, \eta)) \quad \text{for } t \in (\tau_1(\eta), 0], \quad \eta = (x^F, p^F) \tag{5.12}$$

where $x_+^*(\cdot, \cdot)$ is the flow of the smooth vector field $\xi(\cdot, \cdot, \alpha)$, its first component, $x^+(\cdot, \cdot)$, being the flow of $f(\cdot, \alpha)$ defined by (5.2) and $\tau_1(\eta) = \inf\{t < 0;\ \chi(t, \eta) < 0, x^+(t, x^F) \in X_0\}$. It is easy to see that if either $p_1^F \geqq 0$ or $p_1^F \leqq -\alpha k h(\eta)/\log(\alpha/g_0 x_3^F)$ then:

$$\tau_1(\eta) = T_0(x^F) = (x_3^F - \alpha/g_0)/\alpha k \tag{5.13}$$

and $x^*(\cdot, \eta)$ is not defined at $\tau_1(\eta)$ (since $x^*(t, \eta) \notin A$ for $t \leqq \tau_1(\eta)$); hence we must take $\tau_1(\eta) = \tau(\eta)$ in this case. For the remaining values of $\eta \in X_{+3}^* \cup X_{+4}^*$ (i.e. $-\alpha k h(\eta)/\log(\alpha/g_0 x_3^F) < p_1^F < 0$) we have $\tau_1(\eta) > T_0(x^F)$ and moreover, for every $x^F = (0, 0, x_3^F) \in X_F$ and any $t \in (T_0(x^F), 0)$ there exists $\eta = (x^F, p^F) \in X_{+3}^* \cup X_{+4}^*$ such that $\tau_1(\eta) = t$; further on, in this case $x^*(\cdot, \eta)$ is continuable on an interval $(\tau(\eta), \tau_1(\eta)]$ where it is given by:

$$x^*(t, \eta) = x_-^*(t - \tau_1(\eta), x_+^*(\eta)), \quad x_+^*(\eta) = x_+^*(\tau_1(\eta), \eta), \quad t \in (\tau(\eta), \tau_1(\eta)] \tag{5.14}$$

where $x_-^*(\cdot, \cdot) = (x^-(\cdot, \cdot), p^-(\cdot, \cdot))$ is the flow of the smooth vector field $\xi(\cdot, \cdot, 0)$ defined by (5.8), its first component, $x^-(\cdot, \cdot)$, being the flow of $f(\cdot, 0)$ defined by (5.2) and $\tau(\eta) = \inf\{t < \tau_1(\eta);\ x_1^-(t, x^+(\tau_1(\eta), x^F)) > 0\}$ which gives: $\tau(\eta) = \tau_1(\eta) + T_1(x^+(\eta))$ where: $T_1(x) = (x_2 - ((x_2)^2 + 2g_0 x_1)^{1/2})/g_0$, $x^+(\eta) = x^+(\tau_1(\eta), x^F)$. From these results it follows that any characteristic is an integral curve of the stratified Hamiltonian flow $x^*(\cdot, \cdot)$ defined by (5.12)-(5.14) and $X_m = X_r$, $W_m(\cdot) = W_r(\cdot)$, defined by (4.9) and (4.10), respectively, may be described "parametrically" as follows: $X_m = \{x^-(s - t, x^+(t, x^F));\ x^F \in X_F, t \in (T_0(x^F), 0], s \in (T_1(x^+(t, x^F)), t]\}$, $W_m(x) = -x_3^F$ if $x = x^-(s - t, x^+(t, x^F)) \in \tilde{X}_m = X_m$. Using implicit functions arguments, it follows easily that $\mathcal{S}_m = \{X_m^{i,j};\ i = 3, 4,\ j = 1, 2, 3\}$ where $X_m^{i,1} = X^i$, $X_m^{i,2} = \{x^+(t, x^F);\ x^F \in X^i,\ t \in (T_0(x^F), 0)\}$, $X_m^{i,3} = \{x^-(s - t, x^+(t, x^F));\ x^F \in X^i,\ t \in (T_0(x^F), 0),\ s \in (T_1(x^+(t, x^F)), t)\}$, is a C^1-stratification of $W_m(\cdot)$ which is continuous and its restriction $W_m(\cdot)|(X_m \setminus X_F)$ is locally Lipschitzian.

From Theorem 4.9 and Remark 4.5 it follows now that the function $v(\cdot): X_m \to U$ defined by:

$$v(x) = \begin{cases} \alpha & \text{if } x \in \cup\{X_m^{i,j};\ i=3,4, j=1,2\}, \\ 0 & \text{if } x \in X_m^{3,3} \cup X_m^{4,3} \end{cases}$$

is an optimal feedback control for the problems $P_m|X_m$, $P_r|X_m$ and $P_{cp}|X_m$.

Finally, using a tangency argument [25], [26], one may prove that any admissible trajectory (with respect to the original problem) that starts at a point in $X_m \setminus X_F$ remains on X_m and moreover, the points in $X \setminus X_m$ do not have admissible trajectories; hence $v(\cdot)$ is the optimal feedback control of the original problem.

REFERENCES

[1] V. G. BOLTYANSKII, *Mathematical Methods of Optimal Control*, Nauka, Moscow, 1969; Holt, Rinehart and Winston, New York, 1971.

[2] ———, *The support principle in problems of optimal control*, Diff. Uravn., 9 (1973), pp. 1363-1370. (In Russian.)

[3] N. BOURBAKI, *Fonctions d'une variable réele*, Hermann, Paris, 1958.

[4] A. M. BRUCKNER, *Differentiation of Real Functions*, Springer, Berlin, 1978.

[5] P. BRUNOVSKY, *Every normal linear system has a regular time-optimal synthesis*, Math. Slovaca, 20 (1978), pp. 81-100.

[6] ———, *Existence of regular synthesis for general control problems*, J. Differential Equations, 38 (1980), pp. 317-343.

[7] L. CESARI, *Optimization—Theory and Applications*, Springer, Berlin, 1983.

[8] F. H. CLARKE, *Optimal control and the true Hamiltonian*, SIAM Rev., 21 (1979), pp. 157-166.

[9] R. P. FEDORENKO, *On the Cauchy problem for the Bellman equation of dynamic programming*, Z. Vyc. Math. Math. Phys., 9 (1969), pp. 426-432. (In Russian.)

[10] W. H. FLEMING AND R. W. RISHEL, *Deterministic and Stochastic Optimal Control*, Springer, Berlin, 1975.

[11] C. GODBILLON, *Géométrie différentielle et mécanique analytique*, Hermann, Paris, 1969.

[12] R. M. HARDT, *Stratifications of real analytic mappings and images*, Invent. Math., 28 (1975), pp. 193-208.

[13] H. HERMES, *On continuous and measurable selections and the existence of solutions of generalized differential equations*, Proc. Amer. Math. Soc., 29 (1971), pp. 535-542.

[14] J. L. KELLEY, *General Topology*, D. Van Nostrand, Princeton, NJ, 1957.

[15] S. LANG, *Introduction to Differential Manifolds*, Interscience, New York, 1962.

[16] J. MEDITCH, *On the problem of optimal thrust programming for a lunar soft landing*, IEEE Trans. Automat. Control, 9 (1964), pp. 477-485.

[17] S. MIRICĂ, *On the admissible synthesis in optimal control theory and differential games*, this Journal, 7 (1969), pp. 292-316.

[18] ———, *An algorithm for optimal synthesis in control problems*, Revue Franç. d'Inf. Rech. Op., R-2 (1971), pp. 55-92.

[19] ———, *Stratified Hamiltonians and the optimal feedback control*, Ann. di Mat. Pura Appl., XXXIII (1983), pp. 51-78.

[20] ———, *Stratified Hamiltonian systems*, Workshop in Differential Equations and Control Theory, INCREST-University of Bucharest, Bucharest, 1983, pp. 71-81.

[21] ———, *A generalization of Cauchy's method of characteristics*, Rev. Roumaine Math. Pures Appl., 29 (1984), pp. 863-870.

[22] H. STALFORD, *Sufficiency theorem for discontinuous optimal cost surfaces*, this Journal, 16 (1978), pp. 63-82.

[23] H. SUSSMANN, *Analytic stratifications and control theory*, Proc. ICM, Helsinki, 1978, pp. 865-871.

[24] ———, *Subanalytic sets and feedback control*, J. Differential Equations, 31 (1979), pp. 31-52.

[25] C. URSESCU, *Carathéodory solutions of ordinary differential equations on locally compact sets in Fréchet Spaces*, Preprint Series in Mathematics, "Al. I. Cuza" University of Iaşi, 18/1982, Iaşi, Romania.

[26] J. A. YORKE, *Invariance for ordinary differential equations*, Math. Systems Theory, 1 (1967), pp. 353-372.

SIAM J. CONTROL OPTIM.
Vol. 24, No. 4, July 1986

008

THE GRADIENT PROJECTION METHOD USING CURRY'S STEPLENGTH*

R. R. PHELPS†

Abstract. It is shown that, using Curry's steplength, the gradient projection method for finding constrained stationary points of a real valued C^1 function on a closed convex subset C of Hilbert space can work for two rather different classes of convex sets: those with C^2 boundary and "orthogonal polyhedra". Both results are applications of Theorem 1, whose hypotheses require that the metric projection onto C possess directional derivatives which are continuous in a rather weak sense.

Key words. Hilbert space, gradient projection method, nonlinear optimization, Curry's steplength, orthogonal polyhedra

AMS(MOS) subject classification. 49D

1. Introduction. Let f be a real-valued C^1 function on the real Hilbert space H and suppose that C is a nonempty closed convex subset of H. The gradient projection minimization method is designed to produce a "constrained stationary point" of f in C. The method starts with any point $x_0 \in C$ and proceeds iteratively to define a sequence $\{x_n\}$ in C by

$$x_{n+1} = P[x_n - t_n \nabla f(x_n)], \qquad n = 0, 1, 2, \cdots, \tag{1}$$

where P is the metric projection of H onto C and the *steplengths* $t_n \geqq 0$ are chosen in some specified manner. If $C = H$, then P is just the identity map and one has the method of steepest descent for unconstrained minimization. In this case, that is, when $x_{n+1} = x_n - t_n \nabla f(x_n)$, there are at least two established ways of choosing the steplength. The *Cauchy steplength* is one for which t_n is a global minimum point for the C^1 function $g(t) = f(x_n - t_n \nabla f(x_n))$, $t \geqq 0$. (This assumes, of course, that such a minimum exists.) For the more general *Curry steplength* one chooses t_n to be the smallest stationary point of the same function. An historical account of these unconstrained methods (and a proof that Curry's method is valid in any normed linear space) is given by Byrd and Tapia [1]. Returning to the constrained method (1), one can analogously define the steplengths in terms of the functions

$$g_n(t) = f(P[x_n - t \nabla f(x_n)]), \qquad t \geqq 0. \tag{2}$$

In the case of Cauchy's steplength, the method works: any cluster point of $\{x_n\}$ is a constrained stationary point (defined below) of f in C. This was proved by McCormick and Tapia [7] for Hilbert space and by the author [9] for a reasonable class of Banach spaces. In this note we will examine the more general but more complicated case of Curry's steplength. The complications arise because, even in finite dimensional Euclidean space, P may fail to have directional derivatives (see [6], [10]), so that differentiability of the function g_n is at issue. In fact, even when P is analytic (outside of C), the function g_n may fail to be differentiable precisely at its global minimum point. (See the Example at the end of this paper). In order to avoid this difficulty, we modify slightly the definition of Curry's steplength; rather than defining t_n to be the smallest nonnegative stationary point of g_n we let

$$t_n = \inf\{t \geqq 0 \colon dg_n(t) \geqq 0\}, \tag{3}$$

* Received by the editors January 22, 1985, and in revised form May 17, 1985.

† Department of Mathematics GN-50, University of Washington, Seattle, Washington 98195. This research was supported in part by a grant from the National Science Foundation.

where dg_n denotes the right-hand derivative of g_n. This latter derivative will exist if P admits directional derivatives at every point, something we assume in our main theorem. When g_n is differentiable, definition (3) yields the smallest nonnegative stationary point of g_n. Moreover, it retains an important property of any reasonable steplength:

$$f(x_{n+1}) \leqq f(x_n) \quad \text{for all } n.$$

(To see this, note first that it is equivalent to $g_n(t_n) \leqq g_n(0)$. By definition, if $t_n > 0$, then for $t \in [0, t_n)$ we must have $dg_n(t) < 0$. By a standard monotonicity result (see, for instance, [3, p. 22]), the desired inequality holds.)

In order to say what we mean by a constrained stationary point for f, we first recall that the metric projection always has directional derivatives at points of C. To be precise, if $x \in C$, the *support cone* $S_C(x)$ (or simply $S(x)$) is defined to be the closure of the cone $\cup \{\lambda(C-x): \lambda > 0\}$. It is well-known that if $x \in C$ and $u \in H$, then $dP(x)u = P_{S(x)}u$, where

$$dP(x)u = \lim_{t \to 0^+} t^{-1}[P(x+tu) - P(x)]$$

and $P_{S(x)}$ is the metric projection of H onto $S(x)$. (See, for instance, [7] or [11, p. 300].) By combining this result with the chain rule one can compute the right-hand derivative at $t = 0$ of

$$g(t) = f(P[x - t\nabla f(x)]), \qquad t \geqq 0$$

whenever $x \in C$; since $Px = x$ we obtain

$$dg(0) = \langle \nabla f(x), P_{S(x)}[-\nabla f(x)] \rangle = \langle \nabla f(x), dP[x](-\nabla f(x)) \rangle.$$

As shown in the corollary in [9], this latter quantity equals $-\|P_{S(x)}[-\nabla f(x)]\|^2$. (Note that, in particular, we always have $dg(0) \leqq 0$.) We say that $x \in C$ is a *constrained stationary point* for f in C provided $P_{S(x)}[-\nabla f(x)] = 0$. This is obviously equivalent to requiring $dg(0) \geqq 0$. Our first result asserts that any cluster point x^* of the sequence (1) (using the steplength (3)) will be a constrained stationary point for f, provided dP exists in H and is continuous in a certain weak sense. Propositions 2 and 3 exhibit classes of sets C for which these latter hypotheses are satisfied.

As above, we define (when the limit exists)

$$dP(x)u = \lim_{t \to 0^+} t^{-1}[P(x+tu) - P(x)], \qquad x, u \in H.$$

The weak continuity property we use is the following. (It assumes that dP exists.)

(WSC) Suppose that $\{x_n\} \subseteq C$, $x \in C$ and $\|x_n - x\| \to 0$. Assume also that $\{u_n\} \subseteq H$, $t_n \to 0^+$, $x_n - t_n u_n \in H \backslash \operatorname{int} C$, $-u \in H \backslash \operatorname{int} S(x)$ and $\|u_n - u\| \to 0$. Then $\limsup_{n \to \infty} \langle u_n, dP[x_n - t_n u_n](-u_n) \rangle \leqq \langle u, dP[x](-u) \rangle$.

This property (which is more like a semi-continuity property) is clearly satisfied if $\{dP[x_n - t_n u_n](-u_n)\}$ converges weakly to $dP[x](-u)$, something which is true (Proposition 2) whenever C has a C^2 boundary. Weak convergence can fail, however, for the orthogonal polyhedra of Proposition 3, while the (WSC) holds.

2. Main theorem.

THEOREM 1. *Let C be a nonempty closed convex subset of the Hilbert space H and let P denote the metric projection of H onto C. Suppose that the directional derivative dP exists in H and satisfies the (WSC). If f is a real-valued C^1 function on H and $x_0 \in C$, inductively define g_n, t_n and x_n as in (1), (2) and (3). Then any cluster point x^* of the sequence $\{x_n\}$ is a constrained stationary point for f.*

Proof. If we define $g(t) = f(P[x^* - t\nabla f(x^*)])$, $t \geqq 0$, then the foregoing discussion shows that we want to prove that $dg(0) \geqq 0$. Let $\{x_i\}$ be a subsequence of $\{x_n\}$ converging to x^* and let t_i be the corresponding steplength used to define the successor to x_i. Let $t^* = \inf\{t_i\}$. Following a standard approach, we consider two cases.

Case 1. $t^* > 0$. Arguing by contradiction, suppose that $dg(0) < 0$. This implies that there exists $0 < \tau < t^*$ such that $\tau^{-1}[g[\tau) - g(0)] < 0$, that is,

$$f(P[x^* - \tau\nabla f(x^*)]) < f(P[x^*]) = f(x^*).$$

By continuity and the fact that x^* is the limit of $\{x_i\}$, there exists x_N such that

$$f(P[x_N - \tau\nabla f(x_N)]) < f(x^*).$$

Since $0 < \tau < t^* \leqq t_N$ and since, by definition of the steplength t_N, we have $dg_N(t) < 0$ for $\tau \leqq t < t_N$, the monotonicity of g_N in this latter interval implies that $g_N(t_N) \leqq g_N(\tau)$. This means that

$$f(x^*) > f(P[x_N - \tau\nabla f(x_N)]) = g_N(\tau) \geqq g_N(t_N) = f(x_{N+1}).$$

But $i > N+1$ for all but finitely many i, hence $f(x_i) \leqq f(x_{N+1}) < f(x^*)$. Since $f(x_i) \to f(x^*)$, this is a contradiction which completes the proof of Case 1.

Case 2. $t^* = 0$. Assume without loss of generality that $t_i \to 0$. By definition of t_i we can choose s_i for each i such that $t_i < s_i$, $dg_i(s_i) \geqq 0$ and $s_i \to 0$. Of course, $x_i \to x^*$. We consider several possibilities. Let

$$y_i = x_i - s_i\nabla f(x_i)$$

and suppose, first that $y_i \in \operatorname{int} C$ for infinitely many i. Since P is the identity map in C, this implies that

$$dP[y_i](-\nabla f(x_i)) = -\nabla f(x_i)$$

for infinitely many i. Also, $\nabla f(P[y_i]) = \nabla f(y_i)$. Since, for every i,

$$(\alpha) \qquad 0 \leqq dg_i(s_i) = \langle \nabla f(P[y_i]),\, dP[y_i](-\nabla f(x_i)) \rangle$$

we conclude that for infinitely many i the right side of (α) is $\langle \nabla f(y_i), -\nabla f(x_i) \rangle$, which converges to $\langle \nabla f(x^*), -\nabla f(x^*) \rangle = -\|\nabla f(x^*)\|^2$, and hence $\nabla f(x^*) = 0$.

Suppose, next, that $-\nabla f(x^*) \in \operatorname{int} S(x^*)$. This implies that $x^* - \tau\nabla f(x^*) \in \operatorname{int} C$ for some $\tau > 0$ and therefore $x_i - \tau\nabla f(x_i) \in \operatorname{int} C$ for all sufficiently large i. Since $s_i < \tau$ for all but finitely many i, the convexity of C implies that $y_i \equiv x_i - s_i\nabla f(x_i) \in \operatorname{int} C$ whenever $s_i < \tau$, which leads, as above, to $\nabla f(x^*) = 0$. Thus, we are reduced to examining the case where $-\nabla f(x^*) \in H \backslash \operatorname{int} S(x^*)$ and $y_i \in H \backslash \operatorname{int} C$ for all but finitely many i. (Assume that this holds for all i.) Since $y_i \to x^*$ and $x_i \to x^*$ we have $P[y_i] \to P[x^*] = x^*$ and $\nabla f(P[y_i]) - \nabla f(x_i) \to 0$. The sequence $\{dP[y_i](-\nabla f(x_i))\}$ is bounded, so $\langle \nabla f(x_i) - \nabla f(P[y_i]),\, dP[y_i](-\nabla f(x_i)) \rangle \to 0$. It follows from this and the inequality (α) above that

$$\limsup \langle \nabla f(x_i),\, dP[y_i](-\nabla f(x_i)) \rangle \geqq 0$$

and hence, by the (WSC), $dg(0) = \langle f(x^*),\, dP[x^*](-\nabla f(x^*)) \rangle \geqq 0$, which completes the proof.

The foregoing theorem can be formulated and proved in a more general class of Banach spaces, using the same definitions and hypotheses as were used in [9]. We have refrained from proving the more general result since our main application is to Hilbert space (via Proposition 2 below). (While Proposition 3 below could be proved in any l_p space, it applies to rather special subsets.)

3. Smooth bodies. Our next proposition shows that dP has the (WSC) if the boundary of C is sufficiently smooth; this is equivalent to smoothness of the Minkowski (or gauge) functional associated with C [2], [4].

PROPOSITION 2. *Suppose that C is a closed convex subset of Hilbert space H with $0 \in \operatorname{int} C$ and suppose that the Minkowski functional μ for C is of class C^2 (at nonzero points). Then the metric projection P of H onto C satisfies the following properties:*

(i) *The directional derivative $dP(x)u = \lim_{t \to 0^+} t^{-1}[P(x+tu) - P(x)]$ exists for all x, $u \in H$.*

(ii) *If $x \in C$, $\{y_n\} \subseteq H \backslash \operatorname{int} C$ and $\|y_n - x\| \to 0$, then $dP[y_n]u \to dP[x]u$ weakly for each $u \in H \backslash \operatorname{int} S_C(x)$.*

Proof. If $x \in \operatorname{int} C$, then the first assertion follows trivially from the fact that P is the identity mapping in C. If $x \in \partial C$, then it is well known (see McCormick and Tapia [7] or Zarantonello [11, p. 300]) that $dP(x)$ exists and is given by the metric projection $P_{S(x)}$ of H onto the support cone $S(x)$ of C at x. (In the present case, $S(x)$ is the translate to the origin of the closed half-space which contains C and supports it at x.) Finally, if $x \in H \backslash C$, then assertion (i) follows from Holmes' result [5] that since μ is C^2, we know that P is C^1 in $H \backslash C$. (See [2] for related results.) Thus, we need only prove assertion (ii), which implicitly assumes that $x \in \partial C$.

Since P is nonexpansive, we have $\|dP[y_n]u\| \leqq \|u\|$ for all n, so by the relative weak compactness of bounded sets, it suffices to prove that every weakly convergent subsequence of $\{dP[y_n]u\}$ has the same limit $w \in H$ and that $w = dP[x]u$. Without loss of generality we can assume that $\{dP[y_n]u\}$ converges weakly to w. Since the form of $dP[y_n]$ depends on whether $y_n \in C$, we consider first the case that y_n *is* in C—that is, $y_n \in \partial C$—for all but finitely many n. As we noted above, this implies that $dP[y_n]u = P_{S(y_n)}u$. Letting $\nabla\mu$ denote the gradient of μ (which exists at each point of ∂C), we have

$$S(y_n) = \{z \in H : \langle \nabla\mu(y_n), z\rangle \leqq 0\}.$$

We want, of course, to show that $w = P_{S(x)}u$. Suppose, first, that $\langle \nabla\mu(y_n), u\rangle > 0$ for infinitely many n. Then $\langle \nabla\mu(x), u\rangle \geqq 0$ and for infinitely many n the vector $P_{S(y_n)}u$ is the orthogonal projection of u onto the hyperplane

$$H(y_n) = \{z \in H : \langle \nabla\mu(y_n), z\rangle = 0\}.$$

For each such n we have the orthogonal decomposition

$$u = P_{S(y_n)}u + \langle \nabla\mu(y_n), u\rangle \|\nabla\mu(y_n)\|^{-2} \nabla\mu(y_n).$$

Since $\nabla\mu(y_n) \to \nabla\mu(x) \neq 0$ we conclude that u also has the decomposition

$$u = w + \langle \nabla\mu(x), u\rangle \|\nabla\mu(x)\|^{-2} \nabla\mu(x);$$

that is, $w = P_{S(x)}u$. Suppose, therefore, that $\langle \nabla\mu(y_n), u\rangle \leqq 0$ for all but finitely many n. It follows that $\langle \nabla\mu(x), u\rangle \leqq 0$ and $u \in S(y_n)$ for these n, so that $P_{S(y_n)}u = u$ and hence $u = w = P_{S(x)}u$.

We can thus assume that $y_n \in H \backslash C$ for infinitely many n. (For simplicity of notation, assume that it holds for all n.) We only have an implicit description of $dP[y]$ at points $y \in H \backslash C$; it is obtained from the identity

$$(\alpha) \qquad \langle y - Py, Py\rangle \nabla\mu(Py) = y - Py.$$

(See, for instance, [2] for a derivation of this relation. Always $\langle y - Py, y\rangle > 0$.) We can differentiate both sides of (α) in the direction $u \in H$ to obtain

$$(\beta) \qquad \begin{aligned} &\{\langle u - dP[y]u, Py\rangle + \langle y - Py, dP[y]u\rangle\} \nabla\mu(Py) + \langle y - Py, Py\rangle \, d\nabla\mu(Py)\{dP[y]\} \\ &\qquad = u - dP(y)u. \end{aligned}$$

If we now let $y = y_n$ and take limits as $n \to \infty$, we have $y_n - Py_n \to x - Px = 0$ and $dP[y_n]u \to w$ weakly; since μ is C^2 and P is C^1, we conclude that

$$(\gamma) \qquad \langle u - w, x\rangle \nabla\mu(x) = u - w, \qquad u \in H.$$

Now, it is known (see, again, [2]) that $\langle y - Py, dP[y]u\rangle = 0$ whenever $y \in H \setminus C$; from (α), this implies that $\langle \nabla\mu(Py), dP[y]u\rangle = 0$. Taking $y = x_n$, letting $n \to \infty$ and using the norm continuity of $\nabla\mu$ at boundary points of C, this yields $\langle \nabla\mu(x), w\rangle = 0$. We are interested in showing that $w = P_{S(x)}u$ whenever $u \in H \setminus \operatorname{int} S(x)$, that is, whenever $\langle \nabla\mu(x), u\rangle \geqq 0$. Certainly w will be the metric projection of u onto $S(x)$ if it is the metric projection $P_{H(x)}u$ of u onto $H(x) = \{z \in H: \langle \nabla\mu(x), z\rangle = 0\}$. We have just shown that $w \in H(x)$. Furthermore, if $z \in H(x)$, then from (γ) we obtain

$$\langle u - w, z - w\rangle = \langle u - w, x\rangle \langle \nabla\mu(x), z - w\rangle = 0$$

so w satisfies the defining inequality for $P_{H(x)}u$ and the proof is complete.

The fact that conclusion (ii) above implies (WSC) follows easily from the facts that $\|u - u_n\| \to 0$ and $\|dP[y]u_n - dP[y]u\| \leqq \|u - u_n\|$ for any y.

A word about the hypothesis that the boundary of C be C^2 smooth: it guaranteed that P would be C^1 in $H \setminus C$, while the *statement* of Proposition 2 requires merely that P have directional derivatives at each point. It is natural to suspect that a weaker smoothness hypothesis on the boundary of C would suffice for this, but C^1 smoothness is not enough. Zamfirescu [10] has shown that, for almost every (in the sense of Baire category, using the Hausdorff metric on compact convex sets) smooth convex body C in Euclidean n-space E, there is a dense G_δ set of points in $E \setminus C$ at which the associated metric projection fails to have a directional derivative (in at least one direction). (This is a very substantial improvement on Kruskal's well-known example [6].) Recall that for convex bodies in finite dimensional spaces, *smoothness*, that is, a unique supporting hyperplane at each boundary point, is equivalent to C^1 smoothness.

4. Orthogonal polyhedra. There is another class of convex sets C which contrasts sharply with the above; typical examples in the plane would be a rectangle with sides parallel to the coordinate axes, or a cone obtained by translating one of the four quadrants. Following McCormick and Tapia [7], we will call them orthogonal polyhedra. They can be defined in any $l_p(A)$ space (for an arbitrary set A and $1 \leqq p < \infty$) as follows.

For each $\alpha \in A$, let a_α be a real number or $-\infty$ and assume that the function a^+ (defined on A by $a_\alpha^+ = a_\alpha$ if $a_\alpha > 0$, $=0$ otherwise) is an element of $l_p(A)$. Similarly, let b_α be a real number or $+\infty$ and assume that $b^- \in l_p(A)$. Finally, suppose that $a_\alpha \leqq b_\alpha$ for each $\alpha \in A$. Define

$$I(a, b) = \{x \in l_p(A): a_\alpha \leqq x_\alpha \leqq b_\alpha \text{ for each } \alpha \in A\}.$$

It is obvious that $I(a, b)$ is closed, nonempty and convex. For each $\alpha \in A$ and $x \in l_p(A)$ we define

$$p_\alpha(x) = \min\{b_\alpha, \max\{a_\alpha, x_\alpha\}\};$$

this is a number between a_α and b_α which equals x_α if the latter is between these two limits (and equals one of the endpoints otherwise). It is elementary to verify that $Px \equiv (p_\alpha(x))_{\alpha \in A}$ is an element of $l_p(A)$; this uses the l_p-summability restrictions on a and b (and is equivalent to the latter). Moreover, *Px is the nearest point in $I(a, b)$ to x.* (For the uniqueness part of this assertion we require $1 < p < \infty$.) Indeed, one can

easily check cases to verify that if $y \in I(a, b)$, then for all $\alpha \in A$,

$$|x_\alpha - (Px)_\alpha| \equiv |x_\alpha - p_\alpha(x)| \leqq |y_\alpha - x_\alpha|$$

and hence $\|x - Px\|_p \leqq \|y - x\|_p$.

PROPOSITION. 3. *If $C = I(a, b) \subseteq l_p(A)$ (with a, b as above), then $dP[x]u$ exists for all x and u and (in the case $p = 2$) dP satisfies the* (WSC).

Proof. Given x, $u \in l_p(A)$ and $\alpha \in A$, let

$$dp_\alpha(x)u = \lim_{t \to 0^+} [p_\alpha(x + tu) - p_\alpha(x)]/t.$$

This limit exists; indeed, with some patience, it can easily be computed and is described as follows: First, one sees that $dp_\alpha(x)u = 0$ under the following conditions: If $a_\alpha = b_\alpha$, or if $x_\alpha > b_\alpha$ or $x_\alpha < a_\alpha$ or $x_\alpha = b_\alpha$ (and $u_\alpha > 0$) or $x_\alpha = a_\alpha$ (and $u_\alpha < 0$). In the remaining cases we have

$$(*) \qquad dp_\alpha(x)u = u_\alpha \text{ provided } a_\alpha < x_\alpha < b_\alpha \text{ or } a_\alpha < x_\alpha = b_\alpha \text{ (and } u_\alpha < 0) \text{ or } a_\alpha = x_\alpha < b_\alpha \text{ (and } u_\alpha > 0).$$

Since $|dp_\alpha(x)u| \leqq |u_\alpha|$, it is clear that $(dp_\alpha(x)u)_{\alpha \in A} \in l_p(A)$. By reviewing the above case-by-case computations, one obtains a bit more information, namely, that for $t > 0$

$$t^{-1}|p_\alpha(x + tu) - p_\alpha(x)| \leqq |u_\alpha| \quad \text{for each } \alpha.$$

Thus, for each α,

$$|t^{-1}[p_\alpha(x + tu) - p_\alpha(x)] - dp_\alpha(x)u|^p \leqq 2^p |u_\alpha|^p;$$

by dominated convergence, we conclude that

$$\|t^{-1}[P(x + tu) - P(x)] - (dp_\alpha(x)u)\|_p \to 0 \quad \text{as } t \to 0^+,$$

which means that *$dP[x]u$ exists and equals $(dp_\alpha(x)u)$ for each x, $u \in l_p(A)$.*

We now prove that, for $p = 2$, the directional derivative dP satisfies (WSC). Suppose, then, that $\{x^n\} \subseteq I(a, b)$ and $x \in I(a, b)$, that $-u \in l_2(A) \backslash \operatorname{int} S_I(x)$ and $\{u^n\} \subseteq l_2(A)$, and that $\|x^n - x\| \to 0$, $\|u^n - u\| \to 0$ and $t_n \to 0^+$. (We need not assume that $x^n - t_n u^n \in l_2(A) \backslash \operatorname{int} I(a, b)$.) We want to show that

$$(**) \qquad \limsup_{n \to \infty} \langle u^n, dP[x^n - t_n u^n](-u^n) \rangle \leqq \langle u, dP[x](-u) \rangle.$$

We use the coordinate-wise description of dP implicit in (*) to compute the relevant quantities. For instance, $\langle u, dP[x](-u) \rangle = -\sum (u_\alpha)^2$, where the sum is taken over the subset B of those indices $\alpha \in A$ for which x_α satisfies the inequalities described in (*). To compute, for each n, the inner-product term in the left side of (**), note that the α-th term is either zero or $-(u_\alpha^n)^2$, the later holding provided either (letting $y_\alpha^n = x_\alpha^n - t_n u_\alpha^n$) we have $a_\alpha < y_\alpha^n < b_\alpha$ or $a_\alpha < y_\alpha^n = b_\alpha$ (and $-u_\alpha^n < 0$) or $a_\alpha = y_\alpha^n < b_\alpha$ (and $-u_\alpha^n > 0$). The equalities in the latter two sets of inequalities cannot be satisfied: Since $x^n \in I(a, b)$, we have, for instance, $x_\alpha^n \leqq b_\alpha$, so if $-u_\alpha^n < 0$, then $y_\alpha^n < b_\alpha$. Similarly, we necessarily have $y_\alpha^n > a_\alpha$ whenever $-u_\alpha^n > 0$. Thus,

$$\langle u^n, dP[x^n - t_n u^n](-u^n) \rangle = -\sum (u_\alpha^n)^2,$$

where the sum is taken over the subset $B_n \subseteq A$ of all those α for which $a_\alpha < y_\alpha^n < b_\alpha$. Since $\|u^n - u\| \to 0$ and $\|x^n - x\| \to 0$ we have $u_\alpha^n \to u_\alpha$ and $x_\alpha^n \to x_\alpha$ for each α. Consider an index $\alpha \in B$. If $a_\alpha < x_\alpha < b_\alpha$, then since $t_n \to 0^+$, there exists n_α such that $a_\alpha < y_\alpha^n < b_\alpha$ for $n \geqq n_\alpha$, that is, $\alpha \in B_n$ for all sufficiently large n. If $a_\alpha < x_\alpha = b_\alpha$ and $-u_\alpha < 0$, then $-u_\alpha^n < 0$ for all sufficiently large n and $x_\alpha^n \leqq b_\alpha$ for all n, so again, $\alpha \in B_n$ for all

sufficiently large n. The same conclusion holds if $a_\alpha = x_\alpha < b_\alpha$ and $-u_\alpha > 0$. By using these computations and multiplying both sides of (**) by a minus sign, the latter becomes

$$(***) \qquad \sum_{\alpha \in B} (u_\alpha)^2 \leqq \liminf_{n\to\infty} \sum_{\alpha \in Bn} (u_\alpha^n)^2,$$

where $u_\alpha^n \to u_\alpha$ for all α and each $\alpha \in B$ is eventually in B_n. These last two conditions imply that, pointwise on the set A, we have

$$u^2\chi \leqq \liminf_{n\to\infty} (u^n)^2\chi_n,$$

where χ is the characteristic function of the set B and χ_n is the characteristic function of the set B_n. The inequality (***) is now seen to be a consequence of Fatou's lemma.

If, for each $\alpha \in A$, we let $b_\alpha = +\infty$, then we obtain the result for orthogonal polyhedra formulated by McCormick and Tapia [7]. (Actually, they chose an orthogonal system $\{\delta_\alpha: \alpha \in A\}$ in an arbitrary Hilbert space on which to base their polyhedra, but their version is canonically equivalent to the one just described.) One can obviously extend Proposition 3 by using the fact that if T is a linear isometry of $l_2(A)$ onto itself (or onto any other Hilbert space H), then the result is valid for the image $C = T[I(a, b)]$, since $P_C = TP_IT^{-1}$.

We remarked at the end of the introduction that, for orthogonal polyhedra, dP need not have the norm-to-weak continuity property exhibited in Proposition 2(ii) (which is valid for C^2 smooth bodies). This is shown by the following simple example: In l_2, let $C = I(a, b)$ be the positive cone (that is, $a \equiv 0$ and $b \equiv +\infty$); this has empty interior. From the description of dP given in the proof of Proposition 3, one sees readily that if $u \in l_2$ with $u_\alpha < 0$ for all α and if $\{y^n\}$ is a sequence which converges in norm to $x = 0$ and for which $y_\alpha^n > 0$ for all n and α, then for all α, we have $dp_\alpha(y^n)u = u_\alpha$ while $dp_\alpha(x)u = 0$. Thus, $x \in C$, $y^n \in l_2 \backslash \text{int } C$ and $u \in l_2 \backslash S_C(x)$ (this latter holding since $S_C(x) = C$), but $dP[y^n]u = u$ while $dP[x]u = 0$.

Proposition 3 is formulated and proved for $l_p(A)$ rather than for L_p (over a finite measure space, say) partly because this avoids the complications of working with equivalence classes of functions and equivalence classes of measurable sets. The proof given above for the existence of dP goes through without change in any L_p space, while the proof of the (WSC) property can be mimicked to provide a proof in L_2. Mignot [8] has proved the existence of directional derivatives for P_I for sets $I = \{f: a \leqq f \leqq b\}$ in certain Hilbert spaces of analytic functions. He has, in fact, obtained some fundamental results on metric projections and their directional derivatives in the more general context of Hilbert spaces provided with a nonsymmetric inner product. This has allowed him to give a number of applications to questions of optimal control.

5. An example. The following simple two-dimensional example shows that the composite function $g(t)$ may fail to be differentiable precisely at its unique minimum point in $t \geqq 0$, even when the set C has analytically smooth boundary.

Example. Let $C = \{(x, y) \in R^2: x^2 + y^2 \leqq 1\}$, let $f(x, y) = x^2y$ and let $\mathbf{x}^0 = (\frac{1}{2}, 0) \in C$. As usual, for $t \geqq 0$ define

$$g(t) = f(P_C[\mathbf{x}^0 - t\nabla f(\mathbf{x}^0)]).$$

Then $g(t)$ attains its minimum at $t = 2\sqrt{3}$, but is not differentiable there.

Proof. Since $\nabla f(x, y) = (2xy, x^2)$, we have $\mathbf{x}^0 - t\nabla f(\mathbf{x}^0) = (\frac{1}{2}, -t/4)$. For $0 \leqq t \leqq 2\sqrt{3}$, this is in C (where P_C is the identity), so $g(t) = -t/16$. For points x outside of C, we have $P_C(\mathbf{x}) = \mathbf{x}/\|\mathbf{x}\|$, so (with a little computation) we see that $g(t) = -4t(4 + t^2)^{-3/2}$ if $t \geqq 2\sqrt{3}$. It is readily calculated that $g'(t) > 0$ for $t \geqq 2\sqrt{3}$, while $g'(t) = -1/16$ for

$0 \leqq t \leqq 2\sqrt{3}$. (We take the appropriate one-sided derivatives at $t = 2\sqrt{3}$.) Thus, g has its minimum at $2\sqrt{3}$ but is not differentiable at this point. Note that the boundary of C is analytic, as is f; it is the nondifferentiability of P at the boundary which causes problems.

The referee has called our attention to the interesting paper of Gafni and Bertsekas [4], which presents a different and more complicated version of the gradient projection method. The complications are offset by the fact that any limit point of the sequence $\{x_n\}$ is a constrained stationary point, for *arbitrary* closed convex sets C. Briefly described, the differences are as follows: First, in the definition of x_{n+1}, the negative gradient $-\nabla f(x_n)$ is replaced by a vector g_n which is a somewhat complicated perturbed version of the latter. Second, an Armijo-like steplength rule is used which guarantees that $f(x_{n+1}) < f(x_n)$. There is considerable latitude in the choice of g_n: It is easy to construct specific examples for which Curry's steplength converges to an optimum in one iteration, while their method will either take an infinite sequence of iterations or, if g_1 is chosen carefully, a single iteration. They show that for the case when C is a polyhedral cone in a finite dimensional space, ∇f is locally Lipschitzian and f has a strict local minimum, then it is possible to guarantee superlinear convergence to the minimum point.

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] S. FITZPATRICK AND R. R. PHELPS, *Differentiability of the metric projection in Hilbert space*, Trans. Amer. Math. Soc., 270 (1982), pp. 483-501.

[3] T. M. FLETT, *Differential Analysis*, Cambridge Univ. Press, Cambridge, England, 1980.

[4] E. M. GAFNI AND D. P. BERTSEKAS, *Two-metric projection methods for constrained optimization*, this Journal, 22 (1984), pp. 936-964.

[5] R. B. HOLMES, *Smoothness of certain metric projections on Hilbert space*, Trans. Amer. Math. Soc., 184 (1973), pp. 87-100.

[6] J. KRUSKAL, *Two convex counterexamples: a discontinuous envelope function and a nondifferentiable nearest point mapping*, Proc. Amer. Math. Soc., 23 (1969), pp. 697-703.

[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 variationnelles elliptiques*, J. Funct. Anal., 22 (1976), pp. 130-185.

[9] R. R. PHELPS, *Metric projections and the gradient projection method in Banach spaces*, this Journal, 23 (1985), pp. 973-977.

[10] T. ZAMFIRESCU, private communication.

[11] E. H. ZARANTONELLO, *Projections on convex sets in Hilbert space and spectral theorem*, Contributions to Nonlinear Functional Analysis, Publ. No. 27, Math. Res. Center. Univ. Wisconsin, Madison, Academic Press, New York-London, 1971, pp. 237-424.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 4, July 1986

009

AN RKH SPACE APPROACH TO STATE FEEDBACK CONTROL FOR A CLASS OF LINEAR STOCHASTIC SYSTEMS*

R. B. MINTON† AND J. A. RENEKE‡

Abstract. Necessary and sufficient conditions are given for the solution of a linear-quadratic stochastic control problem, with the optimal control written explicitly as a function of the output. System dynamics are of the state-space form $h = \alpha Ch + z + u$, $h(0) = 0$, where h is an observable output, α is a constant matrix, C represents antiderivative, z is a random disturbance (e.g., Wiener process), and the control function $u = Dh$ for some Volterra integral operator D. A standard quadratic cost functional is defined in terms of a Hellinger integral, and the corresponding reproducing kernel Hilbert space is utilized to solve directly for the regulator

Key words. stochastic control theory, state-space system, reproducing kernel, Hellinger integral

AMS(MOS) subject classifications. 49, 60, 93

1. Introduction. In a linear-quadratic-Gaussian control problem, the goal is to find a control function which minimizes a quadratic cost functional subject to linear system dynamics with Gaussian noise and specified initial and terminal conditions. As discussed below, classical results in the literature on linear-quadratic-Gaussian problems require the solution of a matrix Riccati differential equation for complete determination of the optimal control law. In this paper, a necessary condition for solution of a general problem is found which gives the control function as an explicit, and uncomplicated, function of the output.

The function and operator spaces utilized in this paper allow consideration of a general linear system $h = Bh + u + z$ on a finite time interval $[-r, T]$, where B is a Volterra integral operator possibly containing delay and state-dependent noise terms, u is the control function to be determined, and z is a random disturbance. However, to facilitate the introduction of the "operator approach" being used in this paper, system dynamics will be restricted to the state-space format

$$h(t) = \int_0^t \alpha h(s)\, ds + u(t) + z(t), \qquad t \in [0, T]. \tag{1.1}$$

Here, h is an n-dimensional, directly observable output with $h(0) = 0$, and α is a constant $n \times n$ matrix. The control u is required to have the form

$$u(t) = [Dh](t), \tag{1.2}$$

where D is a Volterra integral operator. The disturbance $z(t)$ is a mean-zero second-order stochastic process with independent increments (e.g., n-dimensional Brownian motion). The cost functional J to be minimized has a standard form

$$J(u, h) = \frac{1}{2} E \int_0^T \left\{ c_1 h^*(t) h(t) + c_2 \frac{du^*}{dt}(t) \frac{du}{dt}(t) \right\} dt + \frac{1}{2} Eh^*(T) h(T), \tag{1.3}$$

where c_1 and c_2 are scalar constants and E denotes the expectation operator. The derivative du/dt appears in the cost functional because the control u is being added to the integral equation (1.1) rather than to the corresponding differential form (1.4).

* Received by the editors January 31, 1984, and in revised form June 12, 1985.

† Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061.

‡ Department of Mathematical Sciences, Clemson University, Clemson, South Carolina 29631.

A problem closely related to the control problem (1.1)-(1.3) is the so-called "linear regulator" problem. In this problem, system dynamics is described by

$$dx = [A(t)x(t) + B(t)u(t)]\,dt + \sigma(t)\,dW, \tag{1.4}$$

where $W(t)$ is a Wiener process. Solutions of (1.4) are interpreted as solutions of the corresponding stochastic (Ito) integral equation, since W is not of bounded variation. The cost functional to be minimized is

$$J = E\int_0^T \{x^*(t)M(t)x(t) + u^*(t)N(t)u(t)\}\,dt + Ex^*(T)Dx(T), \tag{1.5}$$

for positive definite matrices M, N, and D. This functional is similar to, but more general than, the cost functional (1.3). The optimal control can be found by the method of dynamic programming [1]. For the relatively simple problem (1.4)-(1.5) the partial differential equation describing the dynamic programming solution implies that the optimal control is given by

$$u(t) = -N(t)^{-1}B^*(t)K(t)x(t), \tag{1.6}$$

where $K(t)$ satisfies the matrix Riccati equation

$$K(t) = -K(t)A(t) - A^*(t)K(t) + K(t)B(t)N(t)^{-1}B^*(t)K(t) - M(t), \tag{1.7}$$
$$K(T) = D.$$

Because K depends on a terminal condition $K(T) = D$, the control (1.6) is not an admissible control in the sense of (1.2). To compute the control (1.6), it is necessary to solve the differential equation (1.7) separately for each t. Some aspects of the relationship between the above result and the result obtained in this paper are examined in an example presented in § 5.

The necessity of solving the auxiliary equation (1.7) provides part of the motivation for reworking the linear regulator problem. The control problem (1.1)-(1.3) is stated in terms of carefully chosen spaces of integral operators and the reproducing kernel Hilbert space induced by a Hellinger integral, and a necessary condition for the optimal control is found directly. These nonstandard spaces also allow consideration of control problems which are not currently in the literature, thus providing further motivation for the study of this approach. Similar spaces have been successfully employed to solve various problems involving deterministic hereditary systems, including system identification [2], parameter estimation [3], optimal control on a finite interval [4], and control of large-scale systems [5].

The spaces of integral operators are defined in § 2 of this paper. A Hellinger integral and its associated Hilbert space and reproducing kernel are introduced in § 3. The control problem (1.1)-(1.3) is solved in § 4. A necessary and sufficient condition for the minimum is given in Theorem 4.1, as well as a necessary condition giving the form of the optimal control. An example is given in § 5 to compare the controls from Theorem 4.1 and from the linear regulator problem.

2. Properties of the operator spaces. The following notation is used throughout the paper. Some definitions are stated in more generality than is required in this paper; in these cases, comments will be given to indicate the nature of the generality.

The underlying probability space is (Ω, F, P), where Ω is a set, F is a σ-algebra of subsets of Ω, and P is a probability measure on (Ω, F). The space $L_2(\Omega)$ contains all square-integrable (with respect to P) functions on Ω. The space $L_{2,n}(\Omega)$ of n-dimensional random variables consists of all functions from Ω into n-space for which

each component belongs to $L_2(\Omega)$. An inner product of elements $f(t)$ and $g(s)$ of $L_{2,n}(\Omega)$ is defined by

$$\langle f(t), g(s)\rangle = \int_\Omega \{f_1(t)g_1(s)+\cdots+f_n(t)g_n(s)\}\, dP,$$

where f_i and g_i are the components of $f(t)$ and $g(s)$, respectively. The corresponding norm satisfies $|f(t)|^2 = \langle f(t), f(t)\rangle$. The expected value of an element x of $L_{2,n}(\Omega)$ is $Ex = \int_\Omega x\, dP$; here, as elsewhere in the paper, Ω-dependence is suppressed in the notation.

The finite time interval to be considered is $S = [0, T]$, where T is a fixed positive constant. The space G of second-order n-dimensional stochastic processes consists of all functions from S into $L_{2,n}(\Omega)$ which are L_2-continuous in each component. The subspace G_0 of G consists of all functions $f \in G$ such that $f(0) = 0$ and $Ef(t) = 0$ for each $t \in S$. The subspace G_z of G_0 consists of all functions $z \in G_0$ such that z has independent increments: $z(t+\Delta t) - z(t)$ is stochastically independent of $z(s)$ for each $0 \leqq \Delta t \leqq T - t$ and $0 \leqq s \leqq t \leqq T$.

The disturbance z in the system equation (1.1) is taken to be a fixed but arbitrary element of G_z. The family of σ-algebras $\{F_t\}_{t\in S}$ corresponding to z is a fixed set of σ-algebras such that $F_s \subset F_t \subset F$ whenever $0 \leqq s \leqq t \leqq T$, and each F_t contains the σ-algebra generated by $\{z(s): 0 \leqq s \leqq t\}$. A canonical example is $z(t) = W(t)$, a Wiener process, and $F_t = \sigma(W(s): 0 \leqq s \leqq t)$. The conditional expectation of $x \in L_{2,n}(\Omega)$ relative to the σ-algebra F_t is denoted by $E(x|F_t)$. A function $f \in G$ is said to be nonanticipating if $E(f(s)|F_t) = f(s)$ whenever $0 \leqq s \leqq t \leqq T$.

For $f \in G$ and $v \in S$, the pseudonorm N_v on G is defined by

$$N_v(f) = \sup\{|f(x)|: 0 \leqq x \leqq v\}.$$

For $f \in G$ and $v \in S$, the projection $P_v f \in G$ is defined by

$$[P_v f](x) = \begin{cases} f(x), & 0 \leqq x \leqq v, \\ f(v), & v \leqq x \leqq T. \end{cases}$$

The above pseudonorms and projections are related by the identity $N_v(f) = N_T(P_v f)$ for each $f \in G$ and $v \in S$.

Integrations will be performed with respect to the measure $k(t) = 1 + t$, so that $dk(t) = dt$. However, all results in this paper hold with k being an arbitrary right continuous increasing function on S with $k(0) = 1$. The notation $k(t) = 1 + t$ will prove useful for writing formulas.

The operator space $\mathscr{A}$ consists of all functions $A: G \to G$ for which the following hold:

(i) A is linear;
(ii) $[Af](0) = f(0)$ for each $f \in G$;
(iii) A is deterministic: $EAf = AEf$ for each $f \in G$;
(iv) there exists a constant $c = c(A)$ such that if $f \in G$ and $0 \leqq u \leqq v \leqq T$,

$$|[Af](v) - f(v) - [Af](u) + f(u)| \leqq c \int_u^v N_t(f)\, dt.$$

The operator space $\mathscr{B}$ consists of all functions $B: G \to G$ for which the following hold:

(i) B is linear;
(ii) $[Bf](0) = 0$ for each $f \in G$;
(iii) B is deterministic: $EBf = BEf$ for each $f \in G$;

(iv) there exists a constant $c = c(B)$ such that if $f \in G$ and $0 \leqq u \leqq v \leqq T$,

$$|[Bf](v) - [Bf](u)| \leqq c \int_u^v N_t(f)\, dt.$$

The inequality $\int_u^v N_t(f)\, dt \geqq 0$ implies that $\mathscr{A}$ and $\mathscr{B}$ are nonempty; in particular, $0 \in \mathscr{B}$ and $I \in \mathscr{A}$, where 0 and I are the zero and identity operators on G, respectively. Another basic element of $\mathscr{B}$ is the operator C given by

$$[Cf](t) = \int_0^t f(s)\, ds, \qquad f \in G. \tag{2.1}$$

As will be shown in § 3, C has an adjoint C^*, with respect to the inner product space induced by a Hellinger integral, given by

$$[C^*f](t) = k(t)f(T) - f(0) - \int_0^t f(s)\, ds. \tag{2.2}$$

Since $[C^*f](0) = f(T) - f(0)$, C^* is not an element of $\mathscr{A}$ or $\mathscr{B}$. However, the quantity $k(t)[C^*Cf](0) - [C^*Cf](t)$ equals $[CCf](t)$, and $C^2 \in \mathscr{B}$ follows from part (vii) of Theorem 2.2. For similar operator spaces defined on $S_r = [-r, T]$, a fundamental element of $\mathscr{B}_r$ is the delay operator B_r given by

$$[B_r f](t) = \begin{cases} 0, & t \leqq 0, \\ \displaystyle\int_0^t b(s) f(s - r)\, ds, & t > 0, \quad f \in G. \end{cases}$$

Delay operators are not considered in this paper.

More examples of elements of $\mathscr{A}$ and $\mathscr{B}$ will be given in § 5. The most important tools for understanding the scope and nature of $\mathscr{A}$ and $\mathscr{B}$ are the fundamental relationship given in Theorem 2.1, and the algebraic structure detailed in Theorem 2.2. Using these results and the Laplace transform technique illustrated in § 5, many elements of $\mathscr{A}$ and $\mathscr{B}$ of practical interest can be computed.

THEOREM 2.1. *If $A \in \mathscr{A}$ and $B \in \mathscr{B}$, then the following hold*:

(i) *$I - B$ is* 1-1 *and onto G, and $(I - B)^{-1} \in \mathscr{A}$*;

(ii) *A is* 1-1 *and onto G, $A^{-1} \in \mathscr{A}$, and $I - A^{-1} \in \mathscr{B}$.*

The proof of Theorem 2.1 is given in § 7. In that proof, the inverse A of $I - B$ is constructed as

$$A = (I - B)^{-1} = I + B + B^2 + \cdots = \sum_{n=0}^{\infty} B^n. \tag{2.3}$$

On appropriate function spaces, (2.3) is identical to the series representation of Bharucha-Reid [3] for Volterra integral operators.

The two parts of Theorem 2.1 imply a 1-1 correspondence between elements of $\mathscr{A}$ and $\mathscr{B}$. The spaces are further related by a powerful algebraic structure, part of which is detailed in the following result.

THEOREM 2.2. *Let AB denote the composition of A and B: $(AB)f = A(Bf)$. If A, A_1, $A_2 \in \mathscr{A}$ and B, B_1, $B_2 \in \mathscr{B}$, then the following hold*:

(i) $-B \in \mathscr{B}$; (iv) $BA \in \mathscr{B}$; (vii) $B_1 B_2 \in \mathscr{B}$;

(ii) $A + B \in \mathscr{A}$; (v) $AB \in \mathscr{B}$; (viii) $P_t B \in \mathscr{B}$ *for* $t \in S$;

(iii) $B_1 + B_2 \in \mathscr{B}$; (vi) $A_1 A_2 \in \mathscr{A}$; (ix) *$\mathscr{A}$, $\mathscr{B}$ are convex.*

The proof of Theorem 2.2 is given in § 7. Theorems 2.1 and 2.2 will be used

extensively in solving equations. For example, the equation

$$x = y + [-B_1 + (I - B_2)^{-1}A]B_1 x$$

has the unique solution, as guaranteed by Theorems 2.1 and 2.2,

$$x = (I - [-B_1 + (I - B_2)^{-1}A]B_1)^{-1} y.$$

Although it is not a representation theorem, the following result provides some insight into the structure of elements of $\mathcal{A}$ and $\mathcal{B}$.

THEOREM 2.3. *Elements of $\mathcal{A}$ and $\mathcal{B}$ are causal; that is, if $f \in G$, $A \in \mathcal{A}$, and $B \in \mathcal{B}$, then $[Bf](t) = [BP_t f](t)$ and $[Af](t) = [AP_t f](t)$.*

Proof. Let $t \in S$ and $f \in G$ be fixed. It is first shown that $P_t B P_t = P_t B$. Clearly, $[P_t B P_t f](0) = [P_t B f](0) = 0$. Suppose that $0 = s_0 < s_1 < \cdots < s_q = t < \cdots < s_N = T$, and let $g = f - P_t f$. Consider $d = |[P_t B g](s_{p+1}) - [P_t B g](s_p)|$. If $p \geqq q$, then $s_{p+1} > s_p \geqq t$ and

$$d = |[Bg](s_{p+1}) - [BG](s_p)| \leqq c_B \int_{s_p}^{s_{p+1}} N_u(g)\, du = 0,$$

since $g(u) = 0$ for $u \leqq t$. Thus, $P_t B P_t = P_t B$, and

$$[Bf](t) = [P_t B f](t) = [P_t B P_t f](t) = [BP_t f](t).$$

Similarly, $P_t A P_t = P_t A$, and $[Af](t) = [AP_t f](t)$.

The above result indicates that the integral-like operators A and B do not use "future" values of their arguments. This property falls short of implying that A and B preserve the nonanticipating property. The operator D_N given in Lemma 4.1 indicates the type of operator that might not preserve nonanticipating functions.

3. The reproducing kernel Hilbert space. The types of integrals known as "Hellinger integrals" were introduced in 1907 in the dissertation of E. D. Hellinger [5]. Many properties of Hellinger integrals, including comparisons of spaces of Hellinger-integrable functions with the spaces of absolutely continuous functions and the functions of bounded variation, are given in [3] and [7], and reproducing kernels on spaces of Hellinger-integrable functions are introduced in [3] and [9].

An inner product space $\{G_H, Q\}$ of Hellinger-integrable functions is defined as follows, where $k(t) = 1 + t$.

The space G_H consists of all functions $f \in G$ for which there exists a constant $b = b(f)$ such that for each partition $\{s_p\}_{p=0,N}$ of $[0, T]$,

$$\sum_{p=1}^{N} |f(s_p) - f(s_{p-1})|^2 / (s_p - s_{p-1}) \leqq b.$$

The smallest such b is the Hellinger integral of f with respect to k, and is denoted $\int_0^T |df|^2 / dk$. The subspace G_{HO} consists of elements of $G_H \cap G_0$.

The functional Q is defined on $G_H \times G_H$ by

$$Q(f, g) = \langle f(0), g(0) \rangle + \int_0^T \langle df, dg \rangle / dk,$$

where the integral is the limit, through refinement of partitions, of the sums

$$\sum_{p=1}^{N} \langle f(s_p) - f(s_{p-1}), g(s_p) - g(s_{p-1}) \rangle / (s_p - s_{p-1}).$$

The corresponding norm is denoted N_H.

It is easy to verify that $\{G_H, Q\}$ is an inner product space. Furthermore, it is complete [3]. Elements of G_H are (in the mean-square sense) absolutely continuous, and G_H is dense with respect to the variation norm in the space of absolutely continuous functions [3]. The similarity of the N_H norm and a Sobolev space norm should be noted; in fact, with $k(t) = 1 + t$, G_H can be thought of as a mean-square Sobolev space.

A fundamental issue to be resolved is the effect of elements of $\mathscr{A}$ and $\mathscr{B}$ on elements of G and G_H. As seen in Theorem 2.1, A and $I - B$ map G onto G. The relationship of $\mathscr{A}$ and $\mathscr{B}$ to G_H is given in the following result.

THEOREM 3.1. *If $f \in G$, $B \in \mathscr{B}$, and $A \in \mathscr{A}$, then $Bf \in G_H$ and A maps G_H onto G_H.*

Proof. Let $0 = s_0 < s_1 < \cdots < s_N = T$, and let c be the constant corresponding to $B \in \mathscr{B}$. Then

$$\sum_{p=1}^{N} |[Bf](s_p) - [Bf](s_{p-1})|^2/(s_p - s_{p-1})$$

$$\leqq \sum_{p=1}^{N} c^2 \int_{s_{p-1}}^{s_p} N_t(f)\, dt^2/(s_p - s_{p-1})$$

$$\leqq c^2 N_T^2(f)\, T.$$

Thus, $Bf \in G_H$. Now, let $A = (I - B)^{-1}$, $g \in G_H$ be arbitrary, and $f = Ag$. Then $f - Bf = g$ and $f = g + Bf$. Since $g \in G_H$ and $Bf \in G_H$, it holds that $f \in G_H$. To show that A is onto G_H, let $f \in G_H$ be arbitrary and call $g = f - Bf$. Clearly, $g \in G_H$ and $f = Ag$.

A classical example of a reproducing kernel is the Green's function of a self-adjoint ordinary differential equation. For a discussion of the development and properties of reproducing kernels, see Aronszajn [1]. A reproducing kernel is defined as follows.

A function K from $S \times S$ into the linear transformations of $L_{2,n}(\Omega)$ is a reproducing kernel for $\{G_H, Q\}$ if for every $t \in S$, $x \in L_{2,n}(\Omega)$, and $f \in G_H$, the following hold:

(i) $K[\cdot, t]x \in G_H$;

(ii) $\langle f(t), x\rangle = Q(f, K[\cdot, t]x)$.

It is now shown that a reproducing kernel for $\{G_H, Q\}$ is given by

$$K(u, v) = \{\min(u, v)\}I, \tag{3.1}$$

where I is the identity transformation of $L_{2,n}(\Omega)$. If $x \in L_{2,n}(\Omega)$ and $t \in S$, it is clear that $K[\cdot, t]x \in G_H$. Let $f \in G$, $t \in S$, and $x \in L_{2,n}(\Omega)$ be arbitrary. Then

$$Q(f, K[\cdot, t]x) = \langle f(0), x\rangle + \int_S \langle df, d(K[\cdot, t]x)\rangle/dk$$

$$= \langle f(0), x\rangle + \int_0^t \langle df, x\, dk\rangle/dk$$

$$= \langle f(0), x\rangle + \int_0^t \langle df, x\rangle$$

$$= \langle f(t), x\rangle.$$

The reproducing kernel is now used to compute the adjoint C^* of C. Let $f \in G_H$ and

$g \in G_H$ be arbitrary. Since $[Cf](0) = 0$, it follows that

$$\begin{aligned} Q(Cf, g) &= \int_0^T \langle d(Cf), dg\rangle / dk = \int_0^T \langle f, dg\rangle \\ &= \langle f(T), g(T)\rangle - \langle f(0), g(0)\rangle - \int_0^T \langle df, g\rangle \\ &= Q(f, K[\cdot, T]g(T) - K[\cdot, 0]g(0) - Cg) \\ &= Q(f, C^*g) \end{aligned}$$

where $[C^*g](t) = k(t)g(T) - g(0) - \int_0^t g(s)\, ds$.

Three lemmas which will be needed to solve the control problem of § 4 are now presented. The first two lemmas are versions of the Hahn-Banach and Riesz representation theorems [11], respectively. The proof of Lemma 3.3. illustrates some of the power of the reproducing kernel.

LEMMA 3.1. *Let $L_0(\Omega)$ be the set of all $x \in L_{2,n}(\Omega)$ such that $Ex = 0$, and let $\{h_p\}_{p=1,N}$ and $\{c_p\}_{p=1,N}$ be sets of elements of $L_0(\Omega)$ such that the h_p's are linearly independent. Then there exists a continuous linear transformation M of $L_{2,n}(\Omega)$ such that $EM = ME$ and $Mh_p = c_p$ for $p = 1, \cdots, N$.*

LEMMA 3.2. *If λ is a continuous linear functional on G, then there exists a function λ of bounded variation (in S) such that for $f \in G$, $\lambda(f) = \int_S \langle f, d\lambda\rangle$.*

LEMMA 3.3. *If $f \in G_H$, $\lambda \in G$ has bounded variation, and $\lambda(T) = 0$, then $\int_S \langle f, d\lambda\rangle = Q(f, C_0\lambda)$, where $[C_0\lambda](t) = \lambda(0) - \int_S \lambda(s)\, ds$.*

Proof. Let $x \in L_{2,n}(\Omega)$, $t \in S$, and $f = K[\cdot, t]x$. Then integration by parts gives

$$\begin{aligned} \int_S \langle f, d\lambda\rangle &= \langle f(T), \lambda(T)\rangle - \langle f(0), \lambda(0)\rangle - \int_S \langle df, \lambda\rangle \\ &= -\langle K[0, t]x, \lambda(0)\rangle - \int_S \langle x dK[\ , t], \lambda\rangle \\ &= -\langle x, \lambda(0)\rangle - \int_S \langle x dk, \lambda\rangle \\ &= -\langle x, \lambda(0)\rangle - \left\langle x, \int_S \lambda(t)\, dt \right\rangle \\ &= Q(f, C_0\lambda). \end{aligned}$$

Thus, the conclusion holds for $f = K[\ , t]x$. For arbitrary $f \in G_H$, the theory of reproducing kernel Hilbert spaces [10] gives $f = \lim_{i\to\infty} f_i$, where f_i, where $f_i = \sum_{p=0}^{N(i)} K[\ , t_i(p)]x_i(p)$ where $x_i \in L_{2,n}(\Omega)$ and the partition $\{t_{i+1}(p)\}_{p=0,N(i+1)}$ is a refinement of the partition $\{t_i(p)\}_{p=0,N(i)}$. Since $\int_S \langle f_i, d\lambda\rangle = Q(f_i, C_0\lambda)$ for each i, it follows that

$$\int_S \langle f, d\lambda\rangle = \lim_{i\to\infty} \int_S \langle f_i, d\lambda\rangle = \lim_{i\to\infty} Q(f_i, C_0\lambda) = Q(f, C_0\lambda),$$

and the proof of the lemma is complete.

4. The control problem. In this section, a linear-quadratic stochastic control problem is stated and a necessary and sufficient condition is found for its solution. System

dynamics is given by the state-space form

$$h(t) = \int_0^t \alpha h(s)\, ds + u(t) + z(t) = \alpha[Ch](t) - + u(t) + z(t),$$

where α is an $n \times n$ constant matrix. The output h can be measured. State-independent noise in any part of the system is absorbed into the disturbance term z. It is assumed that $z \in G_z$, so that z could be Gaussian (i.e., z equals a multiple of a Wiener process on $[0, T]$), but is not required to be. The general nature of z allows a specific control problem to be formulated in terms of a "natural" model. That is, the state vector (and, in particular, its dimension) can be chosen without regard to the resulting form of the noise term, as long as the properties of G_z are retained. This modeling issue is discussed in more detail in § 5.

The objective of the control problem is to choose a feedback control $u = Dh$ for some $D \in \mathcal{B}$, such that a quadratic cost functional given by

$$\begin{aligned} J(D, h) &= \tfrac{1}{2}c_1 N_H^2(Dh) + \tfrac{1}{2}c_2 N_H^2(Ch) + \tfrac{1}{2}|h(T)|^2 \\ &= \tfrac{1}{2}E\left\{\int_0^T c_1 |du/dt|^2 + c_2 |h(t)|^2\, dt + |h(T)|^2\right\} \end{aligned}$$

is minimized.

To effectively apply the Lagrange multiplier theorem to this problem, it is necessary to have the following result.

LEMMA 4.1. *If $B \in \mathcal{B}$, $z \in G_z$, and $h = Bh + z$, then $\{f = Dh\colon D \in \mathcal{B}\}$ is dense in G_{HO}.*

Proof. By Theorem 3.1, each $f = Dh$ is in G_H, and since $f = DAz$, $Ef = 0$. Thus, $Dh \in G_{HO}$. Let $0 = t_0 < t_1 < \cdots < t_N = T$, and let $\{c_p\}_{p=1,N}$ be any sequence of elements of $L_0(\Omega)$. By Lemma 3.1, there exists a continuous linear transformation M of $L_{2,n}(\Omega)$ such that $Mh(t_p) = c_p$ for $p = 1, \cdots, N$. For this M, define a linear transformation D_N on G_H by

$$[D_N f](t) = \begin{cases} 0, & t = 0, \\ (t - t_{p-1})Mf(t_{p-1}) + [D_N f](t_{p-1}), & t_{p-1} < t \leqq t_p. \end{cases}$$

Note that $[D_N h](0) = [D_N h](t_1) = 0$, $[D_N h](t_2) = (t_2 - t_1)c_1$, etc. We show that $D_N \in \mathcal{B}$: if $t_{p-1} \leqq u \leqq v \leqq t_p$, then

$$|[D_N f](v) - [D_N f](u)| = |(v-u)Mf(t_{p-1})| \leqq (v-u)\|M\| N_{t_{p-1}}(f) \leqq \|M\| \int_u^v N_t(f)\, dt,$$

and $D_N E = E D_N$ follows from $EM = ME$. By definition, $D_N h$ is a K-polygon in G_{HO}: a K-polygon is any function that can be written in the form $\sum_{p=0}^{N} K[t, s_p]x_p$, where $\{x_p\}$ is a sequence of random variables in $L_{2,n}(\Omega)$, and $\{s_p\}$ is any sequence of numbers in $(0, T]$. As shown in [2], the set of K-polygons is dense in G_{HO}. Since N, $\{t_p\}$, and $\{c_p\}$ are arbitrary, $\{f = Dh\colon D \in \mathcal{B}\}$ contains all K-polygons in G_{HO}, and hence is dense in G_{HO}.

A complete statement of the control problem is now presented, followed by the main result of this paper. The control function is written explicitly as $u = Dh$ for some D.

Problem P. Let α be a constant $n \times n$ matrix, and let $z \in G_z$ be arbitrary but fixed. Let Γ be all pairs $(D, h) \in \mathcal{B} \times G_0$, and define a functional J on Γ by

$$J(D, h) = \tfrac{1}{2}c_1 N_H^2(Dh) + \tfrac{1}{2}c_2 N_H^2(Ch) + \tfrac{1}{2}|h(T)|^2$$

for fixed positive constants c_1 and c_2. Find the pair (D, h) that minimizes J over Γ subject to $h = \alpha Ch + Dh + z$.

THEOREM 4.1. *A necessary and sufficient condition for* $(D, h) \in \Gamma$ *to solve Problem* P *is that*

$$(I - P_0)\{c_2 C^* Ch + kh(T) + c_1(I - \alpha^* C^*)Dh\} = 0 \tag{4.1}$$

where C^* *is the adjoint of* C *in* $\{G_H, Q\}$ *and* α^* *is the transpose of* α. *Furthermore, if* (4.1) *holds, then*

$$Dh = (c_2/c_1)[I + \alpha^* C]^{-1} C^2 h. \tag{4.2}$$

Note that the expression (4.2) *is independent of the end-time* T.

Proof. It is first shown that (4.1) is necessary. The variation of the cost functional $J(D, h)$ in the direction of $(D, h) \in \Gamma$ is given by

$$\delta J(D, h; \tilde{D}, \tilde{h}) = c_1 Q(\tilde{D}h + D\tilde{h}, Dh) + c_2 Q(C\tilde{h}, Ch) + \langle \tilde{h}(T), h(T) \rangle.$$

The constraint is $K_1(D, h) = h - \alpha Ch - Dh - z = 0$, with variation $\delta K_1 = \tilde{h} - \alpha C\tilde{h} - \tilde{D}h - D\tilde{h}$. The Lagrange multiplier thorem [12] gives the following necessary condition for a minimizing pair (D, h): there exists a continuous linear functional $\tilde{\lambda}$ defined on G_0 such that for every $(\tilde{D}, \tilde{h}) \in \Gamma$,

$$c_1 Q(\tilde{D}h + D\tilde{h}, Dh) + c_2 Q(C\tilde{h}, Ch) + \langle \tilde{h}(T), h(T) \rangle + \tilde{\lambda}(\tilde{h} - \alpha C\tilde{h} - \tilde{D}h - D\tilde{h}) = 0.$$

By Lemmas 3.2 and 3.3, there exists a function λ_1 of bounded variation such that

$$\tilde{\lambda}(\tilde{h} - \alpha C\tilde{h} - \tilde{D}h - D\tilde{h}) = Q(\tilde{h} - \alpha C\tilde{h} - \tilde{D}h - \tilde{D}h, C_0\lambda_1) = Q(\tilde{h} - \alpha C\tilde{h} - \tilde{D}h - D\tilde{h}, \lambda)$$

where $\lambda = (I - P_0)C_0\lambda_1$. Furthermore, if $\tilde{h} \in G_H$, then

$$\langle \tilde{h}(T), h(T) \rangle = Q(\tilde{h}, K(\ , T)h(T)).$$

Thus, the necessary condition becomes: there exists $\lambda \in G_H$ such that

$$\begin{aligned} c_1 Q(\tilde{D}h + D\tilde{h}, Dh) + c_2 Q(C\tilde{h}, Ch) + Q(\tilde{h}, K(\ , T)h(T)) \\ + Q(\tilde{h} - \alpha C\tilde{h} - \tilde{D}h - D\tilde{h}, \lambda) = 0 \end{aligned} \tag{4.3}$$

for each $(\tilde{D}, \tilde{h}) \in \mathcal{B} \times \mathcal{G}_{HO}$. Letting $\tilde{h} = 0$, (4.3) implies that

$$Q(\tilde{D}h, c_1 Dh) - Q(\tilde{D}h, \lambda) = 0, \qquad \tilde{D} \in \mathcal{B}. \tag{4.4}$$

By Lemma 3.1, it follows that $c_1 Dh = \lambda$, since $[Dh](0) = \lambda(0) = 0$. Letting $\tilde{D} = 0$, (4.3) implies that

$$Q(\tilde{h}, c_1 D^* Dh) + Q(\tilde{h}, c_2 C^* Ch) + Q(\tilde{h}, K(\ , T)h(T)) + Q(h, (I - \alpha C - D)^* \lambda) = 0, \qquad \tilde{h} \in G_{HO}.$$

It follows that

$$(I - P_0)\{c_1 D^* Dh + c_2 C^* Ch + K(\ , T)h(T) + (I - \alpha^* C^* - D^*)\lambda\} = 0. \tag{4.5}$$

Combining (4.4) and (4.5) and setting $\lambda = c_1 Dh$ gives the desired equation (4.1). We now proceed to derive (4.2). Using the representations (2.2) and (3.1) of C^* and K, respectively, it is clear that (4.1) is equivalent to

$$c_2(k-1)[Ch](T) - c_2 C^2 h + (k-1)h(T) + \lambda - \alpha^*(k-1)\lambda(T) + \alpha^* C\lambda = 0.$$

Calling $\gamma = -c_2[Ch](T) - h(T) - h(T) + \alpha^*\lambda(T)$, it follows that

$$\lambda + \alpha^* C\lambda = c_2 C^2 h + (k-1)\gamma,$$

and $\alpha^* C \in \mathcal{B}$ implies that

$$\lambda = (I + \alpha^* C)^{-1}(c_2 D^2 h + (k-1)\gamma). \tag{4.6}$$

Intuitively, $\lambda = c_1 Dh$ for some $D \in \mathcal{B}$ implies that $\gamma = 0$. This is formalized below. Since $\lambda = c_1 Dh$, (4.6) and the constraint $h = (I - \alpha C - D)^{-1} z$ imply that

$$(I + \alpha^* C)^{-1}(k-1)\gamma = c_1 Dh - (I + \alpha^* C)^{-1} c_2 C^2 h = D_1 h = Bz,$$

where D_1 and B are the appropriately defined elements of $\mathcal{B}$. Solving for γ, we find that

$$\gamma = \frac{[(I + \alpha^* C)Bz](t)}{t}, \qquad t \in (0, T].$$

Taking a limit of the above equation as $t \to 0$, and observing that $(I + \alpha^* C)^{-1}(k-1)\gamma$, and hence Bz, is differentiable, l'Hôpital's rule gives

$$= \lim_{t \to 0} \frac{[(I + \alpha^* C)Bz](t)}{t} = \lim_{t \to 0} \{[Bz]'(t) + \alpha^* [Bz](t)\}$$

$$= [Bz]'(0) = \lim_{\varepsilon \to 0} \frac{[Bz](\varepsilon) - [Bz](0)}{\varepsilon} = \lim_{\varepsilon \to 0} \frac{[Bz](\varepsilon)}{\varepsilon}.$$

If c is the constant corresponding to $B \in \mathcal{B}$, then

$$\left| \frac{[Bz](\varepsilon)}{\varepsilon} \right| \leqq \frac{c}{\varepsilon} \int_0^{\varepsilon} N_t(z)\, dt \leqq c N_\varepsilon(z) \to 0 \quad \text{as } \varepsilon \to 0.$$

Thus, $\gamma = 0$, and (4.2) follows from (4.6). Finally, we show that (4.1) is sufficient. Clearly, (4.1) implies that (4.3) holds with $\lambda = c_1 Dh$. Suppose that (D, h) satisfies (4.3) with $h = \alpha C h + Dh + z$, and that $(\hat{D}, \hat{h}) \in \Gamma$ satisfies $\hat{h} = \alpha C \hat{h} + \hat{D}\hat{h} + z$. Then $h - \hat{h} \in G_{HO}$, and

$$0 = c_1 Q(\hat{D}h + D(h - \hat{h}), Dh) + c_2 Q(C(h - \hat{h}), Ch) + O(h - \hat{h}, kh(T))$$
$$+ Q(h - \hat{h} - \alpha C(h - \hat{h}) - \hat{D}h - D(h - \hat{h}), c_1 Dh).$$

Since $h - \hat{h} - \alpha C(h - \hat{h}) = Dh - \hat{D}\hat{h}$,

$$0 = c_1 Q(\hat{D}h + D(h - \hat{h}), Dh) + c_2 Q(C(h - \hat{h}), Ch) + Q(h - \hat{h} kh(T))$$
$$+ Q(-\hat{D}\hat{h} - \hat{D}h + D\hat{h}, c_1 Dh)$$
$$= c_1 N_H^2(Dh) + c_2 N_H^2(Ch) - c_1 Q(\hat{D}\hat{h}, Dh) - c_2 Q(C\hat{h}, Ch) + Q(h - \hat{h}, kh(T)).$$

Thus,

$$0 \leqq \tfrac{1}{2} c_1 N_H^2(\hat{D}\hat{h} - Dh) + \tfrac{1}{2} c_2 N_H^2(C\hat{h} - Ch) + \tfrac{1}{2} |\hat{h}(T) - h(T)|^2$$
$$= J(\hat{D}, \hat{h}) + J(D, h) - c_1 Q(\hat{D}\hat{h}, Dh) - c_2 Q(C\hat{h}, Ch) - \langle \hat{h}(T), h(T) \rangle$$
$$= J(\hat{D}, \hat{h}) + J(D, h) - c_1 N_H^2(Dh) - c_2 N_H^2(Ch) - Q(h, kh(T))$$
$$= J(\hat{D}, \hat{h}) - J(D, h).$$

Thus, the pair (D, h) satisfying (4.1) also satisfies $J(D, h) \leqq J(\hat{D}, \hat{h})$, with equality only if $\hat{h} = h$ and $\hat{D} = D$. This completes the proof of Theorem 4.1.

5. An application. In this section, the above result is applied to a pendulum model adapted from Russell [13]. His results are given to allow comparisons of some modeling and computational aspects of the approach taken in this paper and of Russell's more traditional approach.

The physical situation to be referred to is that of a pendulum which is constrained to move in a one-dimensional arc. The pendulum is thought of as a surveyor's instrument subjected to a random force w, which may represent wind or other disturbances. The connection of the pendulum to its housing causes friction, so that a simple linear model of the displacement angle $\theta(t)$ is

$$\theta''+\theta'+\theta=w, \quad \theta(0)=a, \quad \theta'(0)=b, \tag{5.1}$$

where a and b are mean-zero random variables. A control function is to be chosen as a function of θ and θ', and is allowed to affect θ and θ'.

To apply most state-space methods, it is necessary that the system noise be Gaussian; this requirement may play a large role in determining the state vector. For example, Russell uses the state vector $X=(\theta, \theta', w)^T$, because wind is not well modeled as white noise (the "derivative" of a Wiener process). If w is driven by white noise (that is, $w'+w=v=$ white noise), then the system is indeed linear and Gaussian:

$$X'=\begin{pmatrix} 0 & 1 & 0 \\ -1 & -1 & 1 \\ 0 & 0 & -1 \end{pmatrix} X+\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} v. \tag{5.2}$$

If w is not driven by white noise, further components (e.g., w') might be added to the state vector. The control only affects θ and θ', however, so components other than θ and θ' are in a practical sense superfluous.

The operator approach allows a more flexible approach to modeling, although it must be assumed that $a=b=0$ in (5.1). Equation (5.2) may be integrated to obtain an admissible 3-dimensional model, and equation (5.1) may be double-integrated to obtain an admissible 1-dimensional model. Since the control is 2-dimensional, the appropriate 2-dimensional model is presented below. Integrating (5.1) and identifying $h=(\theta, \theta')^*$ gives

$$h=\alpha Ch+z, \quad \alpha=\begin{pmatrix} 0 & 1 \\ -1 & -1 \end{pmatrix}, \quad z(t)=\left(0 \cdot \int_0^t w(s)\,ds\right)^*, \tag{5.3}$$

The control function (4.2) and the resulting output $h=(I-\alpha C-D)^{-1}$ may be computed using Laplace transforms. If $g=(I+\alpha * C)^{-1}f$, $G=(g)$, and $F=(f)$, then $(I+\alpha */s)G=F$. Inverting the matrix $I+\alpha^*/s$ and taking inverse transforms gives

$$g(t)=f(t)+\int_0^t m(t-u)f(u)\,du,$$

$$m(x)=\exp(x/2)\begin{pmatrix} 1 \\ 3 \end{pmatrix}\sin\sqrt{3}/4\times\begin{pmatrix} -2 & 1 \\ -1 & -1 \end{pmatrix}+\cos\sqrt{3}/4\times\begin{pmatrix} 0 & 1 \\ -1 & 1 \end{pmatrix}.$$

A similar computation gives

$$\theta(t)=\int_0^t n(t-u)w(u)\,du,$$

$$n(x)=\cosh(\gamma x/2)[\cos\beta x-(1/2)\sin(\beta x)]$$

$$+\sinh(\gamma x/2)\left[\left(\frac{-1}{\gamma}\right)\cos(\beta x)+\left(\frac{5}{2\beta\gamma}\right)\sin(\beta x)\right],$$

$$\gamma^2=1+2\sqrt{5}, \qquad \beta^2=2\sqrt{5}-\tfrac{1}{2},$$

from which it is relatively easy to compute $E(\theta^2)$.

The corresponding calculations in Russell are done numerically. For the sake of comparison, it is assumed that the disturbance w in (5.1) is Gaussian, so that the covariance X of the optimal (Russell) output satisfies

$$X' = SX + XS* + \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \quad X(0) = 0, \quad S = \alpha + K,$$

$$K' = -\alpha * K - K\alpha + I - K^2, \qquad K(T) = -I.$$

Some values of $E(\theta^2)$ for the two methods are given below for $T = 1$.

(5.4)

t	.25	.5	.75	1.0
Russell	.0013	.0128	.0519	.1391
Operator	.0024	.0197	.0671	.1491

Another comparison of the two methods can be made by differentiating the optimal Russell control $u_r = Kh$, and substituting for K' and h'. After integrating twice, it can be shown that $u_0 = u_r$ satisfies

$$u_0 = (I + \alpha * C)^{-1} C^2 h + (I + \alpha * C)^{-1} C^2 K z'.$$

The difference between u_0 and the optimal control (4.2), then, is the final term in the above expression. This difference is reflected in the table (5.4).

6. Conclusion. A general framework for the study of linear stochastic systems was presented, characterized by a reproducing kernel Hilbert space and the operator spaces $\mathscr{A}$ and $\mathscr{B}$. A state-space control problem was solved within this framework, and a necessary condition was found giving the optimal control explicitly as a function of the output. The function spaces used admit systems with delays and state-department noise, and it is hoped that results similar to Theorem 4.1 of this paper can be found in these more general settings.

7. Proofs of operator properties. The proofs of Theorems 2.1 and 2.2 are presented in this section.

Proof of Theorem 2.1. Let $f \in G$ be arbitrary, and c be the constant for fixed $B \in \mathscr{B}$. If $0 \leqq v \leqq T$, $Bf(0) = 0$ implies that

$$|Bf(v)| \leqq c \int_0^v N_t(f)\, dt. \tag{7.1}$$

Since $N_t(f)$ is increasing in t, it is clear from (7.1) that $N_v(Bf) \leqq c \int_0^v N_t(f)\, dt$. Applying this inequality to (7.1) and integrating by parts gives

$$|B^2 f(v)| \leqq c \int_0^v c \int_0^t N_s(f)\, ds\, dt = c^2 \int_0^v (v-t) N_t(f)\, dt.$$

By induction, we have that

$$|B^n f(v)| \leqq c^n \int_0^v \frac{(v-t)^{n-1}}{(n-1)!} N_t(f)\, dt \leqq \frac{c^n T^n}{(n-1)!} N_T(f).$$

Summing over n gives, for each p,

$$\left| \sum_{n=0}^{p} B^n f(v) \right| \leqq N_T(f) \sum_{n=1}^{\infty} c^n T^n / (n-1)!,$$

where the infinite series converges by the Ratio test. Thus, $\sum_{n=0}^{\infty} B^n f$ converges, and

$h = f + Bh$ has a unique solution. This establishes that $I - B$ is $1-1$ and onto, and hence has an inverse. Call $D = (I-B)^{-1}$. Then $D - I = BD$. Let $0 \leqq t \leqq T$. Since $B \in \mathscr{B}$,

$$|Df(t) - f(t)| \leqq c \int_0^t N_s(f) + N_s((D-I)f)\, ds. \tag{7.2}$$

Equation (7.2) implies $N_s((D-I)f) \leqq c \int_0^s N_u(f) + N_u((D-I)f)\, ds$, so that integration by parts gives

$$|Df(t) - f(t)| \leqq c \int_0^t (1 + c(t-s)) N_s(f) + c(t-s) N_s((D-I)f)\, ds.$$

Induction gives

$$|Df(t) - f(t)| \leqq c \int_0^t \left[1 + c(t-s) + \cdots + \frac{c^n (t-s)^n}{n!}\right] N_s(f) + \frac{c(t-s)^n}{n!} N_s(BDf)\, ds.$$

Letting $n \to \infty$ and simplifying, we get

$$|Df(t) - f(t)| \leqq c e^{ct} \int_0^t N_s(f)\, ds.$$

This inequality is used in the second of the inequalities below. Arguments similar to those given above yield

$$\begin{aligned} |((D-I)f)(v) - ((D-I)f)(u)| &\leqq c \int_u^v N_s(f) + N_s((D-I)f)\, ds \\ &\leqq c \int_u^v N_s(f) + c e^{ct} \int_0^s N_w(f)\, dw\, ds \\ &\leqq cT(1 + c e^{cT}) \int_u^v N_s(f)\, ds. \end{aligned}$$

Thus, $D = (I-B)^{-1} \in \mathscr{A}$, and part (i) is proven.

It is clear that $A \in \mathscr{A}$ implies $I - A \in \mathscr{B}$. From above, then, $A = I - (I - A)$ is $1-1$ and onto. If $A^{-1} \in \mathscr{A}$, it is clear that $I - A^{-1} \in \mathscr{B}$. It remains to show that $A^{-1} \in \mathscr{A}$.

Let c be the constant corresponding to $A \in \mathscr{A}$. Let $f \in G$ be arbitrary. Since A is onto, $f = Ag$ for some $g \in G$, and

$$|A^{-1} f(t) - f(t)| \leqq c \int_0^t N_s(g)\, ds \leqq c \int_0^t N_s(f) + N_s((A^{-1} - I)f)\, ds.$$

Replacing $N_s((A^{-1} - I)f)$ with its bound from the above inequality and integrating by parts, we get

$$|A^{-1} f(t) - f(t)| \leqq c \int_0^t (1 + c(t-s)) N_s(f) + c(t-s) N_s((A^{-1} - I)f)\, ds.$$

It follows that

$$|A^{-1} f(t) - f(t)| \leqq c e^{cT} \int_0^t N_s(f)\, ds.$$

The above inequality is used to show that

$$|A^{-1}f(v)-f(v)-A^{-1}f(u)+f(u)| \leqq c\int_u^v N_t(g)\,dt$$
$$\leqq c\int_u^v N_t(f)+ce^{cT}\int_0^t N_s(f)\,ds\,dt$$
$$\leqq c(1+cTe^{cT})\int_u^v N_t(f)\,dt.$$

Thus, $A^{-1}\in\mathscr{A}$, and the proof is complete.

Proof of Theorem 2.2. Let $f\in G$, and c_A, c_B, $\cdots$ be the constants corresponding to A, B, $\cdots$, respectively.

Part (i) is obvious.

Part (ii) follows from

$$|(A+B)f(v)-f(v)-(A+B)f(u)+f(u)| \leqq (c_A+c_B)\int_u^v N_t(f)\,dt.$$

Part (iii) follows from

$$|(B_1+B_2)f(v)-(B_1+B_2)f(u)| \leqq (c_{B_1}+c_{B_2})\int_u^v N_t(f)\,dt.$$

To prove (iv), let $A=I-B_1$. Then $A-I\in\mathscr{B}$. Let c_1 be the constant for $A-I$. Then

$$|BAf(v)-BAf(u)| \leqq c_B\int_u^v N_t(f)+N_t(B_1f)\,dt$$
$$\leqq c_B(1+c_1T)\int_u^v N_t(f)\,dt.$$

Part (v) follows from

$$|ABf(v)-ABf(u)| \leqq (c_Ac_BT+c_B)\int_u^v N_t(f)\,dt.$$

For (vi), let $A_1=(I-B_1)^{-1}$ and $A_2=(I-B_2)^{-1}$. Since $I-B\in\mathscr{A}$, (iv) gives $B_2(I-B_1)\in\mathscr{B}$, and (i), (ii) give $(I-B_1)-B_2(I-B_1)\in\mathscr{A}$. Then, Theorem 2.1(i) gives $A_1A_2\in\mathscr{A}$.

For (vii), $I-B_2\in\mathscr{A}$, so that (iv) gives $B_1(I-B_2)\in\mathscr{B}$, and (i), (iii) give $B_1B_2=B_1-B_1(I-B_2)\in\mathscr{B}$.

For (viii), it is sufficient to consider $0\leqq u\leqq t\leqq v\leqq T$. Then

$$|P_tBf(v)-P_tBf(u)|=|Bf(t)-Bf(u)| \leqq c_B\int_u^v N_s(f)\,ds.$$

For (ix), $cB\in\mathscr{B}$ for any constant c and (iii) imply that $\mathscr{B}$ is convex. If $a+b=1$, $A_1=I-B_1$, and $A_2=I-B_2$, then $aA_1+bA_2=I-aB_1-bB_2\in\mathscr{A}$, so that $\mathscr{A}$ is convex.

REFERENCES

[1] N. ARONSZAJN, *Theory of reproducing kernels*, Trans. Amer. Math. Soc., 68 (1950), pp. 337-404.

[2] S. L. BENZ AND R. E. FENNEL, *Parameter estimation in a reproducing kernel Hilbert space for linear hereditary systems*, Ph.D thesis, Clemson Univ. Clemson, SC, August 1981.

[3] A. T. BHARUCHA-REID, *Random Integral Equations*, Academic Press, New York, 1972.

[4] W. H. FLEMING AND R. RISHEL, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, Berlin, 1975.

[5] E. D. HELLINGER, *Die Orthogonalinvarianten Quadratischer Formen von Unendlichvielen Variablen*, Ph.D. thesis, Gottingen, 1907.
[6] F. N. HUGGINS, *Some interesting properties of the variation function*, Amer. Math. Monthly, 83 (1976), pp. 538-546.
[7] E. KREYSZIG, *Introductory Functional Analysis With Application*, John Wiley, New York, 1978.
[8] D. G. LUENBERGER, *Optimization by Vector Space Methods*, John Wiley, New York, 1969.
[9] J. S. MACNERNEY, *Hellinger integrals in inner product spaces*, J. Elisha Mitchell Sci. Soc., 76 (1960), pp. 252-273.
[10] J. A. RENEKE, *Control of a large scale hereditary system*, Proc. 15th S.E. Symposium on System Theory, IEEE Press, 1983.
[11] ———, *Optimal control of an hereditary system*, Clemson Univ. Tech. Rep., 341, Clemson, SC, 1980.
[12] J. A. RENEKE AND R. E. FENNELL, *An identification problem for hereditary systems*, Int. J. Appl. Anal., 11 (1981), pp. 167-183.
[13] D. L. RUSSELL, *Mathematics of Finite-Dimensional Control Systems*, Lecture Notes in Pure and Applied Mathematics, 43, Marcel Dekker, New York, 1979.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 4, July 1986

010

HIGHER ORDER CONDITIONS WITH AND WITHOUT LAGRANGE MULTIPLIERS*

J. WARGA†

Abstract. Let Q be a convex subset of a vector space, $\mathcal{U} \subset Q$, $\mathcal{Z}$ a topological vector space, C a convex subset of $\mathcal{Z}$ with a nonempty interior, $\phi = (\phi_1, \phi_2)$: $Q \to \mathbb{R}^m \times \mathcal{Z}$, $\bar{q} \in Q$ and $\phi_2(\bar{q}) \in C$. We assume that ϕ has a pth order Taylor approximation at $\bar{q}$ when it is restricted to an arbitrary finite-dimensional simplex in Q with a vertex at $\bar{q}$. In the case when $\mathcal{U}$ is a proper subset of Q we also assume that Q is a uniform space, ϕ continuous and $\mathcal{U}$ "abundant." We establish a number of higher order sufficient conditions, not involving any Lagrange multipliers, for the existence of neighborhoods G_1 and G_2 of the origins such that $\phi_1(\bar{q}) + G_1 \subset \{\phi_1(u) | u \in \mathcal{U}, \phi_2(u) + G_2 \subset C\}$. These sufficient conditions, involving nonconvex sets of variations, are shown by an example to be stronger than those in the literature. We also generalize prior "Lagrangian" conditions, more akin to the usual necessary conditions.

Key words. local controllability, inclusion restrictions, equality restrictions, Lagrange multipliers

AMS(MOS) subject classifications. 49B27, 49E15, 49E30

1. Introduction. The usual roles of first order and of higher order necessary conditions in optimization are quite different. First order conditions provide candidates for a restricted minimum. Higher order necessary conditions are applied in an attempt to eliminate some of these "first order" candidates from competition. These opposite tendencies also appear in the proofs of the necessary conditions. If we wish to minimize $\phi_0(x)$ on some convex set X subject to the restriction $\phi_1(x) = 0$ then the usual argument proceeds by contradiction: the assumption that the point $\phi(\bar{x}) = (\phi_0, \phi_1)(\bar{x})$ does not lie on the boundary of some "convex version" of the set $\phi(X)$ (that is, the denial of what is essentially the statement of necessary conditions) enables us to construct a set $X_1 \subset X$ such that $\phi(\bar{x})$ lies in the interior of $\phi(X_1)$.

Thus various necessary conditions, of first order and of higher orders, can be stated as (nonexclusive) alternatives [3], [4], [5]: either certain extremality relations hold at $\bar{x}$ or ϕ is controllable at $\bar{x}$, i.e., $\phi(\bar{x})$ belongs to the interior of $\phi(X)$. (We use the term "controllable" instead of "open" because of certain additional inclusion restrictions which are considered in the general problem.) The extremality conditions are formulated in terms of a Lagrange multiplier which represents a functional supporting some convex set of variations. In the case of first order conditions, the largest set of variations is convex and the conditions are quite general. However, in the case of higher order conditions [1], [5] that are also presented in a Lagrangian formulation, certain special convex subsets of the nonconvex set of nonlinear variations are effectively singled out and much information is sacrificed in the process.

In an attempt to discover stronger higher order conditions, we at first search directly for controllability theorems that involve possibly nonconvex sets of variations and do not involve Lagrange multipliers. We state such theorems in § 2 and then, in § 3, present more specialized "Lagrangian" conditions which generalize prior results [1], [5]. In § 4 we demonstrate by a simple example that the non-Lagrangian second order controllability conditions of § 2 are more powerful than their Lagrangian predecessors. The proofs are contained in § 5.

Our approach to non-Lagrangian controllability conditions is based on the observation (which follows from Brouwer's fixed point theorem) that if g is a homeomorphism

* Received by the editors December 17, 1984, and in revised form June 12, 1985. This work was partly supported by the National Science Foundation under grant DMS 8400025.

† Department of Mathematics, Northeastern University, Boston, Massachusetts 02115.

of some nhd (neighborhood) V of $\bar{x}$ in $\mathbb{R}^k$ into $\mathbb{R}^k$ then $(g+e)(V)$ contains a nhd of $(g+e)(\bar{x})$ provided e is continuous and its sup norm is sufficiently small. Now let Q be a convex subset of some vector space, $\bar{q} \in Q$, and $\phi_1 \colon Q \to \mathbb{R}^m$. We determine some finite-dimensional subset V of Q and a pth order Taylor approximation of $\phi_1|_V$ at $\bar{q}$ such that only the pth degree part of the Taylor approximation of $\phi_1 - \phi_1(\bar{q})|_V$ fails to vanish identically, and use that part, under appropriate conditions, to construct a related homeomorphism which plays the role of g. We use the error term to construct e and we can conclude that $\phi_1(V)$ contains a nhd of $\phi_1(\bar{q})$. This basic argument is suitably modified to take account of an additional "unilateral" restriction of the form $\phi_2(q) \in C$ in some topological vector space and to apply when we replace the convex set Q by a (possibly nonconvex) "abundant" subset $\mathcal{U}$. (In the case of optimal control, Q may be the set of relaxed controls or any convex set of ordinary controls and $\mathcal{U}$ any set of ordinary control functions in Q that contains concatenations ("splices") of any two of its elements [2, Thm. IV.3.9, p. 285].)

2. Higher order conditions without Lagrange multipliers. Let Q be a convex subset of a vector space, $\mathcal{U} \subset Q$, $\mathcal{Z}$ a topological vector space, $\mathcal{Z}^*$ the topological dual of $\mathcal{Z}$, C a convex subset of $\mathcal{Z}$ with a nonempty interior, $\phi = (\phi_1, \phi_2) \colon Q \to \mathbb{R}^m \times \mathcal{Z}$, $\bar{q} \in Q$ and $\phi_2(\bar{q}) \in C$. Let

$$\mathcal{T}_k = \left\{ (\theta_1, \cdots, \theta_k) \in \mathbb{R}^k \middle| \theta_j \geqq 0, \sum_j \theta_j \leqq 1 \right\}.$$

We shall say that ϕ has a *pth order finite Taylor approximation at* $\bar{q}$ if, for every choice of a positive integer k and of $q_1, \cdots, q_k \in Q$, the function

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

admits a pth order Taylor approximation at 0, i.e. there exists, for each $n = 1, 2, \cdots, p$, (a restriction of) an n-linear symmetric operator $\psi^{(n)}(0) \colon \mathcal{T}_k^n \to \mathbb{R}^m \times \mathcal{Z}$ such that

$$\lim |\theta|^{-p} \left(\psi(\theta) - \left[\psi(0) + \psi'(0)\theta + \cdots + \frac{1}{p!} \psi^{(p)}(0)\theta^p \right] \right) = 0 \quad \text{as } \theta \to 0, \quad \theta \in \mathcal{T}_k.$$

(Observe that this does not require the existence of higher order derivatives of ψ.) The existence of the operators $\psi^{(n)}(0)$ for each choice of k and $q_1, \cdots, q_k$ guarantees that for each $n = 1, \cdots, p$ there exists (a restriction of) an n-linear symmetric operator $\phi^{(n)}(\bar{q}) \colon (Q - \bar{q})^n \to \mathbb{R}^m \times \mathcal{Z}$ such that

$$\phi^{(n)}(\bar{q})(q_1 - \bar{q})^{\alpha_1} \cdots (q_k - \bar{q})^{\alpha_k} = \psi^{(n)}(0)\delta_1^{\alpha_1} \cdots \delta_k^{\alpha_k}$$

if $\alpha_i \geqq 0$, $\alpha_1 + \cdots + \alpha_k = n$ and δ_i is the ith column of the $k \times k$ unit matrix. As customary, we write ϕ', ϕ'', ϕ''' for $\phi^{(1)}$, $\phi^{(2)}$, $\phi^{(3)}$.

Our theorems will be applicable to a proper subset $\mathcal{U}$ of Q if the following condition is satisfied.

Condition 2.1. Q has a uniform structure, ϕ is continuous 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

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

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

We write $\triangleq$ for "is defined as", A^0, co, A, $B(c, r)$, $\bar{B}(c, r)$ for the interior and the convex hull of A and the open and closed ball of center c and radius r, respectively. We identify $\mathbb{R}^k$ with its dual and write $l_1 x$ for $l_i \cdot x$ if $l_1, x \in \mathbb{R}^k$. We denote by $\mathcal{U}$ a subset of Q which is either Q itself or, if Q is a uniform space and ϕ continuous, one satisfying condition 2.1. We also write $\hat{C} \triangleq C - \phi_2(\bar{q})$.

We say that ϕ_1 is *strongly locally* $(\mathcal{U}, \phi_2, C)$*-controllable at* $\bar{q}$ if there exist nhds G_1, G_2 of 0 in $\mathbb{R}^m$, $\mathcal{Z}$ such that

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

Most of our specific second and third order conditions will follow from the following theorem.

THEOREM 2.2. *Let H be a nhd of some $\bar{x} \in \mathbb{R}^k$, $\alpha_1, c > 0$, $t \geqq 1$,*

$$I_n \triangleq \{1, 2, \cdots, i_n\}, \quad y_{ni} \in Q - \bar{q} \quad \forall n = 1, \cdots, p, \quad i \in I_n,$$

and let the functions

$$b_{ni}: H \to [0, c], \qquad f^n = (f_1^n, f_2^n): H \to \mathbb{R}^m \times \mathcal{Z}$$

be continuous. Assume that

(a) $f^n(x) \in \{0\} \times t\hat{C} \quad \forall n = 1, \cdots, p-1, \quad x \in H, \quad f^p(\bar{x}) \in \{0\} \times t\hat{C}^0,$

(b) $\sum_{n=1}^{p} \alpha_1^n \sum_{i \in I_n} b_{ni}(x) \leqq 1 \quad \forall x \in H$

so that

$$\tilde{q}(x, \alpha) \triangleq \bar{q} + \sum_{n=1}^{p} \alpha^n \sum_{i \in I_n} b_{ni}(x) y_{ni} \in Q \quad \forall x \in H, \quad \alpha \in [0, \alpha_1],$$

(c) $\phi(\tilde{q}(x, \alpha)) = \phi(\bar{q}) + \sum_{n=1}^{p} \frac{1}{n!} \alpha^n f^n(x) + \alpha^p d(x, \alpha),$

where $\lim_{\alpha \to 0+} d(x, \alpha) = 0$ *uniformly for* $x \in H$.

Assume, further, that either

(d) f_1^p *is continuously differentiable and the set* $\{\partial f_1^p(\bar{x})/\partial x_j | j = 1, \cdots, k\}$ *spans* $\mathbb{R}^m$,

or

(d′) *there exist a nhd V of 0 in $\mathbb{R}^{m+1}$ and a continuous function $a = (a_1, \cdots, a_k): V \to H$ such that the function*

$$\theta = (\theta_1, \cdots, \theta_{m+1}) \to \left(\sum_{\mu=1}^{m+1} \theta_\mu, f_1^p(a(\theta)) \right),$$

defined for $\theta \in V$, is a homeomorphism and $a(0) = \bar{x}$.

Then ϕ_1 is strongly locally $(\mathcal{U}, \phi_2, C)$-controllable.

Remark. In the special case where $k = m$ and there exists a nhd W of $\bar{x}$ such that $f_1^p|_W$ is a homeomorphism, condition (d′) is satisfied. Indeed, let $h_1, \cdots, h_{m+1}$ be the vertices of a simplex in $\mathbb{R}^m$ containing 0 in its interior and let V be such that

$$a(\theta) \triangleq (f_1^p)^{-1} \left(f_1^p(\bar{x}) + \sum_{\mu=1}^{m+1} \theta_\mu h_\mu \right) \in W \quad \forall \theta \in V.$$

Then $a(\theta)$ yields the function

$$\theta \to (\textstyle\sum \theta_\mu, f_1^p(a(\theta))) = \left(\sum_{\mu=1}^{m+1} \theta_\mu, f_1^p(\bar{x}) + \sum_{\mu=1}^{m+1} \theta_\mu h_\mu \right)$$

on V which is clearly a homeomorphism.

Second and third order conditions derived from Theorem 2.2 are presented below.

THEOREM 2.3. *Let* $I \triangleq \{1, 2, \cdots, i_1\}$, $J \triangleq \{1, 2, \cdots, j_1\}$ *and assume that* ϕ *has a second order finite Taylor approximation at* $\bar{q}$ *and there exist* $y_i, z_j \in Q - \bar{q}$ *for* $i \in I$, $j \in J$ *such that*

(a) $\phi'(\bar{q})y_i \in \{0\} \times \hat{C}$,

(b) $\phi''(\bar{q})\left(\sum_i y_i\right)^2 + 2\phi'(\bar{q})\sum_j z_j \in \{0\} \times \hat{C}^0$,

(c) *the vectors*

$$\phi_1''(\bar{q})\left(\sum_i y_i\right)y_s, \qquad \phi_1'(\bar{q})z_j \quad \text{for } s \in I, j \in J$$

span $\mathbb{R}^m$.

Then ϕ_1 *is strongly locally* $(\mathcal{U}, \phi_2, C)$*-controllable at* $\bar{q}$. *Furthermore, this conclusion remains valid if assumption* (c) *is replaced by*

(c′) *there exist a nhd* V *of* 0 *in* $\mathbb{R}^{m+1}$ *and continuous functions* $a_i, b_j: V \to (0, \infty)$ *such that the function*

$$\theta = (\theta_1, \cdots, \theta_{m+1}) \to \left(\sum_j \theta_j, \phi_1''(\bar{q})\left(\sum_i a_i(\theta)y_i\right)^2 + 2\phi_1'(\bar{q})\sum_j b_j(\theta)z_j\right)$$

is a homeomorphism and $a_i(0) = b_j(0) = 1$ $\forall i$ *and* j.

The following corollary of Theorem 2.3 can be used to rule out candidates for a restricted minimum.

THEOREM 2.4. *Let the assumptions of Theorem* 2.2 *be satisfied with* $\mathcal{Z}$, ϕ_2, C *replaced by* $\mathcal{Z} \times \mathbb{R}$, (ϕ_2, ϕ_0), $C \times (-\infty, \phi_0(\bar{q})]$. *Then there exists* $u \in \mathcal{U}$ *such that*

$$\phi_0(u) < \phi_0(\bar{q}), \quad \phi_1(u) = \phi_1(\bar{q}), \quad \phi_2(u) \in C.$$

Third order controllability conditions, analogous to those of Theorem 2.3, follow.

THEOREM 2.5. *Let* $I = \{1, 2, \cdots, i_1\}$, $J = \{1, \cdots, j_1\}$, $K = \{1, \cdots, k_1\}$, $R = \{1, \cdots, r_1\}$ *and* ϕ *have a third order finite Taylor approximation at* $\bar{q}$. *Assume that, for* $i, s \in I$, $j \in J$, $k \in K$ *and* $r \in R$, *there exist* $y_i, z_j, h_k \in Q - \bar{q}$ *and real numbers*

$$\chi_{isj} = \chi_{sij} \geqq 0, \quad \bar{\zeta}_j = \sum_{i,s} \chi_{isj} > 0 \quad \text{and } v_{rj}$$

such that

(a) $\phi'(\bar{q})y_i \in \{0\} \times \hat{C}$,

(b) $\sum_j v_{rj}\phi_1'(\bar{q})z_j = 0$,

(c) $\phi''(\bar{q})y_iy_s + 2\sum_j \chi_{isj}\phi'(\bar{q})z_j \in \{0\} \times \hat{C}^0$,

(d) $\phi'''(\bar{q})\left(\sum_i y_i\right)^3 + 6\phi''(\bar{q})\left(\sum_i y_i\right)\left(\sum_j \bar{\zeta}_j z_j\right) + 6\phi'(\bar{q})\left(\sum_k h_k\right) \in \{0\} \times \hat{C}^0$,

(e) *the set whose elements are the vectors*

$$3\phi_1'''(\bar{q})\left(\sum_i y_i\right)^2 y_s + 6\phi_1''(\bar{q})\left[\left(\sum_j \bar{\zeta}_j z_j\right)y_s + 2\left(\sum_i y_i\right)\left(\sum_{i,j} \chi_{isj} z_j\right)\right] \quad \text{for } s \in I,$$

$$\sum_j v_{rj}\phi_1''(\bar{q})\left(\sum_i y_i\right)z_j \quad \text{for } r \in R,$$

$$\phi_1'(\bar{q})h_k \quad \text{for } k \in K$$

spans $\mathbb{R}^m$.

Then ϕ_1 is strongly locally $(\mathcal{U}, \phi_2, C)$-controllable at $\bar{q}$. Furthermore, this conclusion remains valid if assumption (e) *is replaced by*

(e′) *there exist a nhd V of 0 in $\mathbb{R}^{m+1}$, continuous functions $a=(a_1, \cdots, a_{i_1})$: $V \to (0,\infty)^{i_1}$, $b=(b_1, \cdots, b_{r_1})$: $V \to \mathbb{R}^{r_1}$, $c=(c_1, \cdots, c_{k_1})$: $V \to (0,\infty)^{k_1}$, and a corresponding function*

$$\theta \to f_1^3(a(\theta), b(\theta), c(\theta)) \triangleq \phi_1'''(\bar{q})\left(\sum_i a_i(\theta) y_i\right)^3 + 6\sum_{i,j} a_i(\theta)\left[\sum_{s,t} \chi_{stij} a_s(\theta) a_t(\theta) + \sum_r v_{rj} b_r(\theta)\right] \phi_1''(\bar{q}) y_i z_j + 6\sum_k c_k(\theta)\phi_1'(\bar{q}) h_k$$

such that

$$\theta = (\theta_1, \cdots, \theta_k) \to \left(\sum_{\mu=1}^{m+1} \theta_\mu, f_1^3(a(\theta), b(\theta), c(\theta))\right)$$

is a homeomorphism and $a_i(0)=1$, $b_r(0)=0$, $c_k(0)=1 \forall i, r, k$.

As in the case of second order conditions, third order nonoptimality conditions analogous to Theorem 2.4 can be derived from Theorem 2.5 by replacing $\mathcal{Z}, \phi_2, C$ with $\mathcal{Z} \times \mathbb{R}, (\phi_2, \phi_0), C \times (-\infty, \phi_0(\bar{q})]$. Fourth and higher order controllability conditions can be similarly derived from Theorem 2.2 but they become progressively more complicated and unwieldy.

3. Lagrangian conditions. Let Y be an arbitrary set, $p \in \{1, 2, \cdots\}$ and $P \subset Y^p$. We refer to P as *symmetric* if $(y_1, \cdots, y_p) \in P$ implies $(y_{\pi(1)}, \cdots, y_{\pi(p)}) \in P$ for every permutation π of $(1, 2, \cdots, p)$.

THEOREM 3.1. *Let ϕ have a second order finite Taylor approximation at $\bar{q}$. Let $Y \subset Q - \bar{q}$, and let P be a symmetric subset of Y^2. Assume that*

(a) $$\phi'(\bar{q}) y \in \{0\} \times \hat{C} \quad \forall y \in Y$$

and, for every finite subset $\{y_1, \cdots, y_{i_1}\}$ of Y and for

$$S \triangleq \{(i,s) | (y_i, y_s) \in P\}, \quad N \triangleq \{1, 2, \cdots, i_1\}^2 \backslash S$$

there exist numbers α_{ab}^{is} $\forall (i,s) \in S$, $(a,b) \in N$ such that

(b) $$\phi''(\bar{q}) y_a y_b - \sum_{(i,s) \in S} \alpha_{ab}^{is} \phi''(\bar{q}) y_i y_s \in \{0\} \times \hat{C} \quad \forall (a,b) \in N,$$

(c) *for each $\sigma = (\sigma_{is})_{(i,s) \in S}$ with $\sigma_{is} = \sigma_{si} > 0$, the system*

$$\eta_i \eta_s + \sum_{(a,b) \in N} \alpha_{ab}^{is} \eta_a \eta_b = \sigma_{is} \quad \forall (i,s) \in S$$

has a solution $\eta(\sigma) = (\eta_1(\sigma), \cdots, \eta_{i_1}(\sigma))$ such that $\eta_i(\sigma) > 0$ and $\sigma \to \eta(\sigma)$ is continuous.

Then either

(d) *ϕ_1 is strongly locally $(\mathcal{U}, \phi_2, C)$-controllable at $\bar{q}$, or*

(e) *there exists $l = (l_1, l_2) \in \mathbb{R}^m \times \mathcal{Z}^*$ such that*

$$l \neq 0, \quad l\phi'(\bar{q}) z \geqq 0 \quad \forall z \in Q - \bar{q}, \qquad l_2[c - \phi_2(\bar{q})] \leqq 0 \quad \forall c \in C,$$

$$l\phi''(\bar{q}) y \bar{y} \geqq 0 \quad \forall (y, \bar{y}) \in P.$$

Somewhat stronger assumptions than in Theorem 3.1 provide some sufficiency criteria for verifying assumptions (b) and (c) above.

DEFINITION 3.2. Let Y be an arbitrary nonempty set and $P \subset Y^p$ symmetric. For any finite set $F \subset Y$ and any ordering $y_1, \cdots, y_{i_1}$ of elements of F, we denote by δ_i the ith row of the $i_1 \times i_1$ unit matrix, and set

$$S \triangleq \{(k_1, k_2, \cdots, k_p) | (y_{k_1}, \cdots, y_{k_p}) \in P\},$$

$$m_{(k_1,\cdots,k_p)} \triangleq \sum_{j=1}^{p} \delta_{k_j} \quad \forall (k_1, \cdots, k_p) \in S, \quad k_1 \leqq k_2 \leqq \cdots \leqq k_p.$$

We shall say that P is *independent* if the set $\{m_{(k_1,\cdots,k_p)} | (k_1, \cdots, k_p) \in S, k_1 \leqq k_2 \leqq \cdots \leqq k_p\}$ is either empty or linearly independent for every finite $F \subset Y$. (We observe that the particular ordering of elements of F does not affect the property of independence of P.)

Remark. Simple examples of independent sets are provided by:

the diagonal set $P \triangleq \{y, y, \cdots, y) | y \in Y\}$;

$p = 2$, $P \triangleq \{(y, p(y)) | y \in Y\}$, where p: $Y \to Y$ is its own inverse; or by

$Y \triangleq \{y_1, y_2, y_3\}$, $p = 2$, $P \triangleq \{(y_1, y_2), (y_2, y_1), (y_1, y_3), (y_3, y_1), (y_2, y_2)\}$.

THEOREM 3.3. *Let ϕ have a second order finite Taylor approximation at $\bar{q}$. Let $Y \subset Q - \bar{q}$ and $P \subset Y^2$ be symmetric and independent. Assume that*

(a) $\phi'(\bar{q})y \in \{0\} \times \hat{C} \quad \forall y \in Y$,

(b) $\phi''(\bar{q})y\bar{y} \in \{0\} \times \hat{C}$ *if* $y, \bar{y} \in Y$ *but* $(y, \bar{y}) \notin P$.

Then the conclusion of Theorem 3.1 *remains valid.*

Remark. Theorems 3.1 and 3.3 generalize [5, Thm. 2.2]. Indeed, the latter is a special case of Theorem 3.3 corresponding to the choice of $P \triangleq \{(y, y) | y \in Y\}$.

Third and higher order Lagrangian conditions with a generality analogous to that of Theorem 3.2 appear to be too complicated to be of much general interest. We therefore provide a third order analogue of Theorem 3.3. Higher order analogues can also be derived in a similar manner but they are more complicated.

THEOREM 3.4. *Let ϕ have a third order finite Taylor approximation at $\bar{q}$. Let $Y \subset Q - \bar{q}$ and $P \subset Y^3$ be symmetric and independent. For every choice of $u, v, w \in Q - \bar{q}$ and every function $(u, v, w) \to F(u, v, w)$: $(Q - \bar{q})^3 \to \mathbb{R}^m \times Z$, let $\mathcal{P}[F(u, v, w)]$ denote the sum $\sum_{(a,b,c)} F(a, b, c)$ over all distinct permutations (a, b, c) of (u, v, w). Assume that for every choice of $y_1, y_2 \in Y$ there exists a point $z(y_1, y_2) = z(y_2, y_1) \in Q - \bar{q}$ such that*

(a) $\phi'(\bar{q})y_1 \in \{0\} \times \hat{C}$,

(b) $\phi''(\bar{q})y_1y_2 + 2\phi'(\bar{q})z(y_1, y_2) \in \{0\} \times \hat{C}$,

(c) $\mathcal{P}[\phi'''(\bar{q})y_1y_2y_3 + 6\phi''(\bar{q})y_3z(y_1, y_2)] \in \{0\} \times \hat{C}$ *if* $y_3 \in Y$, $(y_1, y_2, y_3) \notin P$.

Then either

(d) ϕ_1 *is strongly locally* $(\mathcal{U}, \phi_2, C)$*-controllable at* $\bar{q}$, *or*

(e) *there exists* $l = (l_1, l_2) \in \mathbb{R}^m \times \mathcal{Z}^*$ *such that*

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

$$l\mathcal{P}[\phi'''(\bar{q})y_1y_2y_3 + 6\phi''(\bar{q})y_3z(y_1, y_2)] \geqq 0 \quad \forall (y_1, y_2, y_3) \in P.$$

Remark. In the special case where C is a cone, Y the singlet $\{y_1\}$, $P = \{(y_1, y_1, y_1)\}$ and ϕ_2 is replaced by (ϕ_2, ϕ_0) as in Theorem 2.4, the above results yield Bernstein's necessary condition III of [1, Thm. 6.1, pp. 231–232].

4. An example. Let $Q=\mathbb{R}^3$, $\bar{q}=(0,0,0)$, $q=(x_1,x_2,x_3)$ and

$$\phi_1(q)=(f_1,f_2)(q)+o(|q|^2),$$

where

$$f_1(q)=x_1^2+x_1x_2+x_2x_3+x_3^2,$$

$$f_2(q)=x_1x_3+x_2^2+2x_2x_3.$$

We shall first show that, by Theorem 2.3, $\phi_1(Q)$ contains a nhd of 0. Indeed, we can eliminate the references to ϕ_2, C by choosing $\mathscr{Z}=C=\mathbb{R}$ and $\phi_2=0$. We let

$$y_1=(1,0,0),\quad y_2=(0,-1,0),\quad y_3=(0,0,1)$$

which yields $\tilde{y}=y_1+y_2+y_3=(1,-1,1)$ and

$$\phi_1''(0)y^2=2(f_1,f_2)(\tilde{y})=(0,0).$$

Thus assumptions (a) and (b) of Theorem 2.3 are satisfied.

We have $\phi_1''(\bar{q})=(f_1'',f_2'')$ and

$$f_1''=\begin{pmatrix}2&1&0\\1&0&1\\0&1&2\end{pmatrix},\qquad f_2''=\begin{pmatrix}0&0&1\\0&2&2\\1&2&0\end{pmatrix},$$

and therefore

$$\phi_1''(\bar{q})\tilde{y}y_1=(1,1),\quad \phi_1''(\bar{q})\tilde{y}y_2=(-2,0),\quad \phi_1''(\bar{q})\tilde{y}y_3=(1,-1).$$

Thus assumption (c) is also satisfied and our conclusion follows from Theorem 2.3.

Next we shall show that prior second order conditions in the literature do not yield this result. The most general of these results appears to be [5, Thm. 2.2] (which generalizes the corresponding results of Bernstein [1]). For the present example, that theorem asserts that if $Y\subset Q-\bar{q}=\mathbb{R}^3$ and $\phi_1''(\bar{q})y_1y_2=0$ for $y_1\neq y_2$, $y_1,y_2\in Y$ then either $\phi_1(Q)$ contains a nhd of $(0,0)$ or there exists $l_1\neq 0$ such that

$$l_1\phi_1''(\bar{q})y^2\geqq 0\quad \forall y\in Y.$$

We shall show that such a vector l_1 exists for every admissible choice of Y which will prove that Theorem 2.2 of [5] cannot shed any light on our example. Indeed, if Y contains at most two distinct elements y_1, y_2 then we can always find some $l_1\neq 0$ in $\mathbb{R}^2$ such that

$$l_1\phi_1''(\bar{q})y_1^2\geqq 0\quad\text{and}\quad l_1\phi_1''(\bar{q})y_2^2\geqq 0.$$

Now if y_1, y_2 are two distinct nonzero elements of Y, then

$$\phi_1''(\bar{q})y_1y_2=(f_1'',f_2'')y_1y_2=(0,0)$$

and therefore $y_2\in\mathbb{R}^3$ is normal to the two vectors $f_1''y_1$ and $f_2''y_1$. If these vectors are not parallel, then y_2 is uniquely determined by y_1 except for a constant factor, and then Y cannot contain more than two distinct elements y for which $\phi_1''(\bar{q})y^2\neq 0$.

It remains, therefore, to consider the case when $f_1''y_1$ and $f_2''y_1$ are parallel. Since these two vectors differ from each other and from 0 if $y_1\neq 0$, they will be parallel if

$$(f_1''-\lambda f_2'')y_1=0$$

for some $\lambda\in\mathbb{R}$. Since $\det(f_1''-\lambda f_2'')=2\lambda^3-4\lambda^2-2\lambda-4$ has only one real root $\lambda_1\approx 2.6589672$ and the corresponding eigenvector $y_1=(1,-0.323812,0.630384)$ is unique,

all vectors normal to $f_1''y_1 = \lambda_1 f_2''y_1$ are linear combinations of two linearly independent vectors $u = (0, 1, -1.7400899)$ and $v = (0.5746829, 0, -1)$. Thus, if there are at least three nonzero vectors y_1, y_2, y_3 in Y, both y_2 and y_3 are linear combinations of u and v, say

$$y_2 = \alpha_1 u + \alpha_2 v, \qquad y_3 = \beta_1 u + \beta_2 v$$

and therefore the requirement $\phi_1''(\bar{q})y_2y_3 = 0$ yields

$$\alpha_1\beta_1\phi_1''(\bar{q})u^2 + (\alpha_1\beta_2 + \alpha_2\beta_1)\phi_1''(\bar{q})uv + \alpha_2\beta_2\phi_1''(\bar{q})v^2 = 0.$$

On the other hand, a direct calculation shows that $\phi_1''(\bar{q})u^2$, $\phi_1''(\bar{q})uv$ and $\phi_1''(\bar{q})v^2$ are each $\neq 0$ and

$$0.6018725\phi_1''(\bar{q})u^2 - 1.3783701\phi_1''(\bar{q})uv + \phi_1''(\bar{q})v^2 = 0,$$

where the first two coefficients are unique. Thus we may assume that $\alpha_2 = \beta_2 = 1$ and therefore

$$\alpha_1\beta_1 = 0.6018725, \qquad \alpha_1 + \beta_1 = -1.3783701$$

which cannot be satisfied by any real α_1, β_1. This shows that Y cannot contain more than two distinct nonzero elements.

5. Proofs. Let ϕ have a pth order finite Taylor approximation at $\bar{q}$, let $c > 0$ and $w_n \in Q - \bar{q}$, $b_n \in [0, c]$ $\forall n = 1, \cdots, p$, and let $\alpha_1 > 0$ be such that

$$c \sum_{n=1}^{p} \alpha_1^n \leqq 1.$$

Then

$$\hat{q}(\alpha) \triangleq \bar{q} + \sum_{n=1}^{p} b_n\alpha^n w_n \in Q \quad \forall \alpha \in [0, \alpha_1]$$

and

$$(*) \qquad \phi(\hat{q}(\alpha)) = \phi(\bar{q}) + \sum_{j=1}^{p} \frac{1}{j!}\phi^{(j)}(\bar{q})\left[\sum_{n=1}^{p} b_n\alpha^n w_n\right]^j + \alpha^p d_1(\alpha, b_1, \cdots, b_p)$$

where $\lim_{\alpha\to 0+} d_1(\alpha, b_1, \cdots, b_p) = 0$ uniformly for all choices of $b_1, \cdots, b_p \in [0, c]$ (because $\lim d_1(\alpha, b_1, \cdots, b_p) = 0$ as $|(b_1\alpha, b_2\alpha^2, \cdots, b_p\alpha^p)|$ converges to 0). If we rearrange the terms in (*), we obtain

$$(**) \qquad \phi(\hat{q}(\alpha)) = \phi(\bar{q}) + \sum_{n=1}^{p} \frac{1}{n!}\alpha^n f^n + \alpha^p d(\alpha, b_1, \cdots, b_p),$$

where $\lim_{\alpha\to 0+} d(\alpha, b_1, \cdots, b_p) = 0$ uniformly for all $b_1, \cdots, b_p \in [0, c]$ and, for $z_i \triangleq b_i w_i$,

$$f^n = \sum_{j=1}^{n} \frac{n!}{j!}\phi^{(j)}(\bar{q}) \sum_{m_1+m_2+\cdots+m_j=n} z_{m_1}z_{m_2}\cdots z_{m_j}$$

so that, in particular,

$$(***) \qquad \begin{aligned} f^1 &= \phi'(\bar{q})z_1, \qquad f^2 = \phi''(\bar{q})z_1^2 + 2\phi'(\bar{q})z_2, \\ f^3 &= \phi'''(\bar{q})z_1^3 + 6\phi''(\bar{q})z_1z_2 + 6\phi'(\bar{q})z_3, \end{aligned}$$

etc. We shall henceforth use the expansions in (**), (***) without any further reference to their derivation.

We shall use the notations 2.3(a), 3.1(c) etc. to denote statements (a) of Theorem 2.3, (c) of Theorem 3.1, etc.

LEMMA 5.1. *Let V be a nhd of 0 in $\mathbb{R}^{m+1}$, $t \geqq 1$, $\alpha_1 \geqq 0$ and, for $n = 1, 2, \cdots$,*

$$g = (g_1, g_2)\colon V \to \mathbb{R}^{m+1} \times \mathscr{Z},$$

$$e = (e_1, e_2)\colon V \times [0, \alpha_1] \to \mathbb{R}^{m+1} \times \mathscr{Z},$$

$$h^n = (h_1^n, h_2^n)\colon V \times [0, \alpha_1] \to \mathbb{R}^{m+1} \times \mathscr{Z}.$$

Assume that $g_1\colon V \to g_1(V)$ is a homeomorphism, that g_2, $e(\cdot, \alpha)$ and $h^n(\cdot, \alpha)$ are continuous for each α and n, and that

$$g(0) \in \{0\} \times t\hat{C}^0, \qquad \lim_{\alpha \to 0+} e(x, \alpha) = 0 \quad \textit{uniformly for } x \in V,$$

$$\forall \alpha > 0 \quad \lim_{n \to \infty} h^n(x, \alpha) = 0 \quad \textit{uniformly for } x \in V.$$

Then there exist $\alpha_0 \in (0, \log(1 + t^{-1})]$, $N \in \{1, 2, \cdots\}$ and nhds V_1, U_1, U_2 of 0 in $\mathbb{R}^{m+1}$, $\mathbb{R}^{m+1}$, $\mathscr{Z}$ such that

$$U_1 \subset \{g_1(x) + e_1(x, \alpha_0) + h_1^N(x, \alpha_0) \mid x \in V_1\},$$

$$g_2(x) + e_2(x, \alpha_0) + h_2^N(x, \alpha_0) + U_2 \subset t\hat{C}^0 \quad \forall x \in V_1.$$

Proof. Let $r, s > 0$ and a nhd U_2 of 0 in $\mathscr{Z}$ be such that $\bar{B}(0, r) \subset V$, $g_2^{-1}(\bar{B}(0, s)) \subset \bar{B}(0, r)$, $g_2(\bar{B}(0, r)) + U_2 + U_2 + U_2 \subset t\hat{C}^0$. Let $\alpha_0 \in (0, \min[\alpha_1, \log(1 + t^{-1})])$ be sufficiently small so that

$$|e_1(x, \alpha_0)| \leqq \tfrac{1}{3}s, \qquad e_2(x, \alpha_0) \in U_2 \quad \forall x \in V.$$

Finally, let $N \in \{1, 2, \cdots\}$ be sufficiently large so that

$$|h_1^N(x, \alpha_0)| \leqq \tfrac{1}{3}s, \qquad h_2^N(x, \alpha_0) \in U_2 \quad \forall x \in V.$$

Next we choose an arbitrary $u \in \mathbb{R}^{m+1}$ with $|u| \leqq \frac{1}{3}s$. The equation

$$g_1(x) + e_1(x, \alpha_0) + h_1^N(x, \alpha_0) = u \tag{1}$$

is equivalent to

$$x = g_1^{-1}(u - e_1(x, \alpha_0) - h_1^N(x, \alpha_0)). \tag{2}$$

Since the right-hand side in (2) is a continuous function of x mapping $\bar{B}(0, r)$ into itself, there exists a point $x = \bar{x} \in \bar{B}(0, r)$ satisfying (2) and therefore also (1). Thus

$$U_1 \triangleq \bar{B}(0, \tfrac{1}{3}s) \subset \{g_1(x) + e_1(x, \alpha_0) + h_1^N(x, \alpha_0) \mid x \in \bar{B}(0, r)\}.$$

Furthermore, for each $x \in \bar{B}(0, r)$, we have

$$g_2(x) + e_2(x, \alpha_0) + h_2^N(x, \alpha_0) + U_2 \subset g_2(x) + U_2 + U_2 + U_2 \subset t\hat{C}^0.$$

This shows that our conclusion is satisfied if we set $V_1 \triangleq \bar{B}(0, r)$. Q.E.D.

Proof of Theorem 2.2. We shall first assume that assumption (d) holds. Let $h_1, \cdots, h_{m+1}$ be the vertices of a simplex in $\mathbb{R}^m$ containing 0 in its interior. By assumption (d), there exist numbers $a_{\mu j}$ for $\mu = 1, \cdots, m+1$, $j = 1, \cdots, k$ such that

$$h_\mu = \sum_{j=1}^{k} a_{\mu j} \partial f_1^p(\bar{x}) / \partial x_j.$$

Let

$$\bar{x} \triangleq (\bar{x}_1, \cdots, \bar{x}_k), \qquad \theta \triangleq (\theta_1, \cdots, \theta_{m+1}),$$
$$a_j(\theta) \triangleq \bar{x}_j + \sum_{\mu} a_{\mu j}\theta_\mu, \qquad a(\theta) \triangleq (a_1, \cdots, a_{m+1})(\theta),$$

and let $\tilde{V}$ be a sufficiently small nhd of 0 in $\mathbb{R}^{m+1}$ so that $a(\theta) \in H$ for $\theta \in \tilde{V}$. We set

$$g_1(\theta) \triangleq \left(\sum_{\mu} \theta_\mu, f_1^p(a(\theta))\right) \quad \forall \theta \in \tilde{V}$$

and observe that

$$g_1(0) = (0, 0) \in \mathbb{R} \times \mathbb{R}^m, \qquad g_2(0) \in t\hat{C}^0.$$

We also have

$$\partial g_1(0)/\partial \theta_\mu = \left(1, \sum_{j=1}^{k} a_{\mu j}\partial f_1^p(\bar{x})/\partial x_j\right) = (1, h_\mu).$$

Thus the matrix $g_1'(0)$ is invertible and, by the inverse function theorem, $\theta \to g_1(\theta)$ is a homeomorphism of some nhd V of 0 in $\mathbb{R}^{m+1}$ onto $g_1(V)$. This shows that assumption (d) implies (d') so that we may assume in all cases that (d') is valid.

Next assume that Condition 2.1 holds. Since the functions b_{ni} are continuous, Condition 2.1 implies that, for all $(\theta, \alpha) \in V \times (0, \alpha_1]$, there exists a sequence $(u_\nu(\theta, \alpha))$ in $\mathcal{U}$ such that, for each α,

$$\lim_{\nu} u_\nu(\theta, \alpha) = \tilde{q}(a(\theta), \alpha) \text{ uniformly for } \theta \in V,$$

$$\theta \to u_\nu(\theta, \alpha) \text{ is continuous for each } \nu.$$

For $\theta \in V$ and $\alpha \in [0, \alpha_1]$, we set

$$h^\nu(\theta, \alpha) \triangleq (0, p!\alpha^{-p}[\phi(u_\nu(\theta, \alpha)) - \phi(\tilde{q}(a(\theta), \alpha))] \in \mathbb{R} \times \mathbb{R}^m \times \mathscr{Z},$$
$$g_2(\theta) \triangleq f_2^p(a(\theta)), \quad g \triangleq (g_1, g_2), \quad e_1(\theta, \alpha) \triangleq (0, p!d_1(a(\theta), \alpha)),$$
$$e_2(\theta, \alpha) \triangleq p!d_2(a(\theta), \alpha), \qquad e \triangleq (e_1, e_2)$$

and observe that t, g, e and h^ν satisfy the assumptions of Lemma 5.1. It follows that there exist $\alpha_0 \in (0, \log(1 + t^{-1})]$, $N \in \{1, 2, \cdots\}$ and nhds V_1, U_1, U_2 of 0 in $\mathbb{R}^{m+1}$, $\mathbb{R}^{m+1}$, $\mathscr{Z}$ such that

$$U_1 \subset \{g_1(\theta) + e_1(\theta, \alpha_0) + h_1^N(x, \alpha_0) | \theta \in V_1\}, \tag{1}$$

$$g_2(\theta) + e_2(\theta, \alpha_0) + h_2^N(\theta, \alpha_0) + U_2 \subset t\hat{C}^0 \quad \forall \theta \in V_1. \tag{2}$$

If we set $G_2 = (1/p!)\alpha_0^p U_2$, then relation (2) yields

$$p!\alpha_0^{-p}[\phi_2(u_N(\theta, \alpha_0)) - \phi_2(\bar{q}) - \sum_{n=1}^{p-1} \frac{1}{n!}\alpha_0^n f_2^n(a(\theta)) + G_2] \subset t\hat{C}^0 \quad \forall \theta \in V_1.$$

Since

$$\sum_{n=1}^{p} \frac{t}{n!}\alpha_0^n < t(e^{\alpha_0} - 1) \leqq 1 \quad \text{and} \quad f_2^n(a(\theta)) \in t\hat{C} \quad \forall n = 1, \cdots, p-1,$$

it follows that

$$\phi_2(u_N(\theta, \alpha_0)) + G_2 \subset C \quad \forall \theta \in V_1. \tag{3}$$

Now let

$$G_1 \triangleq \frac{1}{p!}\alpha_0^p\{(\tau_2, \cdots, \tau_{m+1}) | (\tau_1, \tau_2, \cdots, \tau_{m+1}) \in U_1\}.$$

Relation (1) and assumption (a) imply that

$$\phi_1(\bar{q}) + G_1 \subset \{\phi_1(u_N(\theta, \alpha_0)) | \theta \in V_1\}$$

which, together with (3), yields

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

This completes the proof for the case when Condition 2.1 is satisfied. The proof when $\mathcal{U} = Q$ is even simpler; it is obtained by replacing each $u_\nu(\theta, \alpha)$ with $\tilde{q}(a(\theta), \alpha)$. Q.E.D.

Proof of Theorem 2.3. Let $\alpha_1 > 0$ be sufficiently small so that

$$2i_1\alpha_1 + 2j_1\alpha_1^2 \leqq 1, \tag{1}$$

and let

$$\hat{q}(x, \alpha) \triangleq \bar{q} + \alpha \sum_{i \in I} \eta_i y_i + \alpha^2 \sum_{j \in J} \zeta_j z_j,$$

where

$$k = i_1 + j_1, \qquad x = (x_1, \cdots, x_k) = (\eta_1, \cdots, \eta_{i_1}, \zeta_1, \cdots, \zeta_{j_1}),$$
$$x \in [0, 2]^{i_1+j_1}, \qquad \alpha \in [0, \alpha_1].$$

Thus $\alpha \sum \eta_i + \alpha^2 \sum \zeta_j \leqq 1$ and therefore $\hat{q}(x, \alpha) \in Q$. Since ϕ has a second order finite Taylor approximation, we have

$$\begin{aligned}\phi(\hat{q}(x, \alpha)) &= \phi(\bar{q}) + \alpha \sum_i \eta_i \phi'(\bar{q}) y_i \\ &\quad + \frac{1}{2}\alpha^2 \left[\sum_{i,s \in I} \eta_i \eta_s \phi''(\bar{q}) y_i y_s + 2 \sum_j \zeta_j \phi'(\bar{q}) z_j\right] + \alpha^2 d(x, \alpha),\end{aligned} \tag{2}$$

where $\lim_{\alpha \to 0+} d(x, \alpha) = 0$ uniformly for all $x \in [0, 2]^{i_1+j_1}$. Then Theorem 2.2 is applicable with $p = 2$, $c = 2$, $\bar{x} = (1, 1, \cdots, 1)$, $t = 1$ and obvious definitions of y_{ni}, b_{ni}, $f^n(y_{1i} = y_i, y_{2j} = z_j, b_{1i} = \eta_i, b_{2j} = \zeta_j$, $f^n(x)$ the coefficient of $\alpha^n/n!$ in (2), etc.). Indeed, if we choose H as a sufficiently small nhd of $\bar{x}$, then assumption 2.2(a) follows from 2.3(a)-(b) and assumptions 2.2(b)-(c) from (1) and (2). The partial derivatives $\partial f_1^2/\partial \eta_s$, $\partial f_1^2/\partial \zeta_j$ evaluated at $\bar{x} = (1, \cdots, 1)$ are

$$f^2_{1,\eta_s}(\bar{x}) = 2\phi_1''(\bar{q})\left(\sum_i y_i\right) y_s, \qquad f^2_{1,\zeta_j}(\bar{x}) = 2\phi'(\bar{q}) z_j,$$

so that assumption 2.2(d) follows from 2.3(c). Similarly, assumption 2.2(d') follows from 2.3(c'). Q.E.D.

Proof of Theorem 2.5. Let

$$\hat{q}(x, \alpha) \triangleq \bar{q} + \alpha \sum_i \eta_i y_i + \alpha^2 \sum_j \zeta_j z_j + \alpha^3 \sum_k \gamma_k h_k, \tag{1}$$

where

$$\kappa = i_1 + r_1 + k_1, \qquad x = (x_1, \cdots, x_\kappa) = (\eta_1, \cdots, \eta_{i_1}, \omega_1, \cdots, \omega_{r_1}, \gamma_1, \cdots, \gamma_{k_1})$$

and

$$\zeta_j = \sum_{i,s \in I} \chi_{isj} \eta_i \eta_s + \sum_{r \in R} v_{rj} \omega_r. \tag{2}$$

We restrict η_i and γ_k to the interval $[0, 2]$ and ω_r to the interval $[-\Omega, \Omega]$, where Ω is small enough so that $|\sum_r v_{rj}\omega_r| \leqq \frac{1}{2}\bar{\zeta}_j$. Thus $\zeta_j \in \frac{1}{2}\bar{\zeta}_j\,[1, 3]$. The corresponding set of values of x will be denoted by H_1. We then restrict α to the interval $[0, \alpha_1]$, where $\alpha_1 > 0$ is small enough so that

$$\alpha_1 \sum \eta_i + \alpha_1^2 \sum \zeta_j + \alpha_1^3 \sum \gamma_k \leqq 1 \quad \text{for } x \in H_1.$$

Thus $\hat{q}(x, \alpha) \in Q$ for $x \in H_1$ and $0 \leqq \alpha \leqq \alpha_1$ and

$$\begin{aligned}\phi(\hat{q}(x, \alpha)) = \phi(\bar{q}) &+ \alpha \sum_i \eta_i \phi'(\bar{q}) y_i \\ &+ \frac{1}{2}\alpha^2 \left[\sum_{i,s \in I} \eta_i \eta_s \phi''(\bar{q}) y_i y_s + 2 \sum_j \zeta_j \phi'(\bar{q}) z_j \right] \\ &+ \frac{1}{6}\alpha^3 \left[\sum_{i,s,t \in I} \eta_i \eta_s \eta_t \phi'''(\bar{q}) y_i y_s y_t \right. \\ &\left. + 6 \sum_{i,j} \eta_i \zeta_j \phi''(\bar{q}) y_i z_j + 6 \sum_k \gamma_k \phi'(\bar{q}) h_k \right] + \alpha^3 d(x, \alpha),\end{aligned} \tag{3}$$

where $\lim_{\alpha \to 0+} d(x, \alpha) = 0$ uniformly for $x \in H_1$. We denote by $f^1(x), f^2(x), f^3(x)$ the coefficients of $\alpha, \frac{1}{2}\alpha^2, \frac{1}{6}\alpha^3$ above.

We have $f_1^1(x) = 0$ by 2.5(a) and, by (2) and 2.5(b)-(c),

$$\begin{aligned}f_1^2(x) = &\sum_{i,s \in I} \eta_i \eta_s \left[\phi_1''(\bar{q}) y_i y_s + 2 \sum_j \chi_{isj} \phi_1'(\bar{q}) z_j \right] \\ &+ \sum_r \omega_r \sum_j v_{rj} \phi_1'(\bar{q}) z_j = 0 \quad \forall x \in H_1.\end{aligned}$$

Furthermore, we may choose a sufficiently small nhd H of $\bar{x} = (1, \cdots, 1, 0, \cdots 0, 1, \cdots, 1)$ contained in H_1 so that the assumption $f_2^2(\bar{x}) \in \hat{C}^0$ of 2.5(c) implies $f_2^2(x) \in \hat{C}^0$ for all $x \in H$. Thus assumptions 2.2(a)-(c) are satisfied with the obvious choice of b_{ni}. A straightforward computation shows that 2.2(d) is now equivalent to 2.5(e). Finally, 2.2(d') follows directly from (2), (3) and 2.5(e'). Q.E.D.

Proof of Theorem 3.1. Let

$$W \triangleq \{\phi''(\bar{q}) \sum{}' \tau_{uv} uv + 2\phi'(\bar{q}) z \,|\, (u, v) \in P, \tau_{uv} \in [0, 1], z \in Q - \bar{q}\},$$

where $\sum'$ denotes finite sums. The set W is convex and contains 0. Thus, by a variant of the convex separation theorem ([5, Proof of Lemma 4.1, pp. 55-56]), either there exists $l = (l_1, l_2) \in \mathbb{R}^m \times \mathcal{Z}^*$ such that

$$l \neq 0, \quad lw \geqq 0 \quad \forall w \in W, \qquad l_2[c - \phi_2(\bar{q})] \leqq 0 \quad \forall c \in C, \tag{1}$$

or there exist points $\xi^j = (\xi_1^j, \xi_2^j) \in W \,\forall j = 1, \cdots, m+1$ such that

$$0 \in [\text{co}\,\{\xi_1^1, \cdots, \xi_1^{m+1}\}]^0, \qquad \xi_2^j \in \hat{C}^0 \tag{2}$$

and therefore, for some $\beta_j > 0$,

$$\sum_j \beta_j = 1, \qquad \sum_j \beta_j \xi_1^j = 0. \tag{3}$$

If the first alternative holds, then relations (1) yield statement (e).

Now assume that the second alternative holds. Let $y_1, \cdots, y_{i_1}$ be the elements u, v of Y that appear in the expressions for ξ^j, and let S and N be correspondingly defined.

Then each ξ^j is of the form

$$\xi^j = \sum_{(i,s)\in S} \tau_{is}^j \phi''(\bar{q}) y_i y_s + 2\phi'(\bar{q}) z_j,$$

where $z_j \in Q - \bar{q}$ and $\tau_{is}^j \in [0, 1]$. Since $\phi''(\bar{q})$ is symmetric, we may assume that $\tau_{is}^j = \tau_{si}^j \ \forall j, i, s$. We can also assume that $\tau_{is}^j > 0 \ \forall j, i, s$ because a sufficiently small perturbation of the ξ^j will not affect the validity of relations (2).

Let

$$\sigma_{si}(\theta) = \sigma_{is}(\theta) \triangleq \sum_j (\beta_j + \theta_j)\tau_{is}^j \quad \forall (i, s) \in S, \quad \theta = (\theta_1, \cdots, \theta_{m+1}) \in \mathbb{R}^{m+1}.$$

Since $\sigma_{is}(0) > 0 \ \forall (i, s) \in S$, we may determine a compact nhd V of 0 in $\mathbb{R}^{m+1}$ such that

$$\tfrac{1}{2}\sigma_{is}(0) \leqq \sigma_{is}(\theta) \leqq 2\sigma_{is}(0), \qquad |\theta_j| \leqq \tfrac{1}{2}\beta_j \quad \forall (i, s) \in S, \quad \theta \in V.$$

Then, by assumption (c), there exist continuous functions

$$\theta \to \hat{\eta}_i(\theta) \triangleq \eta_i(\sigma(\theta)) \colon V \to \mathbb{R} \quad \forall i = 1, \cdots, i_1$$

such that

$$\hat{\eta}_i(\theta)\hat{\eta}_s(\theta) + \sum_{(a,b)\in N} \alpha_{ab}^{is} \hat{\eta}_a(\theta)\hat{\eta}_b(\theta) = \sigma_{is}(\theta) \quad \forall (i, s) \in S. \tag{4}$$

We deduce from (a)-(c) and (4) that there exist $t \geqq 1$ and continuous $v_1, v_2 \colon V \to t\hat{C}$ such that

$$\sum_{i=1}^{i_1} \hat{\eta}_i(\theta)\phi'(\bar{q}) y_i = (0, v_1(\theta)),$$

$$\begin{aligned}
&\sum_{i,s=1}^{i_1} \hat{\eta}_i(\theta)\hat{\eta}_s(\theta)\phi''(\bar{q}) y_i y_s + 2\sum_j (\beta_j + \theta_j)\phi'(\bar{q}) z_j \\
&\quad = \sum_{(a,b)\in N} \hat{\eta}_a(\theta)\hat{\eta}_b(\theta) \sum_{(i,s)\in S} \alpha_{ab}^{is} \phi''(\bar{q}) y_i y_s + (0, v_2(\theta)) \\
&\qquad + \sum_{(i,s)\in S} \hat{\eta}_i(\theta)\hat{\eta}_s(\theta)\phi''(\bar{q}) y_i y_s + 2\sum_j (\beta_j + \theta_j)\phi'(\bar{q}) z_j \\
&\quad = \sum_{(i,s)\in S} [\hat{\eta}_i(\theta)\hat{\eta}_s(\theta) + \sum_{(a,b)\in N} \alpha_{ab}^{is} \hat{\eta}_a(\theta)\hat{\eta}_b(\theta)]\phi''(\bar{q}) y_i y_s \\
&\qquad + 2\sum_j (\beta_j + \theta_j)\phi'(\bar{q}) z_j + (0, v_2(\theta)) \\
&\quad = \sum_{(i,s)\in S} \sigma_{is}(\theta)\phi''(\bar{q}) y_i y_s + 2\sum_j (\beta_j + \theta_j)\phi'(\bar{q}) z_j + (0, v_2(\theta)) \\
&\quad = \sum_j (\beta_j + \theta_j)\xi^j + (0, v_2(\theta)).
\end{aligned} \tag{5}$$

Now let

$$\tilde{q}(\theta, \alpha) \triangleq \bar{q} + \alpha \sum_i \hat{\eta}_i(\theta) y_i + \alpha^2 \sum_j (\beta_j + \theta_j) z_j, \tag{6}$$

where $\alpha \in [0, \alpha_1]$ and $\alpha_1 > 0$ is chosen sufficiently small so that

$$\alpha_1 \sum_i \hat{\eta}_i(\theta) + \alpha_1^2 \sum_j (\beta_j + \theta_j) \leqq 1 \quad \forall \theta \in V. \tag{7}$$

Then $\tilde{q}(\theta, \alpha) \in Q$ and, by (5),

$$\begin{aligned}
\phi(\tilde{q}(\theta, \alpha)) &= \phi(\bar{q}) + \alpha(0, v_1(\theta)) \\
&\quad + \frac{1}{2}\alpha^2 \left[\sum_j (\beta_j + \theta_j)\xi^j + (0, v_2(\theta)) \right] + \alpha^2 d(\theta, \alpha),
\end{aligned} \tag{8}$$

where $\lim_{\alpha\to 0+} d(\theta,\alpha)=0$ uniformly for $\theta\in V$. We can verify that Theorem 2.2 is applicable, with 0, θ, V replacing $\bar{x}$, x, H and with

$$p=2,\quad y_{1i}=y_i,\quad y_{2j}=z_j,\quad b_{1i}(\theta)\triangleq\hat{\eta}_i(\theta),\quad b_{2j}(\theta)\triangleq\beta_j+\theta_j,$$

$$f_2^1(\theta)=v_1(\theta),\qquad f^2(\theta)=\sum_j(\beta_j+\theta_j)\xi^j+(0,v_2(\theta)).$$

Indeed, assumption 2.2(a) follows from $v_i(\theta)\in t\hat{C}$, (2), (3) and 3.1(a); 2.2(b) from (6) and (7); 2.2(c) from (8); and 2.2(d) from the first relation of (2) which implies the linear independence of $\{(1,\xi_1^1),\cdots,(1,\xi_1^{m+1})\}$. Thus Theorem 2.2 implies alternative 3.1(d). Q.E.D.

Proof of Theorem 3.3. This theorem follows directly from Theorem 3.1. Indeed, let $\{y_1,\cdots,y_{i_1}\}$ be a finite subset of Y, and let S and N be defined as in Theorem 3.1 and $m_{(i,s)}$ as in Definition 3.2. Set $\alpha_{ab}^{is}=0$ for all $(i,s)\in S$, $(a,b)\in N$. Then assumptions 3.1(a), 3.1(b) follow from 3.3(a), 3.3(b), respectively. Furthermore, the system

$$\eta_i\eta_s=\sigma_{is}\quad\forall(i,s)\in S,\quad i\leqq s, \tag{1}$$

for arbitrary $\sigma_{is}>0$ is equivalent to

$$\log\eta_i+\log\eta_s=\log\sigma_{is}\quad\forall(i,s)\in S,\quad i\leqq s$$

which is a linear system in $\log\eta_i$ with a matrix whose rows are the $m_{(i,s)}$ and therefore independent. Thus system (1) has a solution $\eta(\sigma)=(\eta_1(\sigma),\cdots,\eta_{i_1}(\sigma))$ for all $\sigma=(\sigma_{is})$ with $\sigma_{is}=\sigma_{si}>0$ such that $\sigma\to\eta(\sigma)$ is continuous. This shows that assumption 3.1(c) is also satisfied. Q.E.D.

Proof of Theorem 3.4. Let P be partitioned into equivalence classes, two elements (u,v,w), (u_1,v_1,w_1) belonging to the same class if (u_1,v_1,w_1) is a permutation of (u,v,w). Let P' be formed by selecting one representative from each equivalence class. We set

$$W\triangleq\{\sum{}'\tau_{uvw}\mathcal{P}[\phi'''(\bar{q})uvw+6\phi''(\bar{q})wz(u,v)]+6\phi'(\bar{q})h\mid$$

$$(u,v,w)\in P',\ \tau_{uvw}\in[0,1],\ h\in Q-\bar{q}\},$$

where $\sum'$ denotes finite sums. The set W is convex and contains 0. Thus (as in the proof of Theorem 3.1) either there exists $l=(l_1,l_2)\in\mathbb{R}^m\times\mathcal{Z}^*$ such that

$$l\neq 0,\quad lw\geqq 0\quad\forall w\in W,\quad l_2[c-\phi_2(\bar{q})]\leqq 0\quad\forall c\in C, \tag{1}$$

or there exist points $\xi^\mu=(\xi_1^\mu,\xi_2^\mu)\in W\,\forall\mu=1,\cdots,m+1$ such that

$$0\in[c\{\xi_1^1,\cdots,\xi_1^{m+1}\}^0,\qquad \xi_2^\mu\in C^0-\phi_2(\bar{q})\subset\hat{C}^0 \tag{2}$$

and therefore, for some $\beta_\mu>0$,

$$\sum_\mu\beta_\mu=1,\qquad\sum_\mu\beta_\mu\xi_1^\mu=0. \tag{3}$$

If the first alternative holds, then relations (1) yield statement (e).

Now assume that the second alternative holds. Let $y_1, \cdots, y_{i_1}$ be the elements u, v, w of Y that appear in the expressions for ξ^μ. Let

$$S \triangleq \{(i, s, t) | (y_i, y_s, y_t) \in P'\}, \qquad z_{is} \triangleq z(y_i, y_s).$$

Then each ξ^μ can be written in the form

$$\xi^\mu = \sum_{(i,s,t)\in S} \tau^\mu_{ist} \mathscr{P}[\phi'''(\bar{q}) y_i y_s y_t + 6\phi''(\bar{q}) y_i z_{st}] + 6\phi'(\bar{q}) h_\mu,$$

where $h_\mu \in Q - \bar{q}$ and $\tau^\mu_{ist} \in [0, 1]$. We can also assume that all $\tau^\mu_{ist} > 0$ because small perturbations of the ξ^μ do not affect the validity of relations (2).

Let

$$\sigma_{ist}(\theta) \triangleq \sum_\mu (\beta_\mu + \theta_\mu) \tau^\mu_{ist} \quad \forall (i, s, t) \in S, \quad \theta = (\theta_1, \cdots, \theta_{m+1}) \in \mathbb{R}^{m+1}.$$

Since $\sigma_{ist}(0) > 0 \forall (i, s, t) \in S$, we may determine a compact nhd V of 0 in $\mathbb{R}^{m+1}$ such that

$$\tfrac{1}{2}\sigma_{ist}(0) \leqq \sigma_{ist}(\theta) \leqq 2\sigma_{ist}(0), \quad |\theta_\mu| \leqq \tfrac{1}{2}\beta_\mu \quad \forall \theta \in V.$$

Then, for each $\theta \in V$, the system

$$\eta_i \eta_s \eta_t = \sigma_{ist}(\theta) \quad \forall (i, s, t) \in S \tag{4}$$

is equivalent to the linear system in $\log \eta_i$,

$$\log \eta_i + \log \eta_s + \log \eta_t = \log \sigma_{ist}(\theta) \quad \forall (i, s, t) \in S$$

and, because of the independence of P, this last system has a matrix with the linearly independent rows $m_{(i,s,t)}$. Therefore, system (4) has a solution $\hat{\eta}(\theta) = (\hat{\eta}_1(\theta), \cdots, \hat{\eta}_{i_1}(\theta))$ which is continuous in θ.

Now let

$$\begin{aligned} \tilde{q}(\theta, \alpha) = \bar{q} &+ \alpha \sum_{i=1}^{i_1} \hat{\eta}_i(\theta) y_i + \alpha^2 \sum_{i,s=1}^{i_1} \hat{\eta}_i(\theta) \hat{\eta}_s(\theta) z_{is} \\ &+ \alpha^3 \sum_{\mu=1}^{m+1} (\beta_\mu + \theta_\mu) h_\mu \quad \forall \theta \in V, \quad \alpha \in [0, \alpha_1], \end{aligned} \tag{5}$$

where $\alpha_1 \in (0, \frac{1}{2}]$ is sufficiently small so that $\tilde{q}(\theta, \alpha) \in Q \forall \alpha \in [0, \alpha_1]$. Then, with the summations below carried out over all $i, s, t \in \{1, \cdots, i_1\}$, $j \in \{1, \cdots, j_1\}$ and $\mu \in \{1, \cdots, m+1\}$, we have

$$\begin{aligned} \phi(\tilde{q}(\theta, \alpha)) = \phi(\bar{q}) &+ \alpha \sum_i \phi'(\bar{q}) y_i \\ &+ \frac{1}{2}\alpha^2 \left[\phi''(\bar{q}) \left(\sum_i \hat{\eta}_i(\theta) y_i \right)^2 + 2\phi'(\bar{q}) \sum_{i,s} \hat{\eta}_i(\theta) \hat{\eta}_s(\theta) z_{is} \right] \\ &+ \frac{1}{6}\alpha^3 \left[\phi'''(\bar{q}) \left(\sum_i \hat{\eta}_i(\theta) y_i \right)^3 + 6\phi''(\bar{q}) \sum_{i,s,t} \hat{\eta}_i(\theta) \hat{\eta}_s(\theta) \hat{\eta}_t(\theta) z_{st} y_i \right. \\ &\qquad \left. + 6\phi'(\bar{q}) \sum_\mu (\beta_\mu + \theta_\mu) h_\mu \right] + \alpha^3 d(\theta, \alpha), \end{aligned}$$

where $\lim_{\alpha \to 0+} d(\theta, \alpha) = 0$ uniformly for $\theta \in V$. It follows, by (a)–(c) and (4) that there

exist $\tau \geqq 1$ and continuous v_1, v_2, v_3: $V \to \tau\hat{C}$ such that

$$\begin{aligned}
\phi(\tilde{q}(\theta,\alpha)) &= \phi(\bar{q}) + \alpha(0, v_1(\theta)) + \alpha^2(0, v_2(\theta)) \\
&\quad + \frac{1}{6}\alpha^3 \Bigg\{ \sum_{(i,s,t)\in S} \hat{\eta}_i(\theta)\hat{\eta}_s(\theta)\hat{\eta}_t(\theta)\mathcal{P}[\phi'''(\bar{q})y_i y_s y_t + 6\phi''(\bar{q})y_i z_{st}] \\
&\qquad\qquad + 6\phi'(\bar{q}) \sum_{\mu} (\beta_\mu + \theta_\mu) h_\mu \Bigg\} \\
&\quad + \alpha^3(0, v_3(\theta)) + \alpha^3 d(\theta,\alpha) \\
&= \phi(\bar{q}) + (0, \alpha v_1(\theta) + \alpha^2 v_2(\theta) + \alpha^3 v_3(\theta)) \\
&\quad + \frac{1}{6}\alpha^3 \sum_{(i,s,t)\in S} \sum_{\mu} (\beta_\mu + \theta_\mu)\tau_{ist}^{\mu}\mathcal{P}[\phi'''(\bar{q})y_i y_s y_t + 6\phi''(\bar{q})y_i z_{st}] \\
&\quad + 6\phi'(\bar{q}) \sum_{\mu} (\beta_\mu + \theta_\mu) h_\mu + \alpha^3 d(\theta,\alpha) \\
&= \phi(\bar{q}) + (0, \alpha v_1(\theta) + \alpha^2 v_2(\theta) + \alpha^3 v_3(\theta)) \\
&\quad + \frac{1}{6}\alpha^3 \sum_{\mu} (\beta_\mu + \theta_\mu)\xi^\mu + \alpha^3 d(\theta,\alpha).
\end{aligned}$$

It follows now from Theorem 2.2 (just as in the proof of Theorem 3.1) that ϕ_1 is strongly locally $(\mathcal{U}, \phi_2, C)$-controllable. 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] J. WARGA, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.

[3] ———, *Optimization and controllability without differentiability assumptions*, this Journal, 21 (1983), pp. 837-855.

[4] ———, *Controllability, extremality and abnormality in nonsmooth optimal control*, J. Optim. Theory Appl., 41 (1983), pp. 239-260.

[5] ———, *Second order controllability and optimization with ordinary controls*, this Journal, 23 (1985), pp. 49-60.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 4, July 1986

011

NILPOTENT APPROXIMATIONS OF CONTROL SYSTEMS AND DISTRIBUTIONS*

HENRY HERMES†

Abstract. A constructive method is given to approximate the vector fields in a nonlinear control system, which is linear in the controls, by a system of similar form and on the same state space, with the describing vector fields of the approximating system generating a Lie algebra which has a certain, relevant, sub-algebra nilpotent. Given a k-dimensional distribution on R^n, say locally defined near zero, the method leads to an approximating k-distribution which has nilpotent basis, agrees with the original distribution at zero, and each derived distribution of the approximating distribution and original distribution also agree at zero.

Key words. nilpotent Lie algebras, system approximation

AMS(MOS) subject classifications. 93B10, 49E99

Introduction. The goal of this paper is to show how to constructively approximate (in a sense to be made precise) the vector fields $X^0, X^1, \cdots, X^k$ in a control system on R^n of the form

$$\dot{x} = X^0(x) + \sum_{i=1}^{k} u_i X^i(x), \quad x(0) = 0, \quad X^0(0) = 0, \tag{1}$$

by vector fields $Y^0, \cdots, Y^k$ for which either the Lie algebra, $L(Y^0, \cdots, Y^k)$, generated by $Y^0, \cdots, Y^k$ or a certain appropriate subalgebra of this algebra, is nilpotent.

Assume all vector fields are real analytic. For any set, $\mathcal{S}$, of vector fields on R^n (or an n-manifold M^n), $L(\mathcal{S})$ will denote the Lie algebra they generate, $\mathcal{S}(0)$ the elements of $\mathcal{S}$ evaluated at 0; $[X, Y] = (\text{ad } X, Y)$ the commutator, or Lie product, of the vector fields X, Y and, inductively, $(\text{ad}^{l+1} X, Y) = [X, (\text{ad}^l X, Y)]$.

For the vector fields of system (1), assign the weight $w_0 = 0$ to X^0 and $w_i = 1$ to X^i for $i = 1, \cdots, k$. The weight assigned to a commutator will be the sum of the weights of its constituent factors. For $i = 1, 2, \cdots$ let

$$\mathcal{S}_X^i \text{ be the set of commutators of } X^0, \cdots, X^k \text{ of weight } i. \tag{2}$$

Explicitly, $\mathcal{S}_X^1 = \{(\text{ad}^j X^0, X^i): j \geqq 0, i = 1, \cdots, m\}$. Recall [1] that dim span $\mathcal{S}_X^1(0) = n$ is the "first" order sufficient condition for local controllability of system (1) along the solution, denoted $(\exp tX^0)(0)$, of the uncontrolled equation $\dot{x} = X^0(x)$, $x(0) = 0$. The sets $\mathcal{S}_X^2, \cdots$ play a significant role in higher order local controllability conditions, [2], [3]. We shall assume, mainly for convenience,

$$\dim L(\mathcal{S}_X^1)(0) = n \tag{3}$$

which implies that for all $t > 0$ the attainable set of system (1) at time t has nonempty interior in R^n.

Assume dim span $\mathcal{S}_X^1(0) = n_1$, dim span $(\mathcal{S}_X^1 \cup \mathcal{S}_X^2)(0) = n_2$ and N is the smallest integer such that dim span $(\mathcal{S}_X^1 \cup \cdots \cup \mathcal{S}_X^N)(0) = n$. The notation X^π will denote a product of specific vector fields from the set $X^0, \cdots, X^k$ taken in a specific order. (If we deal with vector fields $Y^0, \cdots, Y^k$, $\mathcal{S}_Y^i$ will denote the equivalent set to $\mathcal{S}_X^i$ and Y^π the equivalent commutator to X^π.)

* Received by the editors February 12, 1985, and in final form June 24. This research was supported by the National Science Foundation under grant DMS-8500941.

† Department of Mathematics, University of Colorado, Boulder, Colorado 80309.

THEOREM 1. *Let* $X^{\pi_1}, \cdots, X^{\pi_{n_1}} \in \mathcal{S}_X^1$ *be such that* $X^{\pi_1}(0), \cdots, X^{\pi_{n_1}}(0)$ *are linearly independent. Adjoint* $X^{\pi_{n_1+1}}, \cdots, X^{\pi_{n_2}} \in \mathcal{S}_X^2$ *such that* $X^{\pi_1}(0), \cdots, X^{\pi_{n_2}}(0)$ *are independent and continue in this way until one has* $X^{\pi_1}, \cdots, X^{\pi_n}$ *with* $X^{\pi_1}(0), \cdots, X^{\pi_n}(0)$ *independent. Then there exist vector fields* $Y^0, \cdots, Y^k$ *on* R^n *which describe the approximating system*

$$\dot{x} = Y^0(x) + \sum_{i=1}^{k} u_i Y^i(x), \qquad x(0) = 0 \tag{4}$$

and satisfy (a) $Y^0(0) = 0$, (b) $Y^{\pi_i}(0) = X^{\pi_i}(0)$, $i = 1, \cdots, n$, (c) $L(\mathcal{S}_Y^1)$ *is nilpotent. Furthermore, the* Y^i *can be explicitly constructed and have polynomial coefficients relative to the preferred coordinates of the construction.*

Remark 1. It is well known (see Example 5) that one cannot, in general, find vector fields $Y^0, \cdots, Y^k$ with $L(Y^0, \cdots, Y^k)$ nilpotent and such that $Y^\pi(0) = X^\pi(0)$ for all commutators X^π of length $\leqq N$ (or, as seen in this example, even of length $\leqq 1$).

Remark 2. If one deals with distributions, say $D^k(x) = \operatorname{span}\{X^1(x), \cdots, X^k(x)\}$ defined in a neighborhood of zero in R^n, one would have $X^1(0), \cdots, X^k(0)$ independent. Then $X^{\pi_1} = X^1, \cdots, X^{\pi_k} = X^k$; $\mathcal{S}_X^1 = \{X^1, \cdots, X^k\}$ while $\mathcal{S}_X^2$ would consist of all products of pairs of elements of $\mathcal{S}_X^1$, etc. Let $D^{k,1}(x) = D^k(x)$, $D^{k,2}(x) = \operatorname{span}(\mathcal{S}_X^1 \cup \mathcal{S}_X^2)(x)$, etc. Theorem 1 then yields the existence of $Y^1, \cdots, Y^k$ such that if $E^k = E^{k,1}(x) = \operatorname{span}\{Y^1(x), \cdots, Y^k(x)\}$, $E^{k,2}(x) = \operatorname{span}(\mathcal{S}_Y^1 \cup \mathcal{S}_Y^2)(x)$, etc. then E^k is a k-distribution having a "nilpotent basis" $Y^1, \cdots, Y^k$ and which approximates D^k in the sense that $E^{k,i}(0) = D^{k,i}(0)$, $i = 1, \cdots, N$, i.e. the derived distributions agree to order N at 0.

Our main concern with Theorem 1 is its application to control systems of the form (1). The solution of (1) at time t for given control u can be expressed (see [4]) as

$$x(t, u) = (\exp tX^0) \circ y(t, u) \tag{5}$$

where $y(t, u)$ satisfies the auxiliary equation

$$\dot{y} = \sum_{i=1}^{k} u_i(t) \sum_{\nu=0}^{\infty} \frac{(-t)^\nu}{\nu!} (\operatorname{ad}^\nu X^0, X^i)(y), \qquad y(0) = 0. \tag{6}$$

If $L(\mathcal{S}_X^1)$ is nilpoent, the solution $y(t, u)$ of (6) can be constructively expressed ([5, § 3] or [3]) as a finite composition

$$y(t, u) = (\exp F_1(t, u) W^1) \circ \cdots \circ (\exp F_m(t, u) W^m)(0) \tag{7}$$

where $\{W^1, \cdots, W^m\}$ is an "ordered basis" for $L(\mathcal{S}_X^1)$. Substituting (7) in (5) yields an expression for the system output as a function of the control, analogous (equivalent) to the Volterra series expansion which is also finite when $L(\mathcal{S}_X^1)$ is nilpotent. Thus $L(\mathcal{S}_X^1)$ nilpotent implies local analysis of system (1) is greatly simplified. Theorem 1 yields $L(\mathcal{S}_Y^1)$ nilpotent for the approximating system (4).

From the assignment of weights, $\mathcal{S}_X^j$ consists of commutators containing j factors from $\{X^1, \cdots, X^k\}$. Each such factor has coefficient a control component so one can view an element of $\mathcal{S}_X^j$ as corresponding to a jth power of the control. In this sense, the approximation can loosely be considered as related to a truncation, to order N, of the expansion of the output $x(t, u)$ in terms of the control.

Nilpotent approximations have received a great deal of attention during the past decade. Krener [6] showed that for any integer $r \geqq 0$ there exists a bilinear system of the form

$$\dot{W} = A_0 W + \sum_{i=1}^{k} u_i A_i W, \qquad W(0) = \mathrm{id}, \tag{8}$$

on $Gl(m, R)$, for sufficiently large m depending on r, having $L(A_0, \cdots, A_k)$ nilpotent, and a linear map $l: gl(m, R) \to R$ such that for all $i_1, \cdots, i_s \in \{1, \cdots, k\}$ with $s \leqq r$, $l([A_{i_1}, [\cdots[A_{i_{s-1}}, A_{i_s}]\cdots]) = [X^{i_1}, [\cdots[X^{i_{s-1}}, X^{i_s}]\cdots](0)$. From this he deduces the existence of a smooth map $\lambda: Gl(m, R) \to R^n$, $\lambda(\mathrm{id}) = 0$, such that for any control u, $|(x(t, u) - \lambda(W(t, u))| \leqq Kt^{r+1}$. That is, the trajectories of the "lifted" nilpotent system (8) when mapped to R^n via λ, locally approximate those of system (1) to order $(r+1)$.

Rothschild and Stein [7] in their study of hypoelliptic operators of the form $L = \sum_{i=1}^{k} (X^i)^2$, again lift (or extend) the vector fields $X^1, \cdots, X^k$ to a sufficiently large space where their Lie algebra is "free to order r" and then achieve a nilpotent approximation theorem.

More recently, Crouch [8], [9], approximates the system (1) by a system of the same form, but on a space of dimension $\sum_{i=1}^{N} \dim \operatorname{span} \mathscr{S}_X^i(0)$, whose describing vector fields generate a solvable algebra containing $L(\mathscr{S}^1)$ as a nilpotent ideal. The input-output map of his "lifted" approximating system is the Nth order truncated Volterra series of the input-output map of system (1).

For computations, the liftings to a higher-dimensional manifold in the above papers create a problem since the map from this manifold back to R^n is, in general, difficult to explicitly obtain. Bressan [10], obtains a nilpotent approximation to a system of the form $\dot{x} = \sum_{i=1}^{k} u_i X^i(x)$ in which the approximating system (again of this form) lives on R^n. His results are closely related to those of this paper. Indeed, it is shown in Example 3 that our construction, when applied to Bressan's example [10] yields the same approximating system as obtained by Bressan. Many of the ideas and methods of this paper were greatly influenced by those in the aforementioned papers of Crouch and Bressan.

1. Dilations and their related nilpotent algebras. Consider R with coordinates $x = (x_1, \cdots, x_n)$. A dilation, δ_t, is a map $\delta_t: R^n \to R^n$ of the form $\delta_t x = (t^{r_1}x_1, \cdots, t^{r_n}x_n)$ where we assume all integers $r_1 \geqq 1$ and $r_1 \leqq r_2 \leqq \cdots r_n$. We first briefly review some standard definitions.

DEFINITION 1. A polynomial h on (R^n, δ_t) is homogeneous of degree j with respect to δ_t if $t^j h(x) = h(\delta_t x)$. Denote by H_j all polynomials homogeneous of degree j; $H_j = 0$ if $j < 0$.

Note that with $\delta_t x$ as above

$$h \in H_j \text{ implies } \frac{\partial h}{\partial x_i} \in H_{j-r_i}. \tag{9}$$

DEFINITION 2. A vector field X on (R^n, δ_t) is homogeneous of degree m if $Xh \in H_{j-m}$ whenever $h \in H_j$, $j = 0, 1, \cdots$. Thus if $X(x) = \sum_{i=1}^{m} a_i(x)\, \partial/\partial x_i$, X is homogeneous of degree m if $a_i \in H_{r_i - m}$, $i = 1, \cdots, n$.

Example 1. Consider R^3 with $\delta_t x = (tx_1, tx_2, t^3x_3)$. Then $h \in H_0$ implies $h(x) = c$, a constant; $h \in H_1$ implies $h(x) = c_1x_1 + c_2x_2$; $h \in H_2$ implies $h(x) = c_1x_1^2 + c_2x_1x_2 + c_3x_2^2$; $h \in H_3$ implies $h(x) = c_1x_1^3 + c_2x_1^2x_2 + c_3x_1x_2^2 + c_4x_2^3 + c_5x_3$, etc.

Remark 3. If X is homogeneous of degree m, Y homogeneous of degree l and $h \in H_j$, then $Xh \in H_{j-m}$, $Yh \in H_{j-l}$ so $[X, Y]h = Y(Xh) - X(Yh) \in H_{j-(m+l)}$ i.e. $[X, Y]$ is homogeneous of degree $(m+l)$.

PROPOSITION 1. *Let $\mathscr{L}_i$ denote the real linear span of all vector fields homogeneous of degree $\geqq i$ with respect to a given dilation $\delta_t x = (t^{r_1}x_1, \cdots, t^{r_n}x_n)$ on R^n. Then each $\mathscr{L}_i$ is a Lie algebra and $\mathscr{L}_1$ is a nilpotent ideal in $\mathscr{L}_0$.*

Indeed, $\mathscr{L}_0 \supset \mathscr{L}_1 \supset \cdots \supset \mathscr{L}_{r_n} \supset \mathscr{L}_{r_n+1} = \{0\}$ while Remark 3 shows $[\mathscr{L}_1, \mathscr{L}_i] \subset \mathscr{L}_{i+1}$ so the descending central series of $\mathscr{L}_1$ terminates showing that $\mathscr{L}_1$ is nilpotent.

Remark 4. If $Y^{\pi}(x)=\sum_{i=1}^{n} b_i(x)\,\partial/\partial x_i \in \mathscr{L}_j$, $1 \leqq j \leqq r_n$ relative to the dilation $\delta_t x = (t^{r_1}x_1, \cdots, t^{r_n}x_n)$, then $b_i(0)=0$ for $1 \leqq i \leqq k$, k the largest integer such that $r_k < j$.

Remark 5. Let $X(x)=\sum_{j=1}^{n} a_j(x)\,\partial/\partial x_j \in \mathscr{L}_1$ with $\mathscr{L}_1$ defined as above relative to the dilation $\delta_t x=(t^{r_1}x_1, \cdots, t^{r_n}x_n)$. Then $\partial^l a_j(x)/\partial x_{i_1}\cdots\partial x_{i_l}=0$ if $r_{i_1}+\cdots+r_{i_l} \geqq r_j$, $l=1, \cdots, n$. If $\alpha=(\alpha_1, \cdots, \alpha_n)$ is a multi-index and we define $|\alpha|=\sum_{i=1}^{n} r_i\alpha_i$ this means $X \in \mathscr{L}_1$ has the form

$$X(x)=\sum_{j=1}^{n}\left(\sum_{|\alpha|<r_j} c_{j,\alpha}x^{\alpha}\right)\frac{\partial}{\partial x_j}$$

where the $c_{j,\alpha}$ are constants and $x^{\alpha}=(x_1^{\alpha_1}\cdots x_n^{\alpha_n})$, Bressan [10] gives a direct (computational) proof that vector fields of this form generate a nilpotent algebra.

2. The constructive proof of Theorem 1.

A. Assume the vector fields $X^0, \cdots, X^k$ are given relative to some coordinates, say $y=(y_1, \cdots, y_n)$. We first make a linear change of coordinates to $x=(x_1, \cdots, x_n)$ such that relative to these coordinates, $X^{\pi_i}(0)=\partial/\partial x_i$, $i=1, \cdots, n$. Constructively, if in the original coordinates we had $X^{\pi_i}(y)=\sum_{j=1}^{n} a_{ij}(y)\,\partial/\partial y_j$, $i=1, \cdots, n$, let $a_i(y)$ denote the vector $a_i(y)=(a_{i1}(y), \cdots, a_{in}(y))$. Form the matrix, B, having columns $a_1(0), \cdots, a_n(0)$. Since $X^{\pi_1}(0), \cdots, X^{\pi_n}(0)$ are linearly independent, B is nonsingular. Make the linear coordinate change $x=B^{-1}y$ in which case $X^{\pi_i}(x)=\sum_{j=1}\alpha_{ij}(x)\,\partial/\partial x_j$ with $\alpha_i(x)=B^{-1}a_i(Bx)$, i.e. in these coordinates $X^{\pi_i}(0)=\partial/\partial x_i$.

B. Choose a dilation on R^n induced by system (1) as follows. Recall the assumption $\dim L(\mathscr{S}_X^1)(0)=n$ and notation $\dim \operatorname{span}(\mathscr{S}_X^1 \cup \cdots \cup \mathscr{S}_X^i)(0)=n_i$. This assumption implies $n_1 \geqq 1$ and there exists a smallest integer N such that $n_N=n$. Choose the dilation (in the preferred coordinates as above).

(i) $\delta_t x=(tx_1, \cdots, tx_{n_1}, t^2x_{n_1+1}, \cdots, t^2x_{n_2}, \cdots, t^N x_n)$ or, for future notational convenience,

(ii) $\delta_t x=(t^{r_1}x_1, \cdots, t^{r_{n_1}}x_{n_1}, t^{r_{n_1+1}}x_{n_1+1}, \cdots, t^{r_N}x_n)$ where r_i is the exponent of t in (i) of the coefficient of x_i.

C. Let $X^0(x)=\sum_{j=1}^{n} a_{0j}(x)\,\partial/\partial x_j$ be our vector field of weight zero. The approximating vector field $Y^0(x)=\sum_{j=0}^{n} b_{0j}(x)\,\partial/\partial x_j$ is obtained by letting b_{0j} be the truncation of the expansion of a_{0j} in homogeneous polynomials (relative to δ_t) of degree $\leqq r_j$. Specifically, if we expand $a_{0j}(x)=\sum_{l=0}^{\infty} a_{0j}^l(x)$ where $a_{0j}^l \in H_l$, then

$$b_{01}(x)=\sum_{l=0}^{r_1} a_{01}^l(x), \cdots, b_{0n}(x)=\sum_{l=0}^{r_n} a_{0n}^l(x). \tag{10}$$

If $X^i(x)=\sum_{j=1}^{n} a_{ij}(x)\,\partial/\partial x_j$, $1 \leqq i \leqq k$, is a vector field of weight one, the approximating vector field $Y^i(x)=\sum_{j=0}^{n} b_{ij}(x)\,\partial/\partial x_j$ where b_{ij} is the truncation of the expansion of a_{ij} in homogeneous polynomials of degree $\leqq r_j-1$. In the above notation

$$b_{i1}(x)=\sum_{l=0}^{r_1-1} a_{i1}^l(x), \cdots, b_{in}(x)=\sum_{l=0}^{r_n-1} a_{in}^l(x). \tag{11}$$

D. (a) From the construction, if $X^{\pi}(x)=\sum_{i=1}^{n} a_i(x)\,\partial/\partial x_i$ is any commutator of weight $j \geqq 1$, the corresponding "approximating" commutator $Y^{\pi}(x)=\sum_{i=1}^{n} b_i(x)\,\partial/\partial x_i$ satisfies $a_i(0)=b_i(0)$, $n_{j-1}+1 \leqq i \leqq n$.

(b) If $Y^{\pi}(x)=\sum_{i=1}^{n} b_i(x)\,\partial/\partial x_i$ is an approximating commutator of weight $j \geqq 1$, then $Y^{\pi} \in \mathscr{L}_j$ so (see Remark 4) $b_i(0)=0$, $1 \leqq i \leqq n_{j-1}$.

(c) Now consider the specifically chosen commutators X^{π_l}, $1 \leqq l \leqq n$, which, in our preferred coordinates, satisfy $X^{\pi_l}(0)=\sum_{i=1}^{n} a_i(0)\,\partial/\partial x_i=\partial/\partial x_l$. If X^{π_l} has weight

j, then $n_{j-1}+1 \leqq l \leqq n_j$ (see (i)) so if $Y^{\pi_l}(x) = \sum_{i=1}^n b_i(x)\,\partial/\partial x_i$ is the approximating commutator, from (b), $b_i(0) = 0$ for $1 \leqq i \leqq n_{j-1}$ while from (a), $b_i(0) = a_i(0)$ for the remaining indices. Thus $Y^{\pi_l}(0) = X^{\pi_l}(0)$.

Finally, $L(\mathscr{S}_Y^1)$ is nilpotent by Proposition 1, while $X^0(0) = 0$ certainly implies $Y^0(0) = 0$.

3. Examples. For notational and printing ease, throughout this section a vector field $X(x) = \sum_{i=1}^n a_i(x)\,\partial/\partial x_i$ will be written as $X(x) = (a_1(x), \cdots, a_n(x))$.

Example 2. The purpose of this example is to show that one cannot replace the statement $Y^{\pi_i}(0) = X^{\pi_i}(0)$, $i = 1, \cdots, n$ of Theorem 1 by $Y^{\pi_i}(0) = X^{\pi_i}(0)$ for all commutators X^π of length $\leqq N$.

Consider R^3 with $X^0(x) = (0, x_1, x_1^2/2)$, $X^1 = (1, 0, 0)$, $X^2(x) = (x_2, 0, 0)$. Then $(\mathrm{ad}X^0, X^1)(x) = (0, 1, x_1)$, $(\mathrm{ad}^2X^0, X^1) = 0$, $[[X^0, X^1], X^1] = (0, 0, 1)$, $[X^2, [X^0, X^1]](x) = (1, 0, -x_2)$. One can choose $X^{\pi_1} = X^1 \in \mathscr{S}_X^1$, $X^{\pi_2} = (\mathrm{ad}\,X^0, X^1) \in \mathscr{S}_X^1$ and $X^{\pi_3} = [[X^0, X^1], X^1] \in \mathscr{S}_X^2$. This system induces the dilation $\delta_t x = (tx_1, tx_2, t^2x_2)$ and the coordinates are already the preferred coordinates which make $X^{\pi_i}(0) = \partial/\partial x_i$, $i = 1, \cdots, 3$.

A polynomial $h \in H_0$ implies $h = c$, $h \in H_1$ implies $h(x) = c_1x_1 + c_2x_2$, $h \in H_2$ implies $h(x) = c_1x_1^2 + c_2x_1x_2 + c_3x_2^2 + c_4x_3$. The formulae (10), (11) then yield the approximating vector fields $Y^0 = X^0$, $Y^1 = X^1$, $Y^2 = 0$. Thus $Y^{\pi_i}(0) = X^{\pi_i}(0)$, $i = 1, 2, 3$, $Y^0(0) = X^0(0)$ but $[X^2, [X^0, X^1]](0) \neq [Y^2, [Y^0, Y^1]](0)$.

Example 3. (An example from [10].) Consider R^2 with $X^0(x) = 0$, $X^1 = (1, 0) \in \mathscr{S}_X^1$, $X^2(x) = (\sin(x_1 - x_2), 1 - \cos(x_1 + x_2)) \in \mathscr{S}_X^1$. Then $[X^1, X^2](x) = (\cos(x_1 - x_2)), \sin(x_1 - x_2)) \in \mathscr{S}_X^2$ and one finds $1 = n_1 = \dim \mathrm{span}\, \mathscr{S}_X^1(0) = \dim \mathrm{span}\, (\mathscr{S}_X^1 \cup \mathscr{S}_X^2)(0)$ so $n_2 = n_1$. Further computation gives $[[X^2, X^1], X^1](0) = (0, 1)$ and this product has weight 3. Thus $\delta_t x = (tx_1, t^3x_2) = (t^{r_1}x_1, t^{r_2}x_2)$ while $X^{\pi_1} = X^1$, $X^{\pi_2} = [[X^2, X^1], X^1]$ with coordinates already as desired.

A polynomial $h \in H_0$ implies $h(x) = c$, $h \in H_1$ implies $h(x) = c_1x_1$, $h \in H_2$ implies $h(x) = c_1x_1^2$, $h \in H_3$ implies $h(x) = c_1x_1^3 + c_2x_2$.

Expanding the components of X^2 in a Taylor series about zero gives $\sin(x_1 - x_2) = (x_1 - x_2) - (x_1 - x_2)^3/3! + \cdots$, $1 - \cos(x_1 + x_2) = (x_1 + x_2)^2/2 + \cdots$. The approximating vector fields are $Y^1 = X^1$; the first component of Y^2 consists of terms in $\sin(x_1 - x_2)$ which belong to $H_{r_1 - 1} = H_0$, the second component of Y^2 consists of terms in $(1 - \cos(x_1 + x_2))$ which are in H_j with $0 \leqq j \leqq r_2 - 1 = 2$. Thus $Y^2(x) = (0, x_1^2/2)$. This agrees with the nilpotent approximation in [10].

Example 4. The case $\dim \mathrm{span}\, \mathscr{S}_X^1(0) = n$, i.e. the system satisfies the first order, local controllability condition.

Consider the system (1) on R^n and assume $\dim \mathrm{span}\, \mathscr{S}_X^1(0) = n$. Then $X^{\pi_1}, \cdots, X^{\pi_n}$ can be chosen from $\mathscr{S}_X^1$, these all have weight 1, and in the preferred coordinates $x = (x_1, \cdots, x_n)$ satisfy $X^{\pi_i}(0) = \partial/\partial x_i$, $i = 1, \cdots, n$. The induced dilation is $\delta_t x = (tx_1, \cdots, tx_n)$; polynomials $h \in H_0$ are constants while $h(x) \in H_1$ implies $h(x) = \sum_{i=1}^n c_i x_i$. The approximation $Y^0(x) = X_x(0)x$ where $X_x(x)$ denotes the Jacobian matrix of partial derivatives. For $1 \leqq i \leqq k$, if $X^i(x) = \sum_{j=1}^n a_{ij}(x)\,\partial/\partial x_j$, then $Y^i = \sum_{j=1}^n b_{ij}\,\partial/\partial x_j$ where b_{ij} is the constant term in the Taylor series expansion of $a_{ij}(x)$ about zero. Thus, here, the approximating system is the "simplest" linearization of the original system.

Example 5. Let $X^0(x) = (x_1, x_1^2/2)$, $X^1 = (1, 0)$ on R^2. Then $(\mathrm{ad}\,X^0, X^1)(x) = (1, x_1)$ while $(\mathrm{ad}^\nu X^0, X^1)(0) = X^1(0)$ for all $\nu \geqq 0$; hence $L(X^0, X^1)$ is not nilpotent nor could one find vector fields Y^0, Y^1 whose commutators agree with those of X^0, X^1 at zero to order one and with $L(Y^0, Y^1)$ nilpotent. Further computation shows $[[X^0X^1], X^1] = (0, 1) \in \mathscr{S}_X^2$ so one may choose $X^{\pi_1} = X^1 \in \mathscr{S}_X^1$; $X^{\pi_2} = [[X^0, X^1], X^1]$

and the induced dilation is $\delta_t x = (tx_1, t^2x_2)$. The approximating vector fields are $Y^0 = X^0$, $Y^1 = X^1$; $L(Y^0, Y^1)$ is not nilpotent but $L(\mathcal{S}_Y^1)$ is nilpotent as promised.

REFERENCES

[1] H. HERMES, *On local and global controllability*, this Journal, 17 (1974), pp. 252-261.

[2] H. SUSSMANN, *Lie brackets and local controllability: A sufficient condition for scalar-input systems*, this Journal, 21 (1983), pp. 686-713.

[3] H. HERMES, *Control systems which generate decomposable Lie algebras*, J. Differential Equations, 44 (1982), pp. 166-187.

[4] K. T. CHEN, *Decomposition of differential equations*, Math. Ann., 146 (1962), pp. 263-278.

[5] H. HERMES, *Local controllability and sufficient conditions in singular problems* II, this Journal, 14 (1976), pp. 1049-1062.

[6] A. J. KRENER, *Bilinear realizations of input-output maps*, this Journal, 13 (1975), pp. 827-832.

[7] L. P. ROTHSCHILD AND E. M. STEIN, *Hypoelliptic differential operators and nilpotent groups*, Acta Math., 137 (1976), pp. 247-320.

[8] P. E. CROUCH, *Solvable approximations to control systems*, this Journal, 22 (1984), pp. 40-54.

[9] ———, *Solvable approximations of control systems*, Proc. 23rd IEEE Conference on Decision and Control, 2 (1984), pp. 775-780.

[10] A. BRESSAN, *Local asymptotic approximation of nonlinear control systems*, preprint.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 4, July 1986

012

LEGENDRE-TAU APPROXIMATIONS FOR FUNCTIONAL DIFFERENTIAL EQUATIONS*

KAZUFUMI ITO† AND RUSSELL TEGLAS‡

Abstract. In this paper we consider the numerical approximation of solutions to linear retarded functional differential equations using the so-called Legendre-tau method. The functional differential equation is first reformulated as a partial differential equation with a nonlocal boundary condition involving time-differentiation. The approximate solution is then represented as a truncated Legendre series with time varying coefficients which satisfy a certain system of ordinary differential equations. The method is very easy to code and yields very accurate approximations. Convergence is established, various numerical examples are presented, and comparison between the latter and cubic spline approximations is made.

Key words. Legendre-tau approximation, retarded systems, numerical convergence

AMS (MOS) subject classifications. Primary 65N20, 65R20; secondary 34K99, 93C20

1. Introduction. In this paper we consider Legendre-tau approximations of solutions of functional differential equations (FDEs). The tau method, invented by Lanczos in 1938 [12], is one of several approximation techniques which are referred to as a spectral method [9]. Spectral methods have been used in numerical computations for a wide class of partial differential equations (PDEs). In this paper, we view the original FDE as a PDE $u_t = u_x$ with boundary conditions which involve time differentiation. The Legendre-tau method is based upon representing the approximate solution as a truncated series of Legendre polynomials. The evolution equation for the expansion coefficients is then determined by substituting the series into the above PDE and by imposing the boundary conditions. To our knowledge, the use of the tau method in approximating solutions of FDEs and the specific manner in which it is used (i.e., applying the tau method to a reformulated PDE with boundary conditions involving time evolution) are new. The idea of formulating FDEs as Cauchy problems on an appropriate Hilbert space is not new. Within this framework, Banks and Kappel [2] make use of approximation results from linear semigroup theory (in particular, the Trotter-Kato theorem) to establish the convergence of numerical schemes based upon splines.

In this paper our considerations are restricted to linear autonomous FDEs of retarded type. Our ideas can be extended to (i) nonautonomous, (ii) nonlinear, (iii) neutral-type or (iv) integro-differential systems. We add here that the tau method should prove useful in dealing with partial differential equations with boundary conditions which involve time differentiation. Our main goal is the application of the tau method to optimal control and parameter estimation problems. As will be discussed in § 7 the tau method may offer considerable improvements over other methods (e.g., those discussed in [2], [3]) in many instances. One reason for this is that the semigroup $\{S(t): t \geqq 0\}$ associated with a retarded FDE has the property that the range of $S(t)$ is contained in $\mathcal{D}(\mathcal{A}^k)$ for each $t \geqq kr$ where $\mathcal{A}$ is the infinitesimal generator of $\{S(t): t \geqq 0\}$ with domain $\mathcal{D}(\mathcal{A})$ and r is the longest delay time appearing in the FDE. Thus, the regularity of solutions increases with time. In § 3, we will see that, in such a case, approximations by orthogonal polynomials are quite powerful.

* Received by the editors August 11, 1983, and in revised form May 29, 1985. This research was supported by the National Aeronautics and Space Administration under NASA contract NAS1-17070 while the authors were in residence at ICASE, NASA Langley Research Center, Hampton, Virginia 23665.

† Division of Applied Mathematics, Brown University, Providence, Rhode Island 02917.

‡ Department of Mathematics, University of Vermont, Burlington, Vermont 05405.

The following is a brief summary of the contents of this paper. In § 2, we review the equivalence results between FDEs and abstract Cauchy problems on the product space $\mathbb{R}^n \times L_2$. In § 3, we recall various properties of Legendre polynomials including certain estimates which are needed to establish convergence. Section 4 is concerned with the development of the numerical scheme based upon the Legendre-tau approximation for solving the class of FDEs under consideration. In order to establish the numerical convergence of such approximations to the actual solution in the $\mathbb{R}^n \times L_2$ norm, we first show, in § 5, that convergence holds in a stronger norm under the special assumptions that the initial data is sufficiently regular and that the inhomogeneous forcing term f is identically zero. In § 6, we then extend our result to the inhomogeneous case wherein $f \in L_2^{\text{loc}}$ and show that the sequence obtained by truncating the last term in each of the Legendre-tau approximations converges to the actual solution in the $\mathbb{R}^n \times L_2$ norm whenever the initial data lies in $\mathbb{R}^n \times L_2$. Finally, in § 7, we present numerical results and compare these with results for cubic spline approximations discussed in [2].

Throughout this paper the following notation will be used. $r > 0$ stands for the longest delay time appearing in the FDE. The Hilbert space of $\mathbb{R}^n$-valued square integrable functions on the interval $[a, b]$ is denoted by $L_2([a, b]; \mathbb{R}^n)$. When the underlying space and interval can be understood from the context, we will abbreviate the notation and simply write L_2. $L_2^{\text{loc}}([0, \infty); \mathbb{R}^n)$, or L_2^{loc}, is the space of $\mathbb{R}^n$-valued locally square integrable functions on the semi-infinite interval $[0, \infty)$. H^k is the Sobolev space of $\mathbb{R}^n$-valued functions f on a compact interval with $f^{(k-1)}$ absolutely continuous and $f^{(k)} \in L_2$, with norm $\|f\|_{H^k} = (\sum_{i=0}^{k} \|f^{(i)}\|_{L_2}^2)^{1/2}$. The Banach space of $\mathbb{R}^n$-valued continuous functions on the interval $[-r, 0]$ is denoted by C. We denote by Z the product space $\mathbb{R}^n \times L_2([-r, 0]; \mathbb{R}^n)$. Given an element $z \in Z$, $\eta \in \mathbb{R}^n$ and $\phi \in L_2$ denote the two coordinates of z: $z = (\eta, \phi)$. The bracket $\langle \cdot, \cdot \rangle_H$ stands for the inner product in the Hilbert space H, and the subscript for the underlying Hilbert space will be omitted when understood from the context. $\|\cdot\|$ denotes the norm for elements of a Banach space and for operators between Banach spaces, while $|\cdot|$ denotes the Euclidean norm in $\mathbb{R}^n$.

If X and Y are Banach spaces, then the space of bounded operators from X to Y is denoted by $\mathscr{L}(X, Y)$. $\mathscr{D}(\mathscr{A})$ denotes the domain of a linear operator $\mathscr{A}$. χ_I denotes the characteristic function of the interval I. Given a measurable function $x: [-r, \infty) \to \mathbb{R}^n$ and $t \geqq 0$, the function $x_t: [-r, 0] \to \mathbb{R}^n$ is defined by $x_t(\theta) = x(t + \theta)$, $\theta \in [-r, 0]$. Finally, for any function ϕ of the independent variable θ, we shall use either $\dot{\phi}$ or $\partial\phi/\partial\theta$ to denote the derivative of ϕ with respect to θ.

2. Linear retarded differential equations. In this section, we state the type of equations to be considered and recall some results for such equations which are important for the discussion to follow.

Given $(\eta, \phi) \in Z$ and $f \in L_2^{\text{loc}}([0, \infty); \mathbb{R}^n)$, we consider the initial value problem

$$\begin{aligned} &\frac{d}{dt}x(t) = Dx_t + f(t), \qquad t \geqq 0, \\ &x(0) = \eta, \qquad x_0 = \phi, \end{aligned} \tag{2.1}$$

where $D: \mathscr{D}(D) \subseteq L_2([-r, 0]; \mathbb{R}^n) \to \mathbb{R}^n$ has the form

$$D\phi = \int_{-r}^{0} d\mu(\theta)\phi(\theta) \tag{2.2}$$

with μ a matrix-valued function of bounded variation on $[-r, 0]$. As an example, consider

$$\mu(\theta) = \sum_{i=0}^{m} A_i \chi_{(\theta_i, 0]}(\theta) + \int_{-r}^{\theta} A(s)\, ds, \tag{2.3}$$

where $-r = \theta_m < \cdots < \theta_0 = 0$ and A_i and $A(\cdot)$ are $n \times n$ matrices, the elements of the latter being integrable on $[-r, 0]$. Then

$$Dx_t = \sum_{i=0}^{m} A_i x(t+\theta_i) + \int_{-r}^{0} A(\theta) x(t+\theta)\, d\theta, \qquad t \geqq 0. \tag{2.4}$$

It is well known [3], [5], [8] that for $(\eta, \phi) \in Z$ and $f \in L_2^{\text{loc}}$, (2.1) admits a unique solution

$$x \in L_2([-r, T]; \mathbb{R}^n) \cap H^1([0, T]; \mathbb{R}^n)$$

for any $T \geqq 0$ such that

$$z(t) = (x(t), x_t) \tag{2.5}$$

is a Z-valued continuous function which depends continuously on $(\eta, \phi) \in Z$ and $f \in L_2^{\text{loc}}$ for each $t \geqq 0$.

For $t \geqq 0$, define $S(t): Z \to Z$ by $S(t)(\eta, \phi) = (x(t), x_t)$ whre x is the homogeneous solution of (2.1) (i.e., $f \equiv 0$). Then $\{S(t): t \geqq 0\}$ forms a strongly continuous semigroup on Z. The following results are now standard if one deals with FDEs in the state space Z ([3], [4], [15]).

LEMMA 2.1.

(i) *If $\mathcal{A}$ denotes the infinitesimal generator of $\{S(t): t \geqq 0\}$, then*

$$\mathcal{D}(A) = \{(\eta, \phi) \in Z: \dot{\phi} \in L_2 \text{ and } \eta = \phi(0)\}$$

and for $(\phi(0), \phi) \in \mathcal{D}(A)$

$$\mathcal{A}(\phi(0), \phi) = (D\phi, \dot{\phi}).$$

(ii) *The spectrum $\sigma(\mathcal{A})$ of $\mathcal{A}$ only consists of point spectrum and $\lambda \in \sigma(\mathcal{A})$ if and only if $\det \Delta(\lambda) = 0$ where*

$$\Delta(\lambda) = \lambda I - \int_{-r}^{0} e^{\lambda\theta}\, d\mu(\theta).$$

For each λ in the resolvent set $\rho(\mathcal{A})$ of $\mathcal{A}$, the resolvent of $\mathcal{A}$ is given by

$$(\lambda I - \mathcal{A})^{-1} z = (\psi(0), \psi) \quad \text{for } z = (\eta, \phi) \in Z, \tag{2.6}$$

with

$$\psi(\theta) = e^{\lambda\theta} b + \int_{\theta}^{0} e^{\lambda(\theta - s)} \phi(s)\, d\theta,$$

$$b = \Delta^{-1}(\lambda)\left[\eta + D\left(\int_{\cdot}^{0} e^{\lambda(\cdot - s)} \phi(s)\, ds\right)\right].$$

(iii) *If $z(0) = (\phi(0), \phi) \in \mathcal{D}(\mathcal{A})$ and $f \in L_2^{\text{loc}}$, then $z(t) = S(t)z(0) + \int_0^t S(t-s)\mathcal{B}f(s)\, ds$ satisfies the equation*

$$\frac{dz}{dt}(t) = \mathcal{A}z(t) + \mathcal{B}f(t), \qquad t \geqq 0, \tag{2.7}$$

in Z where $\mathcal{B}:\mathbb{R}^n \to Z$ is defined by

$$\mathcal{B}f = (f, 0) \in Z \quad \text{for } f \in \mathbb{R}^n.$$

(iv) *For $k \geqq 2$, the domain of the kth power of $\mathcal{A}$ is contained in the set*

$$\{(\phi(0), \phi) \in Z: \dot{\phi}(0) = D\phi \text{ and } \phi \in H^k\},$$

and is dense in Z.

Since $\mathcal{A}$ is closed, $\mathcal{D}(\mathcal{A})$ itself with the graph norm

$$\|z\|^2_{\mathcal{D}(\mathcal{A})} = \|z\|^2_Z + \|\mathcal{A}z\|^2_Z$$

is a Hilbert space. Then the restriction of the semigroup $\{S(t): t \geqq 0\}$ to $\mathcal{D}(\mathcal{A})$ also forms a C_0-semigroup on $\mathcal{D}(\mathcal{A})$. Let us denote by X the space $H^1[-r, 0]$ equipped with the norm

$$\|\phi\|_1^2 = |\phi(0)|^2 + \int_{-r}^{0} |\dot{\phi}|^2 \, d\theta$$

and inner product

$$\langle \phi, \psi \rangle_1 = \langle \phi(0), \psi(0) \rangle_{\mathbb{R}^n} + \int_{-r}^{0} \langle \dot{\phi}(\theta), \dot{\psi}(\theta) \rangle_{\mathbb{R}^n} \, d\theta.$$

It is readily established that X is isomorphic to $\mathcal{D}(\mathcal{A}) \subset Z$: consider the map $E: \mathcal{D}(\mathcal{A}) \to X$ defined by

$$E(\phi(0), \phi) = \phi \in X \quad \text{for } (\phi(0), \phi) \in \mathcal{D}(\mathcal{A}).$$

Then $E^{-1}: X \to \mathcal{D}(\mathcal{A}) \subset Z$ is given by

$$E^{-1}\phi = (\phi(0), \phi) \in \mathcal{D}(\mathcal{A}) \quad \text{for } \phi \in X,$$

and it is easy to show that there exists constants $0 < c \leqq C < \infty$ such that

$$c\|\phi\|_1 \leqq \|E^{-1}\phi\|_{\mathcal{D}(\mathcal{A})} \leqq C\|\phi\|_1, \tag{2.8}$$

i.e., E is an isomorphism.

Since $\mathcal{A}$ generates a C_0-semigroup $\{S(t): t \geqq 0\}$ on $\mathcal{D}(\mathcal{A})$ with graph norm, $E\mathcal{A}E^{-1}$ generates a C_0-semigroup $\{ES(t)E^{-1}: t \geqq 0\}$ on X. Moreover, it is easily seen that

$$\mathcal{D}(E\mathcal{A}E^{-1}) = \{\phi \in X: \dot{\phi} \in X \text{ and } \dot{\phi}(0) = D\phi\}$$

and for $\phi \in \mathcal{D}(E\mathcal{A}E^{-1})$.

$$E\mathcal{A}E^{-1}\phi = \dot{\phi}. \tag{2.9}$$

For the sake of convenience, we will use the notation

$$\tilde{S}(t) = ES(t)E^{-1} \quad \text{and} \quad \tilde{\mathcal{A}} = E\mathcal{A}E^{-1}. \tag{2.10}$$

3. Properties of Legendre polynomials. In this section, we review some properties of Legendre polynomials ([11], [13], e.g.).

The Legendre polynomial of degree k, $p_k(x)$, $-1 \leqq x \leqq 1$, can be defined as the solution of the differential equation

$$\frac{d}{dx}\left((1-x^2)\frac{dp}{dx}(x)\right) + k(k+1)p(x) = 0, \tag{3.1}$$

which satisfies $p(1) = 1$. Thus, $p_0(x) = 1$, $p_1(x) = x$, $p_2(x) = \frac{1}{2}(3x^2 - 1)$, and so on. The

Legendre polynomials $\{p_k\}_{k\geqq 0}$ satisfy the orthogonality relation

$$\int_{-1}^{1} p_k(x)p_l(x)\,dx = \frac{2}{2k+1}\delta_{kl}, \tag{3.2}$$

and they form a basis for $L_2(-1,1)$: any $f\in L_2(-1,1)$ can be written as

$$f=\sum_{k\geqq 0} f_k p_k$$

where

$$f_k=\frac{2k+1}{2}\int_{-1}^{1} f(x)p_k(x)\,dx$$

with

$$\|f\|_{L_2}^2=\sum_{k\geqq 0}\frac{2}{2k+1}f_k^2.$$

They possess the recursion formula

$$(k+1)p_{k+1}(x)=(2k+1)xp_k(x)-kp_{k-1}(x). \tag{3.3}$$

From this, we have

$$p_k(\pm 1)=(\pm 1)^k, \qquad |p_k(x)|\leqq 1, \quad |x|\leqq 1,$$

and

$$\dot{p}_k(\pm 1)=(\pm 1)^{k+1}k(k+1)/2.$$

If f is represented as

$$f=\sum_{k=0}^{N} f_k p_k,$$

then

$$\dot{f}=\sum_{k=0}^{N-1} b_k p_k \quad \text{where } b_k\equiv(2k+1)\sum_{\substack{j=k+1\\ j+k\,\text{odd}}}^{N} f_j. \tag{3.4}$$

For any positive integer N, let P^N be the orthogonal projection of L_2 onto the subspace spanned by $\{p_k\}_{k=0}^N$. Then we have the following error estimates [6].

LEMMA 3.1. *For any real $s\geqq 0$, there exists a constant K such that*

$$\|f-P^N f\|_{L_2}\leqq KN^{-s}\|f\|_{H^s}.$$

LEMMA 3.2. *For any real s and σ such that $1\leqq s\leqq\sigma$, there exists a constant K such that*

$$\|f-P^N f\|_{H^s}\leqq KN^{2s-\sigma-1/2}\|f\|_{H^\sigma}.$$

The next lemma gives an error estimate in the supremum norm.

LEMMA 3.3. *For any positive integer m there exists a constant K such that*

$$|f(x)-P^N f(x)|\leqq KN^{-2m+1}\|f\|_{H^{2m}}.$$

Proof. Let us denote by $\mathscr{L}$ the differential operator

$$(\mathscr{L}f)(x)=\frac{d}{dx}\left((1-x^2)\frac{df}{dx}(x)\right).$$

From (3.1), we have

$$\int_{-1}^{1} f(x)p_k(x)\,dx = \left(-\frac{1}{k(k+1)}\right)^m \int_{-1}^{1} (\mathscr{L}^m p_k)(x) f(x)\,dx, \qquad k \geqq 1.$$

Since $\mathscr{L}$ is symmetric,

$$\int_{-1}^{1} f(x)p_k(x)\,dx = \left(-\frac{1}{k(k+1)}\right)^m \int_{-1}^{1} (\mathscr{L}^m f)(x) p_k(x)\,dx, \qquad k \geqq 1.$$

It follows that if $\{f_n\}_{n\geqq 0}$ are the Legendre coefficients of f, then

$$f_k = \left(-\frac{1}{k(k+1)}\right)^m \frac{2k+1}{2} \int_{-1}^{1} (\mathscr{L}^m f)(x) p_k(x)\,dx \equiv \left(-\frac{1}{k(k+1)}\right)^m g_k,$$

where $\{g_k\}_{k\geqq 0}$ are the Legendre coefficients of $\mathscr{L}^m f$. Thus, for any $M > N$ and $|x| \leqq 1$,

$$\begin{aligned}
|P^M f(x) - P^N f(x)| &= \left| \sum_{k=N+1}^{M} f_k p_k(x) \right| \\
&\leqq \sum_{k=N+1}^{M} |f_k| = \sum_{j=N+1}^{M} \left(\frac{1}{k(k+1)}\right)^m |g_k| \\
&\leqq \left(\sum_{k=N+1}^{M} \left(\frac{1}{k(k+1)}\right)^{2m} \frac{2k+1}{2} \right)^{1/2} \left(\sum_{k=N+1}^{M} \frac{2}{2k+1} |g_k|^2 \right)^{1/2} \\
&\leqq \left(\int_N^{\infty} \left(\frac{1}{x(x+1)}\right)^{2m} \frac{2x+1}{2}\,dx \right)^{1/2} \|\mathscr{L}^m f\|_{L_2} \\
&= (4m-2)^{-1/2} \left(\frac{1}{N(N+1)}\right)^{m-1/2} \|\mathscr{L}^m f\|_{L_2}.
\end{aligned}$$

Here we have used $|p_k(x)| \leqq 1$ for $|x| \leqq 1$ and

$$\left(\frac{1}{k(k+1)}\right)^{2m} \frac{2k+1}{2} \leqq \int_{k-1}^{k} \left(\frac{1}{x(x+1)}\right)^{2m} \frac{2x+1}{2}\,dx.$$

It now follows that $\{P^N f\}_{N\geqq 0}$ is a Cauchy sequence in $C[-1, 1]$ and hence that $P^N f$ converges uniformly to f. Letting $M \to \infty$ above, we obtain

$$|f(x) - P^N f(x)| \leqq (4m-2)^{-1/2} \left(\frac{1}{N(N+1)}\right)^{m-1/2} \|\mathscr{L}^m f\|_{L_2}.$$

Since $\mathscr{L}^m$ is a differential operator of order $2m$ with continuous coefficients on $[-1, 1]$, there exists a constant C_m for each $m \geqq 0$ such that

$$\|\mathscr{L}^m f\|_{L_2} \leqq C_m \|f\|_{H^{2m}},$$

which completes the proof.

If f is C^∞, then from Lemma 3.1, the error $\|f - P^N f\|$ decreases more rapidly than any power of $1/N$. This is usually referred to as "infinite order" approximation.

4. Legendre–tau approximation. In this section, we discuss the Legendre-tau approximation of solutions to (2.1).

For simplicity of exposition, we assume that $r = 2$. See the remark at the end of the section for the general case. If $z(t, \theta) = x(t+\theta)$, $\theta \in [-2, 0]$, where x is the solution

to (2.1) with initial data $(\eta, \phi) \in \mathcal{D}(\mathcal{A})$, then, according to Lemma 2.1, z satisfies

$$\frac{\partial z}{\partial t}(t, \theta) = \frac{\partial z}{\partial \theta}(t, \theta), \qquad \theta \in [-2, 0], \tag{4.1}$$

$$\frac{dz}{dt}(t, 0) = \int_{-2}^{0} d\mu(\theta) z(t, \theta) + f(t). \tag{4.2}$$

The approximate solution $z^N(t, \theta)$ is assumed to be expanded in a Legendre series:

$$z^N(t, \theta) = \sum_{k=0}^{N} a_k^N(t) p_k(\theta + 1), \qquad a_k^N \in \mathbb{R}^n. \tag{4.3}$$

The tau approximation [9] of (4.1) and (4.2) is as follows. Note that from (3.4),

$$\frac{\partial z^N}{\partial \theta}(t, \theta) = \sum_{k=0}^{N-1} b_k^N(t) p_k(\theta),$$

where

$$b_k^N(t) \equiv (2k+1) \sum_{\substack{j=k+1 \\ j+k \text{ odd}}}^{N} a_j^N(t). \tag{4.4}$$

Equating $\partial/\partial t$ (z^N) with $\partial/\partial \theta$ (z^N) in the sense that

$$\left\langle \frac{\partial}{\partial t} z^N - \frac{\partial}{\partial \theta} Z^N, g \right\rangle_{L_2} = 0$$

for all polynomials g on $[-2, 0]$ of degree at most $N-1$ leads to the N equations

$$\frac{d}{dt} a_k^N(t) = b_k^N(t), \qquad 0 \leqq k \leqq N-1. \tag{4.5}$$

The essence of the tau method is that the boundary condition (4.2) is then imposed to determine an equation for a_N^N. From (4.2) we obtain

$$\frac{d}{dt}\left(\sum_{k=0}^{N} a_k^N(t) \right) = \int_{-2}^{0} d\mu(\theta) z^N(t, \theta) + f(t),$$

or

$$\frac{d}{dt} a_N^N(t) = -\sum_{k=0}^{N-1} b_k^N(t) + \int_{-2}^{0} d\mu(\theta) z^N(t, \theta) + f(t). \tag{4.6}$$

Hence, from (4.3)-(4.6), we obtain a system of ordinary differential equations for $a_0^N, \cdots, a_N^N$:

$$\frac{d}{dt} \alpha^N(t) = A^N \alpha^N(t) + B^N f(t), \tag{4.7}$$

where $\alpha^N = \operatorname{col}(a_0^N, \cdots, a_N^N)$ and

$$B^N = e_N \otimes I,$$

where $e_N \in \mathbb{R}^{N+1}$ is given by $e_N = \operatorname{col}(0, \cdots, 0, 1)$, I is the $n \times n$ identity matrix, and $\otimes$ denotes Kronecker product. For the case where N is even, A^N is given by

$$A^N = A_0^N + A_\mu^N,$$

where

$$A_0^N = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 3 & 0 & 3 & \cdots & 0 & 3 \\ 0 & 0 & 0 & 5 & 0 & \cdots & 5 & 0 \\ \vdots & \vdots & & & & & \vdots & \vdots \\ & & & & & & 2N-3 & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 2N-1 \\ 0 & -1 & -3 & -6 & -10 & \cdots & & -\dfrac{N(N+1)}{2} \end{bmatrix} \otimes I \tag{4.8}$$

and

$$A_\mu^N = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \\ D_0 & D_1 & \cdots & D_N \end{bmatrix} \tag{4.9}$$

with

$$D_k = \int_{-2}^{0} d\mu(\theta) p_k(\theta+1), \qquad 0 \leqq k \leqq N. \tag{4.10}$$

For N odd, only the last column of A_0^N is different.

For $N \geqq 1$, let

$$z^N(t) = (z^N(t,0), z^N(t,\cdot)) \in Z$$

where

$$z^N(t,\theta) = \sum_{k=0}^{N} a_k^N(t) p_k(\theta+1)$$

as in (4.3). Then the approximate solution $z^N(t)$ is the exact solution to the modified equation:

$$\frac{d}{dt} z^N(t) = \mathcal{A} z^N(t) + \mathcal{B} f(t) + \tau_N(t), \tag{4.11}$$

in Z, where

$$\tau_N(t) = \left(0, \frac{d}{dt} a_N^N(t) p_N\right) \in Z.$$

Indeed,

$$\frac{d}{dt} z^N(t) = \left(\frac{d}{dt} \sum_{k=0}^{N} a_k^N(t), \frac{d}{dt} \sum_{k=0}^{N} a_k^N(t) p_k\right)$$

and, using (4.5) and (4.6),

$$= \left(\frac{d}{dt} \sum_{k=0}^{N} a_k^N(t), \sum_{k=0}^{N-1} b_k^N(t) p_k\right) + \tau_N(t),$$

$$= \left(Dz^N(t,\cdot) + f(t), \sum_{k=0}^{N-1} b_k^N(t) p_k\right) + \tau_N(t).$$

But

$$\mathscr{A}z^N(t)+\mathscr{B}f(t)=\left(Dz^N(t,\cdot)+f(t),\ \sum_{k=0}^{N-1} b_k^N(t)p_k\right), \tag{4.12}$$

whence (4.11) follows. Note that

$$\begin{aligned}\frac{d}{dt}a_N^N(t)&=-\sum_{k=0}^{N-1} b_k^N(t)+Dz^N(t,\cdot)+f(t)\\&=-\dot{z}^N(t,0)+Dz^N(t,\cdot)+f(t).\end{aligned}$$

This fact suggests the introduction of the projection operator L^N on Z, defined as follows: for $z=(\eta,\phi)\in Z$.

$$L^N z=\left(\eta,\ \sum_{k=0}^{N} a_k^N p_k\right), \tag{4.13}$$

where, for $0\leqq k\leqq N-1$,

$$a_k^N=\frac{2k+1}{2}\int_{-2}^{0}\phi(\theta)p_k(\theta+1)\,d\theta,$$

and

$$a_N^N=\eta-\sum_{k=0}^{N-1} a_N^N p_k(1)=\eta-\sum_{k=0}^{N-1} a_k^N.$$

It is easy to show that L^N is a projection on Z (i.e., $L^NL^N=L^N$) and $L^NZ\subset\mathscr{D}(\mathscr{A})$. The tau method can now be interpreted as follows. The approximate solution $z^N(t)$ satisfies

$$\frac{d}{dt}z^N(t)=L^N\mathscr{A}z^N(t)+L^N\mathscr{B}f(t) \tag{4.14}$$

in Z. Indeed, applying L^N to both sides of (4.12) and using (4.6),

$$L^N\mathscr{A}z^N(t)+L^N\mathscr{B}f(t)=\mathscr{A}z^N(t)+\mathscr{B}f(t)+\tau_N(t).$$

Remark. For the general case, i.e., $r\neq 2$, the matrix A^N appearing in (4.7) is replaced by

$$A^N=\frac{2}{r}A_0^N+\tilde{A}_\mu^N,$$

where

$$\tilde{D}_k=\int_{-r}^{0} d\mu(\theta)p_k\left(\frac{2\theta+r}{r}\right),\qquad 0\leqq k\leqq N,$$

replaces D_k in (4.9). In particular, if the delay operator D is given by (2.3) and (2.4), then $\tilde{D}_k$ has the form

$$\tilde{D}_k=\sum_{i=0}^{m} A_i p_k\left(\frac{2\theta_i+r}{r}\right)+\int_{-r}^{0} A(\theta)p_k\left(\frac{2\theta+r}{r}\right)d\theta,$$

for $0\leqq k\leqq N$.

5. Convergence proof. Note that the projection operators L^N are neither orthogonal nor uniformly bounded and therefore do not converge strongly to the

identity operator in Z. This fact underlies an essential difficulty in trying to prove convergence in the space Z. To overcome this difficulty, we will begin with a convergence result in the space X which is isomorphic to $\mathcal{D}(\mathcal{A})$: our motivation here is the fact that (4.14) is also valid in $\mathcal{D}(\mathcal{A})$.

Without loss of generality, we can assume that $r=2$. For each $N\geqq 1$, let us define the orthogonal projection Π^N on Z by

$$\Pi^N(\eta,\phi)=(\eta,P^{N-1}\phi) \quad \text{for } (\eta,\phi)\in Z, \tag{5.1}$$

where P^N is the orthogonal projection on $L_2([-2,0];\mathbb{R}^n)$:

$$(P^N\phi)(\theta)=\sum_{k=0}^{N} a_k p_k(\theta+1) \tag{5.2}$$

with

$$a_k=\frac{2k+1}{2}\int_{-2}^{0}\phi(\theta)p_k(\theta+1)\,d\theta.$$

The orthogonality of Π^N follows from that of P^N. Since the projection operators L^N, $N\geqq 1$, given by (4.13), can be written as

$$L^N z=\Pi^N z+(0,(\eta-(P^{N-1}\phi)(0))p_N) \quad \text{for } z=(\eta,\phi)\in Z, \tag{5.3}$$

it follows immediately that

$$L^N\Pi^N=L^N \quad \text{and} \quad \Pi^N L^N=\Pi^N. \tag{5.4}$$

This fact, when combined with the convergence result for the space X, will lead to the result that the solution $z^N(t)$ to (4.14) with initial data $z^N(0)=L^N(\eta,\phi)$ converges to $z(t)$, the solution to (2.7) with initial data $z(0)=(\eta,\phi)$ in Z in the sense that

$$\|\Pi^N z^N(t)-z(t)\|_Z\to 0 \quad \text{as } N\to\infty. \tag{5.5}$$

For $N\geqq 1$, we define the finite-dimensional subspace $X^N\subset X$ by

$$X^N=\left\{\phi\in X:\ \phi(\theta)=\sum_{k=0}^{N} a_k p_k(\theta+1),\ a_k\in\mathbb{R}^n\right\},$$

and the orthogonal projection Q^N of X onto X^N by

$$(Q^N\phi)(\theta)=\phi(0)+\int_0^\theta (P^{N-1}\dot{\phi})(s)\,ds.$$

The orthogonality of Q^N with respect to the inner product $\langle\cdot,\cdot\rangle_1$, follows from that of P^N on $L_2(-2,0)$. If

$$Q^N\phi=\sum_{k=0}^{N} a_k p_k,$$

then

$$\frac{d}{d\theta}Q^N\phi=\sum_{k=0}^{N} a_k\dot{p}_k=P^{N-1}\dot{\phi},$$

so that, from (4.4), a_k, $1\leqq k\leqq N$, are uniquely determined. a_0 is then determined by

$$(Q^N\phi)(0)=\phi(0)=\sum_{k=0}^{N} a_k \Rightarrow a_0=\phi(0)-\sum_{k=1}^{N} a_k.$$

Let us define $\mathcal{A}^N : X \to X^N$ by

$$\mathcal{A}^N\phi = EL^N\mathcal{A}E^{-1}Q^N\phi \quad \text{for } \phi \in X, \tag{5.6}$$

where E is the isomorphism between X and $\mathcal{D}(\mathcal{A})$ defined in § 2. Then (4.14) can be written as

$$\begin{aligned} &\frac{d}{dt}z^N(t) = \mathcal{A}^N z^N(t) + \mathcal{B}^N f(t) \quad \text{in } X, \\ &z^N(0) = Q^N\phi, \end{aligned} \tag{5.7}$$

where, for $N \geqq 1$, $\mathcal{B}^N = EL^N\mathcal{B}$.

Our goal in this section is to prove the semigroup convergence of $S^N(t) = e^{\mathcal{A}^N t}$ to $\tilde{S}(t)$, defined by (2.10). The case where $f \neq 0$ and the initial data lies merely in Z will be dealt with in the following section. The principal result of this section is based upon the Trotter-Kato theorem (see [14, Thm. 4.6]).

THEOREM 5.1. *Let $S(t)$ and $S^N(t)$, $N \geqq 1$, be C_0-semigroups acting on a Banach space X, with infinitesimal generators $\mathcal{A}$ and $\mathcal{A}^N$ respectively. Assume that the following conditions are satisfied*:

(i) (*stability*) *There exists a constant ω such that*

$$\|S(t)\|_X \leqq e^{\omega t} \quad \text{and} \quad \|S^N(t)\|_X \leqq e^{\omega t}, \qquad t \geqq 0.$$

(ii) (*consistency*) *There exists a subset $\mathcal{D}$ contained in $\mathcal{D}(\mathcal{A}) \cap \bigcap_{N=1}^{\infty} \mathcal{D}(\mathcal{A}^N)$ which together with $(\lambda I - \mathcal{A})\mathcal{D}$ for some $\lambda > 0$ is dense in X and such that $\mathcal{A}^N\phi \to \phi$ for all $\phi \in \mathcal{D}$ as $N \to \infty$. Then for all $\phi \in X$, $\|S^N(t)\phi - S(t)\phi\|_X \to 0$ uniformly on bounded t-intervals.*

LEMMA 5.2 (stability). *There exists a constant $\omega \in \mathbb{R}$ such that, for all $N \geqq 1$, $\mathcal{A}^N - \omega I$ is dissipative on X, i.e.,*

$$\langle \mathcal{A}^N\phi, \phi\rangle_1 \leqq \omega\|\phi\|_1^2 \quad \text{for } \phi \in X.$$

Moreover, if $z^N(t)$ is the solution to (5.7) with $z^N(0) = Q^N\phi$ for $\phi \in X$, then

$$\|z^N(t)\|_1^2 \leqq e^{2\omega t}\|z^N(0)\|_1^2 + \int_0^t e^{2\omega(t-s)}|f(s)|^2\, ds. \tag{5.8}$$

Proof. For any $\phi^N \in X^N$ and $f \in \mathbb{R}^n$, it follows from (5.6) and (4.13) that

$$\langle \mathcal{A}^N\phi^N + \mathcal{B}^N f, \phi^N\rangle_1 = \langle D\phi^N + f, \phi^N(0)\rangle + \int_{-2}^{0} \langle \ddot{\phi}^N(\theta) + b_N\dot{p}_N(\theta), \dot{\phi}^N(\theta)\rangle\, d\theta \tag{5.9}$$

where

$$b_N = D\phi^N + f - \dot{\phi}^N(0). \tag{5.10}$$

Let I denote the integral term in (5.9). Then

$$\begin{aligned} I &= \int_{-2}^{0} \frac{1}{2}\frac{d}{d\theta}|\dot{\phi}^N(\theta)|^2\, d\theta + \int_{-2}^{0} \langle b_N, \dot{\phi}^N(\theta)\rangle\dot{p}_N(\theta+1)\, d\theta \\ &= \frac{1}{2}|\dot{\phi}^N(0)|^2 - \frac{1}{2}|\dot{\phi}^N(-2)|^2 + \langle b_N, \dot{\phi}^N(0) - (-1)^N\dot{\phi}^N(-2)\rangle \\ &\quad + \int_{-2}^{0} \langle \ddot{\phi}^N(\theta), b_N\rangle p_N(\theta+1)\, d\theta \end{aligned} \tag{5.11}$$

where we have used $p_N(\pm 1) = (\pm 1)^N$. Note that the last term on the right-hand side

vanishes because $\ddot{\phi}^N$ is a vector in $\mathbb{R}^N$ whose elements are polynomials of degree less than N and hence orthogonal to p_N. Thus, from (5.11),

$$\begin{aligned} I &= \tfrac{1}{2}|\dot{\phi}^N(0)|^2 - \tfrac{1}{2}|\dot{\phi}^N(-2)|^2 + \langle D\phi^N + f - \dot{\phi}^N(0), \dot{\phi}^N(0) - (-1)^N\dot{\phi}^N(-2)\rangle \\ &= -\tfrac{1}{2}|\dot{\phi}^N(0)|^2 + (-1)^N\langle \dot{\phi}^N(0), \dot{\phi}^N(-2)\rangle - \tfrac{1}{2}|\dot{\phi}^N(-2)|^2 \\ &\quad + \langle D\phi^N + f, \dot{\phi}^N(0) - (-1)^N\dot{\phi}^N(-2)\rangle \\ &\leqq \tfrac{1}{2}|D\phi^N + f|^2 \end{aligned}$$

where we have used the inequality $2\langle x, y\rangle_{\mathbb{R}^n} \leqq |x|^2 + |y|^2$. From this and (5.9), we find

$$\begin{aligned} \langle \mathscr{A}^N\phi^N + \mathscr{B}^N f, \phi^N\rangle_1 &\leqq \langle D\phi^N + f, \phi^N(0)\rangle + \tfrac{1}{2}|D\phi^N + f|^2 \\ &= \langle D\phi^N, \phi^N(0)\rangle + \langle f, D\phi^N + \phi^N(0)\rangle + \tfrac{1}{2}|f|^2 + \tfrac{1}{2}|D\phi^N|^2 \\ &= \langle D\phi^N, \phi^N(0)\rangle + \tfrac{1}{2}\{|f|^2 + 2\langle f, D\phi^N + \phi^N(0)\rangle + |D\phi^N + \phi^N(0)|^2\} \\ &\quad - \tfrac{1}{2}|D\phi^N + \phi^N(0)|^2 + \tfrac{1}{2}|D\phi^N|^2 \\ &= \tfrac{1}{2}|f + D\phi^N + \phi^N(0)|^2 - \tfrac{1}{2}|\phi^N(0)|^2 \\ &\leqq |D\phi^N + \phi^N(0)|^2 + |f|^2. \end{aligned}$$

Note that $D \in \mathscr{L}(X, \mathbb{R}^n)$ with

$$|D\phi| \leqq \beta\|\phi\|_1 \tag{5.12}$$

for some positive constant $\beta < \infty$. Thus

$$\langle \mathscr{A}^N\phi^N + \mathscr{B}^N f, \phi^N\rangle_1 \leqq \omega\|\phi^N\|_1^2 + |f|^2 \tag{5.13}$$

with $\omega = (1+\beta)^2$.

Since Q^N is symmetric with respect to $\langle\cdot,\cdot\rangle_1$, we have, for all $\phi \in X$,

$$\langle \mathscr{A}^N\phi, \phi\rangle_1 = \langle Q^N\mathscr{A}^NQ^N\phi, \phi\rangle_1 = \langle \mathscr{A}^NQ^N\phi, Q^N\phi\rangle_1.$$

It then follows from (5.13) that

$$\langle \mathscr{A}^N\phi, \phi\rangle_1 \leqq \omega\|Q^N\phi\|_1^2 \leqq \|\phi\|_1^2$$

which implies that $\mathscr{A}^N - \omega I$ is dissipative on X.

From (5.7) and (5.13),

$$\frac{1}{2}\frac{d}{dt}\|z^N(t)\|_1^2 = \langle \mathscr{A}^Nz^N(t) + \mathscr{B}^Nf(t), z^N(t)\rangle_1 \leqq \omega\|z^N(t)\|_1^2 + |f(t)|^2.$$

A standard argument using Gronwall's inequality then yields (5.8). Q.E.D.

If $\tilde{\mathscr{A}}$ is defined by (2.10), (i.e., the infinitesimal generator of the semigroup $\tilde{S}(t)$, on X), then Lemma 5.3 follows.

LEMMA 5.3 (consistency). *Let*

$$\mathscr{D}^k = \{\phi \in X: \phi \in H^k \text{ and } \dot{\phi}(0) = D\phi\}.$$

Then for $k \geqq 2$, $\mathscr{D}^k$ and $(\lambda I - \tilde{\mathscr{A}})\mathscr{D}^k$ for $\lambda \in \mathbb{R}$ sufficiently large are dense in X and, for $k \geqq 5$, $\mathscr{A}^N\phi \to \tilde{\mathscr{A}}\phi$ in X for $\phi \in \mathscr{D}^k$.

Proof. Since $\mathscr{D}^k$ densely contained in $\mathscr{D}(\tilde{\mathscr{A}})$ for $k \geqq 2$, the denseness of $\mathscr{D}^k$ follows from that of $\mathscr{D}(\tilde{\mathscr{A}})$. Since $\mathscr{D}^2 = \mathscr{D}(\tilde{\mathscr{A}})$ and, for $\lambda \in \mathbb{R}$ sufficiently large

$$(\lambda I - \tilde{\mathscr{A}})\mathscr{D}(\tilde{\mathscr{A}}) = X,$$

$(\lambda I - \tilde{\mathscr{A}})\mathscr{D}^k$ is dense in X for $k \geqq 2$.

For $\phi \in \mathcal{D}(\tilde{\mathcal{A}})$ and $Q^N\phi \equiv \phi^N$,

$$(\mathcal{A}^N - \tilde{\mathcal{A}})\phi = \dot{\phi}^N - \dot{\phi} + (D\phi^N - \dot{\phi}^N(0))p_N \tag{5.14}$$

in X. From (5.10), the definition of Q^N,

$$\dot{\phi}^N = p^{N-1}\dot{\phi} \quad \text{and} \quad \phi^N(0) = \phi(0).$$

Thus, from Lemma 3.3,

$$|\dot{\phi}^N(0) - \dot{\phi}(0)| \leqq KN^{-2m+1}\|\dot{\phi}\|_{H^{2m}}.$$

From Lemma 3.2,

$$\|\dot{\phi} - \dot{\phi}^N\|_{H^1} = \|\dot{\phi} - p^{N-1}\dot{\phi}\|_{H^1} \leqq KN^{-2m+3/2}\|\dot{\phi}\|_{H^{2m}}.$$

From (5.12),

$$|D(\phi^N - \phi)|_{\mathbb{R}^n} \leqq \beta\|\phi - \phi^N\|_1 = \beta\|\dot{\phi} - \dot{\phi}^N\|_{L_2} = \beta\|\dot{\phi} - p^{N-1}\dot{\phi}\|_{L_2};$$

so that, by Lemma 3.1,

$$\leqq \beta KN^{-2m}\|\dot{\phi}\|_{H^{2m}}.$$

Since $\dot{\phi}(0) = D\phi$ for $\phi \in \mathcal{D}(\tilde{\mathcal{A}})$,

$$\begin{aligned}|D\phi^N - \dot{\phi}^N(0)| &\leqq |\dot{\phi}^N(0) - \dot{\phi}(0)| + |D(\phi^N - \phi)| \\ &\leqq (1+\beta)KN^{-2m+1}\|\dot{\phi}\|_{H^{2m}}.\end{aligned}$$

Note that

$$\int_{-1}^{1} |\dot{p}_N(\theta)|^2 = p_N(\theta)\dot{p}_N(\theta)|_{-1}^{1} - \int_{-1}^{1} p_N(\theta)\ddot{p}_N(\theta)\, d\theta$$

where the second term vanishes as before. Hence

$$\int_{-1}^{1} |\dot{p}_N(\theta)|^2\, d\theta = N(N+1).$$

It now follows from these estimates and (5.14) that

$$\begin{aligned}\|(\mathcal{A}^N - \tilde{\mathcal{A}})\phi\|_1 &\leqq \|\dot{\phi}^N - \dot{\phi}\|_1 + |D\phi^N - \dot{\phi}^N(0)|(1+\sqrt{N(N+1)}) \\ &\leqq (2+\beta)KN^{-2m+2}\|\dot{\phi}\|_{H^{2m}}.\end{aligned}$$

Therefore, if $\phi \in \mathcal{D}^k$, $k \geqq 5$,

$$\|(\mathcal{A}^N - \tilde{\mathcal{A}})\phi\|_1 \leqq (2+\beta)KN^{-2}\|\phi\|_{H^5} \to 0,$$

since $|\dot{\phi}\|_{H^4} \leqq \|\phi\|_{H^5}$. Q.E.D.

Remark. Using the same "completing the square" argument as the one employed in the proof of Lemma 5.2, it is easily verified that $\tilde{\mathcal{A}} - \omega I$ is dissipative on X as well.

Combining Theorem 5.1 with Lemmas 5.2 and 5.3, we have the following theorem.

THEOREM 5.4. *If $\{S^N(t): t \geqq 0\}$ denotes the semigroup on X generated by $\mathcal{A}^N$, $N \geqq 1$, then, for all $\phi \in X$,*

$$\|S^N(t)\phi - \tilde{S}(t)\phi\|_1 \to 0$$

uniformly on bounded t-intervals.

6. Convergence proof (continued). In this section, we prove the convergence of our scheme for the cases where $f \neq 0$ and where the initial data (η, ϕ) is merely assumed to lie in Z.

Let us first consider the case where $f \in L_2^{\text{loc}}([0,\infty);\mathbb{R}^n)$ and $(\eta,\phi)\equiv 0$. The solution (2.1) is given by

$$z(t)=\int_0^t S(t-s)\mathscr{B}f(s)\,ds. \tag{6.1}$$

Recall that for any $f\in\mathbb{R}^n$, $\mathscr{B}f=(f,0)\in Z$. Using formula (2.6) for the resolvent of $\mathscr{A}$,

$$\mathscr{A}^{-1}\mathscr{B}f=(\Delta^{-1}f,\Delta^{-1}p_0)\in Z \tag{6.2}$$

provided that $\Delta=\Delta(0)=\int_{-r}^0 d\mu(\theta)$ is invertible.

When Δ is not invertible (i.e., $0\in\sigma(\mathscr{A})$), we can choose $\omega\in\rho(\mathscr{A})$ and consider $y(t)=e^{-\omega t}x(t)$, $x(t)$ being the solution to the initial value problem (2.1). Then $y(t)$ satisfies

$$\frac{d}{dt}y(t)=-\omega y(t)+\int_{-r}^0 e^{\omega\theta}\,d\mu(\theta)y(t+\theta)+e^{-\omega t}f(t)\equiv D_\omega y_t+e^{-\omega t}f(t)$$

with initial data $y(\theta)=e^{-\omega\theta}\phi(\theta)$, $-r\leqq\theta\leqq 0$. Clearly, this problem may be formulated as before on the product space Z; the corresponding generator $\mathscr{A}_\omega$ will, by construction, satisfy $0\in\rho(\mathscr{A}_\omega)$:

$$\Delta_\omega(\lambda)=(\lambda+\omega)I-\int_{-r}^0 e^{(\lambda+\omega)\theta}\,d\mu(\theta),$$

$$\det[\Delta_\omega(0)]=\det[\Delta(\omega)]\neq 0.$$

Thus, without loss of generality, we may consider (6.2) above.

Let us rewrite (6.1) as

$$z(t)=\int_0^t \mathscr{A}S(t-s)\mathscr{C}f(s)\,ds$$

where $\mathscr{C}f\equiv\mathscr{A}^{-1}\mathscr{B}f$. If f is continuously differentiable, it follows from [10, p. 487] that

$$z(t)=S(t)\mathscr{C}f(0)-\mathscr{C}f(t)+\int_0^t S(t-s)\mathscr{C}\dot{f}(s)\,d.$$

Since $\mathscr{C}f\in\mathscr{D}(\mathscr{A})$, $z(t)\in\mathscr{D}(\mathscr{A})$ for all $t\geqq 0$ and we can write

$$\begin{aligned} Ez(t)&=ES(t)E^{-1}E\mathscr{C}f(0)-E\mathscr{C}f(t)+\int_0^t ES(t)E^{-1}E\mathscr{C}\dot{f}(s)\,ds \\ &=\tilde{S}(t)E\mathscr{C}f(0)-E\mathscr{C}(t)+\int_0^t \tilde{S}(t-s)E\mathscr{C}\dot{f}(s)\,ds. \end{aligned} \tag{6.3}$$

LEMMA 6.1. *For $z=(\eta,\phi)\in Z$,*

$$(\mathscr{A}^N)^{-1}EL^Nz=E\mathscr{A}^{-1}\Pi^Nz.$$

Proof. Recall that the isomorphism $E:\mathscr{D}(\mathscr{A})\to X$ and that $\mathscr{A}^N=EL^N\mathscr{A}E^{-1}Q^N$: $X\to X^N\equiv Q^NX$; thus, there will in general be many solutions $w\in\mathscr{D}(\mathscr{A})$ to

$$\mathscr{A}^NEw=EL^Nz. \tag{6.4}$$

We shall interpret $(\mathscr{A}^N)^{-1}EL^Nz$ to be the unique solution of (6.4) lying in X^N. We must then show that a unique solution exists and has the indicated form.

For $z\in Z$, let

$$\Pi^Nz=\left(\eta,\sum_{k=0}^{N-1}b_kp_k\right)\equiv(\eta,\phi).$$

From (2.6), $\mathscr{A}^{-1}(\eta, \phi) = (\psi(0), \psi)$, where

$$\psi(\theta) = \Delta^{-1}\left(\eta + D\int_{\cdot}^{0} \phi(s)\,ds\right) - \int_{\theta}^{0} \phi(s)\,ds,$$

and thus, since ϕ is a polynomial of degree $N-1$, ψ has degree N. This shows that

$$E\mathscr{A}^{-1}\Pi^N z \in X^N.$$

Since $Q^N : X \to X^N$ is a projection,

$$\begin{aligned}\mathscr{A}^N E\mathscr{A}^{-1}\Pi^N z &= EL^N \mathscr{A} E^{-1} Q^N E\mathscr{A}^{-1}\Pi^N z \\ &= EL^N \Pi^N z \\ &= EL^N z \quad \text{by (5.4).}\end{aligned}$$

Thus, $Ew = E\mathscr{A}^{-1}\Pi^N z \in X^N$ is a solution of (6.4).

To establish uniqueness, we must show that

$$\mathscr{A}^N Ev = 0 \text{ and } Ev \in X^N \Rightarrow v = 0.$$

Let $Ev = \sum_{k=0}^{N} a_k p_k \equiv \phi$ so that

$$v = E^{-1}\phi = (\phi(0), \phi) \in \mathscr{D}(\mathscr{A}).$$

Thus, $\mathscr{A}v = (D\phi, \dot{\phi})$, and so

$$\begin{aligned}\mathscr{A}^N Ev = EL^N \mathscr{A} E^{-1} Q^N Ev = EL^N \mathscr{A} v &= E(D\phi, \dot{\phi} + (D\phi - \dot{\phi}(0))p_N) \\ &= \dot{\phi} + (D\phi - \dot{\phi}(0))p_N.\end{aligned}$$

Since $\dot{\phi}$ is a polynomial of degree $N-1$, $\mathscr{A}^N Ev = 0$ and the orthogonality of the Legendre polynomials yields the fact that $\phi \equiv \phi_0 = \text{constant}$ and $D\phi = 0$. But

$$D\phi = D(\phi_0 p_0) = \Delta\phi_0$$

whence (cf. (6.2)) $\phi_0 = 0$, or $v = 0$. Q.E.D.

From Lemma 6.1,

$$(\mathscr{A}^N)^{-1} EL^N \mathscr{B} f = E\mathscr{A}^{-1}\Pi^N \mathscr{B} f = E\mathscr{A}^{-1}\mathscr{B} f = E\mathscr{C} f \quad \text{for } f \in \mathbb{R}^n.$$

From the definition of $\mathscr{A}^N$ and $\mathscr{B}^N$, the same argument as above applied to

$$z^N(t) = \int_0^t \mathscr{A}^N S^N(t-s)(\mathscr{A}^N)^{-1}\mathscr{B}^N f(s)\,ds$$

yields

$$z^N(t) = S^N(t)E\mathscr{C} f(0) - E\mathscr{C} f(t) + \int_0^t S^N(t-s)E\mathscr{C}\dot{f}(s)\,ds. \tag{6.5}$$

It then follows from Theorem 5.4 that

$$S^N(t)E\mathscr{C} f \to \tilde{S}(t)E\mathscr{C} f \quad \text{in } X \text{ for } f \in \mathbb{R}^n.$$

Since $\dim(\mathbb{R}^n) < \infty$, this implies

$$\|(S^N(t) - \tilde{S}(t))E\mathscr{C}\|_{\mathscr{L}(\mathbb{R}^n, X)} \to 0.$$

Hence, from (6.3) and (6.5), if f is continuously differentiable,

$$\|z^N(t) - Ez(t)\|_1 \to 0, \tag{6.6}$$

uniformly on bounded t-intervals.

According to the stability result of Lemma 5.2 (c.f. (5.8)),

$$\left\| \int_0^t S^N(t-s)\mathcal{B}^N f(s)\, ds \right\|_1 \leqq e^{\omega t} \|f\|_{L_2([0,t];\mathbb{R}^n)}. \tag{6.7}$$

Using the same argument as that given in the proof of Lemma 5.2, the above estimate holds true for (2.7); i.e.,

$$\left\| E \int_0^t S(t-s)\mathcal{B}f(s)\, ds \right\|_1 \leqq e^{\omega t} \|f\|_{L_2([0,t];\mathbb{R}^n)}. \tag{6.8}$$

Since the space of continuously differentiable functions on $[0, t]$ is dense in $L_2(0, t)$ for all $t > 0$, it follows from the Banach-Steinhaus theorem that (6.6) holds true for any $f \in L_2^{\text{loc}}$ as well. Thus, combining the above with Theorem 5.4, we have the following theorem.

THEOREM 6.2. *For any initial data $(\phi(0), \phi) \in Z$ with $\phi \in H^1$ and $f \in L_2^{\text{loc}}$, the approximate solution $z^N(t)$ to (5.7) converges strongly to $Ez(t)$ in X, uniformly on bounded t-intervals.*

Remark. The theorem yields much stronger convergence results with regard to the nature of the inhomogeneous term than those stated in [2] and [3] where convergence results in Z are obtained by other methods.

Note that Theorem 6.2 is not applicable when the initial data (η, ϕ) is merely assumed to lie in Z. To complete our study of convergence, we thus need to consider the case $(\eta, \phi) \in Z$ and $f \equiv 0$.

THEOREM 6.3. *For any $z = (\eta, \phi) \in Z$,*

$$\|\Pi^N(E^{-1}S^N(t)EL^N z - S(t)z)\|_Z \to 0$$

uniformly on bounded t-intervals.

Proof. Let

$$d = d(t) = \Pi^N(E^{-1}S^N(t)EL^N z - S(t)z),$$

thus

$$d = \Pi^N E^{-1}\mathcal{A}^N S^N(t)(\mathcal{A}^N)^{-1}EL^N z - \Pi^N S(t)z.$$

It follows from Lemma 6.1 that

$$\begin{aligned} d &= \Pi^N E^{-1}\mathcal{A}^N S^N(t)E\mathcal{A}^{-1}\Pi^N z - \Pi^N S(t)z \\ &= \Pi^N(E^{-1}\mathcal{A}^N - \mathcal{A}E^{-1})S^N(t)E\mathcal{A}^{-1}\Pi^N z \\ &\quad + \Pi^N \mathcal{A}E^{-1}(S^N(t) - \tilde{S}(t))E\mathcal{A}^{-1}z + \Pi^N \mathcal{A}E^{-1}S^N(t)E\mathcal{A}^{-1}(\Pi^N z - z) \\ &\equiv d_1 + d_2 + d_3. \end{aligned}$$

Since $S^N(t)E\mathcal{A}^{-1}\Pi^N z \in X^N$,

$$\begin{aligned} d_1 &= \Pi^N(E^{-1}EL^N\mathcal{A}E^{-1}Q^N - \mathcal{A}E^{-1})S^N(t)E\mathcal{A}^{-1}\Pi^N z \\ &= \Pi^N(L^N - I)\mathcal{A}E^{-1}S^N(t)E\mathcal{A}^{-1}\Pi^N z = 0 \end{aligned}$$

as $\Pi^N L^N y = \Pi^N y$ for any $y \in Z$. Since Π^N is orthogonal on Z,

$$\begin{aligned} \|d_2\|_Z &\leqq \|\mathcal{A}E^{-1}(S^N(t) - \tilde{S}(t))E\mathcal{A}^{-1}z\|_Z \leqq \|E^{-1}(S^N(t) - \tilde{S}(t))E\mathcal{A}^{-1}z\|_{\mathcal{D}(\mathcal{A})} \\ &\leqq C\|(S^N(t) - \tilde{S}(t))E\mathcal{A}^{-1}z\|_1 \end{aligned}$$

by (2.8). Thus, by Theorem 5.4, $\|d_2(t)\|_Z \to 0$ uniformly on bounded t-intervals.

Likewise,

$$\|d_3\|_Z \leqq C\|S^N(t)E\mathscr{A}^{-1}(\Pi^N z - z)\|_1 \leqq C \cdot \|S^N(t)\|_{\mathscr{L}(X)} \cdot \|E\mathscr{A}^{-1}(\Pi^N z - z)\|_1$$
$$\leqq C e^{\omega t} \cdot c^{-1}\|\mathscr{A}^{-1}(\Pi^N z - z)\|_{\mathscr{D}(\mathscr{A})}$$

by Lemma 5.2 and the first inequality in (2.8). Thus

$$\|d_3\|_Z \leqq \left(\frac{C}{c}\right) e^{\omega t}\{\|\mathscr{A}^{-1}(\Pi^N z - z)\|_Z + \|\Pi^N z - z\|_Z\}$$
$$\leqq \left(\frac{C}{c}\right) e^{\omega t} \cdot (\|\mathscr{A}^{-1}\|_{\mathscr{L}(Z)} + 1)\|\Pi^N z - z\|_Z,$$

and thus $\|d_3(t)\|_Z \to 0$ uniformly on bounded t-intervals as well. This completes the proof of Theorem 6.3.

Remark. Theorem 6.3 would not hold without Π^N as pointed out in § 5: the sequence of projections $\{L^N\}_{N\geqq 1}$ does not converge strongly to the identity operator in Z. However, the theorem implies that the $\mathbb{R}^n$-component of $z^N(t)$ (i.e., $z^N(t, 0)$) does in fact converge to $x(t)$, the solution of (2.1), uniformly on bounded t-intervals, and this is all that is needed as far as numerical convergence of our scheme is concerned. Moreover, it follows from (5.4) that $\{\Pi^N E^{-1} S^N(t) E L^N, t \geqq 0\}$ forms a strongly continuous semigroup on $Z^N \equiv \Pi^N Z$ and its generator is given by

$$\Pi^N E^{-1}\mathscr{A}^N E L^N = \Pi^N E^{-1} E L^N \mathscr{A} E^{-1} Q^N E L^N = \Pi^N \mathscr{A} L^N$$

by (5.6).

7. Numerical results and conclusions. In this section we discuss some numerical examples which demonstrate the feasibility of the Legendre-tau approximation. For the purpose of comparison, we have also computed approximate solutions by using the cubic spline approximation (S_3) which is discussed in [2]. All computations were performed on a Control Data Corporation Cyber 170 model 730 at NASA Langley Research Center (LaRC) using software written in Fortran. The integration of the system of ordinary differential equations (ODEs) (4.6) was carried out by an IMSL routine (DVERK) employing the Runge-Kutta-Verner fifth and sixth order method.

For the Legendre-tau approximation, implementation of the algorithms is almost as easy for the averaging approximation (AV) which is discussed in [3] and the first-order spline approximation (S_1). In Table 1, we give the number of operations for each approximation to compute the right-hand side of (4.6) for the scalar system:

$$\frac{d}{dt}x(t) = ax(t) + bx(t-r) + f(t). \tag{7.1}$$

SP stands for the Legendre-tau approximation. For S_3, C_1 and C_2 are some positive constants (independent of N). Note that, for S_1 and S_3, the operations which are required in order to perform the Cholesky decomposition of Q_k^N (for the definition,

TABLE 1

Method	Number of additions	Number of multiplications
SP	$2N+4$	$N+2$
AV	$N+2$	$N+2$
S_1	$3N+4$	$3N+5$
S_3	$11N+C_1$	$9N+C_2$

see [2, p. 511]) are not included. However these numbers may not reflect directly the CPU time required to solve the approximating system of ODEs. According to our calculations, for the same value of N, the Legendre-tau approximation is about four times as fast as S_3. In addition, no storage space is necessary for the matrix A^N appearing in (4.6) for system (7.1) because of its simple structure. For the general system (2.6) in $\mathbb{R}^n$, the last n rows of A^N need to be stored in order to integrate the approximating system of ODEs using the DVERK routine.

As will be evident from the numerical results for the examples presented below, both approximation methods (SP and S_3) behave about the same initially. The typical feature of the Legendre-tau method is that the relative error decreases with increasing time. This is not unexpected due to two facts: (i) the regularity of the solution to a problem which has smooth inhomogeneous terms increases in time (as pointed out in the Introduction) and (ii) Lemma 3.1. For all examples, on the interval where the solution is infinitely differentiable, the rate of convergence seems to be infinite order. In contrast, spline approximation methods and AV yields finite-order rates of convergence, which is observed in our calculations for S_3.

In order to approximate the initial data, or, in the case when the system involves distributed delays, the expansion coefficients of certain functions need to be computed. For AV and spline approximations such computations are relatively easy, since each element has local support. In our calculations for S_3 we used a Gauss quadrature rule [7]. In contrast, the Legendre polynomials are supported on the whole interval $[-2, 0]$, and, for k large, $p_k(x)$ is a rapidly oscillating function. A feasible algorithm for computing Legendre coefficients will be discussed in a forthcoming paper. However for AV and spline approximations, the expansion coefficients must be recomputed if N is changed. For the Legendre-tau approximation we can use the values which have already been computed for smaller N.

In the tables below we will use the following notation. δ_{SP}^N is one of the differences

$$\left| \sum_{k=0}^{N} (a_k^N)_j(t) - x_j(t) \right|, \qquad j = 1, 2, \cdots, n,$$

where $a_k^N(t)$ is the kth segment of dimension n in the solution vector α^N of the approximating system of ODEs (4.6), and $x(t)$ is the true solution. Similarly, $\delta_{S_3}^N$ denotes one of the differences

$$|(\beta^N(0) w_j^N)(t) - x_j(t)|, \qquad j = 1, 2, \cdots, n.$$

Example 1 (Banks-Kappel [2, Example 1]). In this example we study the equation for a damped oscillator with delayed restoring force and constant external force,

$$\frac{d^2}{dt^2} x(t) + \frac{d}{dt} x(t) + x(t-1) = 10,$$

with initial conditions

$$x(\theta) = \cos\theta, \quad \frac{d}{dt} x(\theta) = -\sin\theta \quad \text{for } \theta \in [-1, 0].$$

Rewriting the above equation as a first order system we have

$$\frac{d}{dt}\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & -1 \end{bmatrix}\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} x_1(t-1) \\ x_2(t-1) \end{bmatrix} + \begin{bmatrix} 0 \\ 10 \end{bmatrix},$$

where $x_1(t) = x(t)$ and $x_2(t) = dx(t)/dt$.

Table 2 and Table 3 show the numerical results for $x(t)$ and $dx(t)/dt$. For this example SP achieves the same accuracy with smaller N that S_3 does. For SP, it appears that the rate of convergence is infinite-order.

TABLE 2

t	$x(t)$	$\delta^4_{S_3}$	$\delta^8_{S_3}$	$\delta^{16}_{S_3}$	δ^4_{SP}	δ^8_{SP}
.25	1.2704759	0.00007	0.00022	0.000024	0.00011	0.0000047
.5	1.9936737	0.00171	0.00021	0.000022	0.00028	0.0000012
.75	3.0614837	0.00247	0.00019	0.000017	0.00011	0.0000176
1.0	4.3927203	0.00083	0.00024	0.000012	0.00027	0.0000106
1.25	5.9259310	0.00082	0.00032	0.000016	0.00020	0.0000016
1.5	7.6000709	0.00090	0.00005	0.000010	0.00007	0.0000009
1.75	9.3440157	0.00054	0.00025	0.000005	0.00006	0.0000002
2.0	11.0833011	0.00013	0.00003	0.000015	0.00003	0.

TABLE 3

t	$\dot{x}(t)$	$\delta^4_{S_3}$	$\delta^8_{S_3}$	$\delta^{16}_{S_3}$	δ^4_{SP}	δ^8_{SP}
.25	2.06969	0.00946	0.00205	0.00048	0.00001	0.00001
.5	3.64428	0.00792	0.00021	0.00186	0.00171	0.00025
.75	4.84445	0.00073	0.00200	0.00045	0.00335	0.00014
1.0	5.76581	0.00790	0.00192	0.00046	0.00348	0.00031
1.25	6.45956	0.00625	0.00119	0.00039	0.00029	0.00008
1.5	6.45956	0.00185	0.00245	0.00062	0.00078	0.00002
1.75	7.0159957	0.00061	0.00003	0.00067	0.00075	0.
2.0	6.8497211	0.00180	0.00200	0.00059	0.00019	0.

Example 2 ([2, Example 4]). Next we use an example due to Popov for a degenerate system where we have $(1, -2, -1)^T x(t) = 0$ for $t \geqq 2$ and all iniital data $(\eta, \phi) \in Z$. The equation is

$$\frac{d}{dt}x(t) = \begin{bmatrix} 0 & 2 & 0 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{bmatrix} x(t) + \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 2 & 0 \end{bmatrix} x(t-1).$$

We choose the initial data:

$$\eta = \operatorname{col}(1, 1, 1) \quad \text{and} \quad \phi(\theta) \equiv 0 \in \mathbb{R}^3 \quad \text{for } -1 \leqq \theta \leqq 0.$$

Note that $(\eta, \phi) \notin \mathcal{D}(\mathcal{A})$.

In Tables 4, 5, and 6 we give the numerical results. Here the initial function is not in the subspace Z^N for either method. With respect to $x_2(t)$ and $x_3(t)$ we can see that S_3 gives slightly better approximations than SP for $t \leqq 1$. This is because x_2 and x_3 have a jump discontinuity in the derivative at $t = 1$. For SP, the convergence of the approximate solutions to the asymptotic solution $(2, 0, 2)$ is quite rapid and seems to be infinite-order.

Example 3 ([2, Example 5]). We consider the scalar equation

$$\frac{d}{dt}x(t) = 5x(t) + x(t-1),$$

with initial function

$$x(\theta) = 5 \quad \text{for } \theta \in [-1, 0].$$

TABLE 4

t	$x_1(t)$	$\delta^4_{S_3}$	$\delta^8_{S_3}$	$\delta^{16}_{S_3}$	δ^4_{SP}	δ^8_{SP}	δ^{16}_{SP}
0.2	1.36	0.01211	0.00493	0.00247	0.00140	0.00041	0.00001
0.4	1.64	0.00889	0.00528	0.00001	0.00146	0.00158	0.00001
0.6	1.84	0.00445	0.00334	0.00190	0.00547	0.00098	0.00016
0.8	1.96	0.00980	0.00074	0.00146	0.00233	0.00111	0.00019
1.0	2.0	0.00020	0.00221	0.00054	0.00339	0.00033	0.00009
1.2	2.0	0.00725	0.00424	0.00175	0.00237	0.00002	0.00001
1.4	2.0	0.00160	0.00460	0.00085	0.00064	0.00003	3×10^{-7}
1.6	2.0	0.00527	0.00406	0.00130	0.00036	0.00003	13×10^{-7}
1.8	2.0	0.00331	0.00281	0.00242	0.00028	0.00001	5×10^{-7}
2.0	2.0	0.00256	0.00117	0.00130	0.00004	5×10^{-6}	1×10^{-7}
2.2	2.0	0.00306	0.00049	0.00096	0.00004	1×10^{-6}	9×10^{-9}
2.4	2.0	0.00531	0.00178	0.00214	0.00003	2×10^{-8}	2×10^{-9}
2.6	2.0	0.00212	0.00246	0.00115	5×10^{-6}	9×10^{-8}	1×10^{-9}
2.8	2.0	0.00055	0.00251	0.00092	3×10^{-6}	5×10^{-8}	3×10^{-9}
3.0	2.0	0.00117	0.00200	0.00192	2×10^{-6}	2×10^{-8}	0.

TABLE 5

t	$x_2(t)$	$\delta^4_{S_3}$	$\delta^8_{S_3}$	$\delta^{16}_{S_3}$	δ^4_{SP}	δ^8_{SP}	δ^{16}_{SP}
0.2	0.8	0.00947	0.00536	0.00230	0.02139	0.00517	0.00295
0.4	0.6	0.01206	0.00451	0.00034	0.02201	0.00650	0.00302
0.6	0.4	0.00526	0.00260	0.00202	0.00021	0.00126	0.00119
0.8	0.2	0.00153	0.00015	0.00139	0.02790	0.00240	0.00039
1.0	0.	0.02133	0.01895	0.00949	0.04081	0.02072	0.01034
1.2	0.	0.01505	0.00168	0.00010	0.00135	0.00265	0.00053
1.4	0.	0.01369	0.00068	0.00094	0.00440	0.00085	0.00013
1.6	0.	0.00105	0.00078	0.00209	0.00071	0.00027	0.00003
1.8	0.	0.01149	0.00105	0.00193	0.00067	0.00006	3×10^{-6}
2.0	0.	0.00444	0.00123	0.00020	0.00042	4×10^{-6}	4×10^{-7}
2.2	0.	0.00636	0.00108	0.00173	0.00004	5×10^{-6}	3×10^{-7}
2.4	0.	0.00521	0.00066	0.00202	0.00007	3×10^{-6}	8×10^{-8}
2.6	0.	0.00205	0.00019	0.00073	0.00003	1×10^{-6}	1×10^{-8}
2.8	0.	0.00407	0.00021	0.00084	0.00005	3×10^{-7}	3×10^{-10}
3.0	0.	0.00039	0.00051	0.00119	0.00005	3×10^{-7}	6×10^{-10}

TABLE 6

t	$x_3(t)$	$\delta^4_{S_3}$	$\delta^8_{S_3}$	$\delta^{16}_{S_3}$	δ^4_{SP}	δ^8_{SP}	δ^{16}_{SP}
0.2	1.0	0.01131	0.00517	0.00235	0.04191	0.01044	0.00592
0.4	1.0	0.01278	0.00471	0.00022	0.04592	0.01267	0.00600
0.6	1.0	0.00628	0.00373	0.00202	0.00799	0.00181	0.00224
0.8	1.0	0.00472	0.00100	0.00164	0.04989	0.00105	0.00105
1.0	1.0	0.04871	0.03399	0.01807	0.08354	0.04269	0.02070
1.2	1.36	0.02598	0.00193	0.00149	0.00955	0.00603	0.00091
1.4	1.64	0.02266	0.00337	0.00228	0.00595	0.00250	0.00014
1.6	1.84	0.00015	0.00167	0.00248	0.00443	0.00143	0.00007
1.8	1.96	0.01932	0.00086	0.00153	0.00314	0.00117	0.00016
2.0	2.0	0.01007	0.00266	0.00049	0.00295	0.00057	0.00008
2.2	2.0	0.00901	0.00349	0.00223	0.00328	0.00002	0.00001
2.4	2.0	0.00127	0.00375	0.00198	0.00098	0.00007	1×10^{-6}
2.6	2.0	0.00071	0.00321	0.00069	0.00052	0.00005	3×10^{-6}
2.8	2.0	0.00714	0.00211	0.00049	0.00050	0.00002	9×10^{-7}
3.0	2.0	0.00465	0.00076	0.00060	0.00011	0.00001	1×10^{-7}

The numerical results for this example can be found in Tables 7 and 8. Again, we observe the quickness of the convergence of the approximate SP solution in this example.

Example 4. Here we deal with the equation which has multiple point delays

$$\frac{d}{dt}x(t) = x(t) + 2x(t-\tfrac{1}{2}) + x(t-1)$$

with initial function

$$x(\theta) = 1 \quad \text{for } \theta \in [-1, 0].$$

The numerical results in Table 9 show that both methods work equally well.

TABLE 7

t	$x(t)$	$\delta_{S_3}^4$	$\delta_{S_3}^8$	$\delta_{S_3}^{16}$	$\delta_{S_3}^{32}$	$\delta_{S_3}^{64}$
0.2	15.309691	0.03	0.00486	0.00028	0.000621	0.000066
0.4	43.334337	0.05	0.00403	0.00229	0.000243	0.000077
0.6	119.513222	0.26	0.01267	0.00080	0.000056	0.000107
0.8	326.588900	0.78	0.01900	0.00136	0.000357	0.000079
1.0	889.478955	2.39	0.09568	0.00206	0.000012	0.000123
1.2	2420.772761	7.28	0.25894	0.00031	0.000845	0.000149
1.4	6588.865818	22.02	0.77234	0.00316	0.001704	0.000036
1.6	17934.153211	65.58	0.71924	0.01228	0.004029	0.000051
1.8	48815.256906	194.12	6.70537	0.03713	0.013520	0.001419
2.0	132871.377933	570.95	19.58949	0.10230	0.039730	0.004005

TABLE 8

t	$x(t)$	δ_{SP}^4	δ_{SP}^8	δ_{SP}^{16}	δ_{SP}^{32}	δ_{SP}^{64}
0.2	15.309691	0.04	0.00823	0.00021	0.000291	0.000046
0.4	43.334337	6×10^{-5}	0.00316	0.00076	0.000368	0.000058
0.6	119.513222	0.22	0.01232	0.00196	0.000193	0.000042
0.8	326.588900	0.60	0.01626	0.00285	0.000452	0.000066
1.0	889.478955	1.67	0.01392	0.00220	0.000334	0.000049
1.2	2420.772761	5.22	0.00216	0.00007	0.000032	0.000003
1.4	6588.865818	15.56	0.00108	0.00004	2×10^{-7}	4×10^{-7}
1.6	17934.153211	46.07	0.00075	0.00003	0.000002	4×10^{-8}
1.8	48815.256906	135.58	0.00002	0.00002	0.000003	2×10^{-7}
2.0	132871.377933	396.72	0.00069	0.00009	0.000014	0.000002

TABLE 9

t	$x(t)$	$\delta_{S_3}^4$	$\delta_{S_3}^8$	$\delta_{S_3}^{16}$	δ_{SP}^4	δ_{SP}^8	δ_{SP}^{16}
0.2	1.885611	0.002409	0.000311	0.000086	0.006696	0.000193	0.000183
0.4	2.967299	0.001213	0.000022	0.000403	0.024093	0.000916	0.000748
0.6	4.331245	0.004842	0.001300	0.000565	0.021832	0.000727	0.000472
0.8	6.342954	0.010185	0.004028	0.000170	0.006117	0.001556	0.000436
1.0	9.278242	0.005778	0.003518	0.000356	0.011526	0.001310	0.000254
1.2	13.563777	0.011080	0.001466	0.000741	0.008807	0.000168	0.000027
1.4	19.903791	0.010571	0.000014	0.000361	0.007846	0.000388	0.000024
1.6	29.212354	0.002034	0.000492	0.000141	0.004891	0.000145	0.000025
1.8	42.845032	0.015807	0.001791	0.000343	0.015039	0.000077	0.000002
2.0	62.841170	0.001850	0.001460	0.000224	0.021896	0.000072	0.000013

Example 5. This example is used in [1] for the identification problem which has a distributed delay and a discontinuous forcing function. This equation is

$$\frac{d}{dt}x(t) = -3x(t) - \int_1^0 x(t+\theta)\, d\theta + u_{.1}(t),$$

with initial function

$$x(\theta) \equiv 1 \quad \text{for } \theta \in [-1, 0],$$

where $u_{.1}(t)$ is defined by

$$u_{.1}(t) = \begin{cases} 1 & \text{on } [0, .1], \\ 0 & \text{otherwise.} \end{cases}$$

In Table 10 we give the numerical results. The approximations by S_3 seem to be slightly better initially than by SP. This is because the solution $x_t(\theta)$ has a discontinuous derivative with respect to θ for $t \leqq 1.1$. The superiority of SP on the interval where the solution is infinitely differentiable is again observed for $t \geqq 1.2$.

TABLE 10

t	$x(t)$	$\delta^8_{S_3}$	$\delta^{16}_{S_3}$	$\delta^{32}_{S_3}$	δ^8_{SP}	δ^{16}_{SP}	δ^{32}_{SP}
0.2	0.304159	0.000447	0.000099	0.000051	0.001490	0.000040	0.0000105
0.4	−0.463775	0.000060	0.000124	0.000024	0.001717	0.000151	0.000089
0.6	−1.173353	0.000424	0.000006	0.000014	0.003747	0.000519	0.000058
0.8	−1.796286	0.000113	0.000018	0.000043	0.004891	0.000624	0.000178
1.0	−2.307740	0.001049	0.000400	0.000101	0.003346	0.000082	0.000115
1.2	−2.497698	0.002418	0.000151	0.000138	0.001002	0.000014	0.000048
1.4	−2.121921	0.001361	0.000205	0.000073	0.000652	0.000013	0.000003
1.6	−1.236591	0.000108	0.000266	0.000022	0.000533	0.000027	0.000002
1.8	0.051303	0.001261	0.000063	0.000019	0.000482	0.000037	0.000003
2.0	1.606226	0.001956	0.000227	0.000014	0.000693	0.000092	0.000002

From the discussion and numerical results presented here we can conclude the following. For the Legendre-tau approximation, the effort in implementing the algorithm is as much as for AV. But the accuracy of approximation is as good as that which can be obtained using high order spline approximation. When the solution to a problem is infinitely differentiable, the rate of convergence is faster than any finite power of $(1/N)$. In addition, from our calculations, our method appears to be about four times as fast as the cubic spline approximation method. These properties are much more evident for examples (not presented in this paper) where optimal control and the approximation of eigenvalues were considered.

REFERENCES

[1] H. T. BANKS, J. A. BURNS AND E. M. CLIFF, *Parameter estimation and identification for systems with delays*, this Journal, 19 (1981), pp. 791-828.

[2] H. T. BANKS AND F. KAPPEL, *Spline approximations for functional differential equations*, J. Differential Equations, 34 (1979), pp. 496-522.

[3] H. T. BANKS AND J. A. BURNS, *Hereditary control problems: Numerical methods based on averaging approximations*, this Journal, 16 (1978), pp. 169-208.

[4] C. BERNIER AND A. MANITIUS, *On semigroups in $\mathbb{R}^n \times L^p$ corresponding to differential equations with delays*, Can. J. Math., XXX (1980), pp. 969-978.

[5] J. G. BORISOVIC AND A. S. TURBABIN, *On the Cauchy problem for linear nonhomogeneous differential equations with retarded argument*, Soviet Math. Dokl., 10 (1969), pp. 401-405.

[6] C. CANUTO AND A. QUARTERONI, *Approximation results for orthogonal polynomials in Sobolev spaces*, Math. Comp., 38 (1982), pp. 67-86.

[7] P. J. DAVIS AND P. RABINOWITZ, *Methods of Numerical Integration*, Academic Press, New York, 1975.

[8] M. C. DELFOUR AND S. K. MITTER, *Hereditary differential systems with constants delays*: I. *General case*, J. Differential Equations, 12 (1972), pp. 213-235.

[9] D. GOTTLIEB AND S. A. ORSZAG, *Numerical Analysis of Spectral Methods: Theory and Applications*, CBMS Regional Conference Series in Applied Mathematics 26, Society for Industrial and Applied Mathematics, Philadelphia, 1977.

[10] T. KATO, *Perturbation Theory for Linear Operators*, Foundations of Mathematical Science, 132, Springer-Verlag, Berlin, 1966.

[11] D. L. KRIEDER, R. G. KULLER, D. R. OSTBERG AND F. W. PERKINS, *An Introduction to Linear Analysis*, Addison-Wesley, Reading, MA, 1966.

[12] C. LANCZOS, *Applied Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1956.

[13] N. N. LEBEDEV, *Special Functions and Their Applications*, Dover, New York, 1972.

[14] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Applied Mathematical Sciences 44, Springer-Verlag, Berlin, 1983.

[15] R. B. VINTER, *On the evolution of the state of linear differential delay equations in M^2: Properties of the generator*, J. Inst. Math. Appl., 21 (1978), pp. 13-23.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 4, July 1986

013

FINITE TIME CONTROLLERS*

V. T. HAIMO†

Abstract. Continuous finite time differential equations are introduced as fast accurate controllers for dynamical systems. These have qualities superior to controllers which are currently in use in such applications as robotics. The structure of the phase portrait for scalar second order finite time systems is determined. This characterization is used to develop a class of second order finite time systems which can be used as controllers.

Key words. control systems, nonlinear stability, finite time control, ordinary differential equations

1. Introduction. A standard problem in system theory is to develop controllers which drive a system to a given position as fast as possible. Consider

$$\dot{x} = f(x) + u(t)g(x), \qquad x \text{ in } R^n$$

where f models the natural dynamics of the system and g the effect of the control u. An example would be the positioning of a robotic manipulator at a set point in space.

There is a considerable body of research on linear feedback input for multi-dimensional systems, that is feedback control laws of the form

$$u(t)g(x) = Kx, \qquad K : R^n \to R^n.$$

This research has been concerned with finding K so that certain performance criteria are met by the feedback system (see e.g., [1] and [2]). Linear feedback may be quite good from the point of view of accuracy of tracking and placement. It has the disadvantage, however, that solutions of the feedback system are exponential functions of time if f is smooth, since the system behaves linearly in a neighborhood of the set point. Thus convergence can never occur in finite time. Whether or not this presents a serious practical problem will depend on the application.

One may ask whether it is possible to control a system to equilibrium in finite time using a bounded control. A standard textbook solution to this problem is to use a bang-bang control strategy (see, for example, [3]). Such controls optimize the time to reach equilibrium for trajectories of

$$\dot{x} = Ax + Bu, \qquad x \text{ in } R^n, \qquad u \text{ in } R, \quad |u| \leqq 1.$$

It turns out that the optimal control $u(x)$ is discontinuous, and switches from $u = 1$ to $u = -1$ on specific contours in x space.

The implementation of such a discontinuous control strategy leads to decreased response time. There may also be unwanted side effects such as introduced vibrations which arise because of repeated overshooting of the switching contour, caused by errors in the implementation of the discontinuous control. We are thus led to rephrase the question posed above. Can one control a system to equilibrium using a bounded and continuous control law? We will develop such finite time controllers.

* Received by the editors January 30, 1985, and in revised form June 17, 1985. This work was supported in part by the National Science Foundation under grant numbers ECF-81-21428 and EFS-84-03923, and by the Office of Naval Research, under the Joint Services Electronics Program contract number N0014-75-C-0648.

† Graduate School of Business Administration, Harvard University, Boston, Massachusetts 02163.

2. Definitions. We will discuss qualitative properties of differential equations. Some vocabulary which arises in this context is herewith defined. Suppose that

$$\dot{x} = f(x), \qquad x \text{ in } R^n,$$

and that $x(t, x_0)$ denotes a solution which passes through x_0 at $t = 0$. We will often call solutions *trajectories.* When $x(t, x_0)$ is regarded as a map from R^{n+1} to R^n_i then it will be called the *flow* of $\dot{x} = f$. A set is *invariant with respect to the flow* if all solutions intersecting the set are contained in it.

An *equilibrium point* is a point, x, such that $f(x) = 0$.

An equilibrium point is *asymptotically stable* if

(i) for any $\varepsilon > 0$ there exists a $\delta > 0$ so that $\|x_0\| < \delta \to \|x(t, x_0)\| < \varepsilon$, for all $t \geqq 0$, and

(ii) there exists a neighborhood of 0, U, so that all trajectories which enter U converge to the origin.

Here we have used the double bar to denote the Euclidean norm, as we shall continue to do throughout the paper.

In studying stability Lyapunov theory is very useful. The idea is to find a function which when restricted to a trajectory is a strictly decreasing function of time. If the restricted function has a unique minimum which is at the origin, then the trajectory must converge to that equilibrium.

More formally: suppose there exists $v(x)$ so that $v(0)$ is the unique minimum of v in a neighborhood of $x = 0$. Suppose also that v is C^1 and $\dot{v} = \langle \text{grad } v, f(x) \rangle < 0$ except at 0 where it vanishes. (Here grad v denotes the gradient of v with respect to the standard Riemannian metric on E^n.) Such a function will be called a *Lyapunov function* for $\dot{x} = f(x)$. Since we are only interested in studying local stability properties of ordinary differential equations, we need only define such a function in a neighborhood of zero. A Lyapunov function is *positive definite* if it is positive except at zero where it vanishes.

3. Finite time systems. It is appropriate to limit discussion to differential equations with an isolated equilibrium point at the origin, and no other equilibria, because we are interested, for example, in the behavior of a robot arm in the neighborhood of a set point. This set point is modeled as an isolated equilibrium and, as we are studying local behavior, other equilibria are not of concern. We will call differential equations with the properties that the origin is asymptotically stable, and all solutions which converge to zero do so in finite time, *finite time differential equations.* Unless otherwise specified all right-hand sides of differential equations will be C^1 everywhere except at zero, where they will be assumed to be continuous, and to have an isolated equilibrium.

One notices immediately that finite time differential equations cannot be Lipschitz at the origin. As all solutions reach zero in finite time, there is nonuniqueness of solutions through zero in backwards time. This, of course, violates the uniqueness condition for solutions of Lipschitz differential equations.

In one dimension necessary and sufficient conditions for the finite time property may be found easily. We have

FACT 1. $\dot{x} = r(x)$, $r(0) = 0$, *x in R, is finite time iff*

(i) $xr(x) \leqq 0$ *and equals* 0 *only at* $x = 0$, *for x in a neighborhood of* 0, *and*

(ii) $\int_p^0 (dx/r(x)) < \infty$ *for all p in R.*

Here (i) determines the asymptotic stability of the origin and (ii) determines the finite time property.

The proof is left to the reader.

Let $\text{sig}^a z = (\text{sgn } z)|z|^a$, for z and a in R.

Example 1. $\dot{x} = -\text{sig}^{1/2} x$ is a finite time equation.

One may use Lyapunov theory to extend Fact 1 to the multidimensional case.

PROPOSITION 1. *Consider $\dot{x} = g(x)$ with x in R^n and g continuous. If v is a positive definite Lyapunov function for $\dot{x} = g$, and if $\dot{v} \leqq r(v)$, where $\dot{z} = r(z)$, z in R, is a finite time equation, then $\dot{x} = g$ is also finite time.*

Proof. $\dot{v} \leqq r(v)$ implies that $dv/r(v) \geqq dt$ since by Fact 1 $r(v) < 0$ for $v > 0$. We then have

$$\infty > \int_p^0 dv/r(v) \quad \text{(again by Fact 1)} \quad \text{and}$$

$$\int_p^0 dv/r(v) \geqq \int_0^T dt = T,$$

where the trajectory of $\dot{x} = g$ with initial condition $x(0) = p$ reaches the origin at $t = T$, for $T \leqq \infty$. Since this time to origin is finite, then $\dot{x} = g$ is finite time.

Example 2.

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} -1 & 1/2 \\ 1/2 & -1 \end{bmatrix} \begin{bmatrix} \text{sig}^{1/2} x_1 \\ \text{sig}^{1/2} x_2 \end{bmatrix}$$

is finite time.

Proof. Let $v = x_1^2 + x_2^2$. Then

$$\dot{v} = -|x_1|^{3/2} - |x_2|^{3/2} + 1/2(x_1 \text{ sig}^{1/2} x_2 + x_2 \text{ sig}^{1/2} x_1).$$

One may show this is negative definite as follows. Let

$$C = \{(x_1, x_2): -|x_1|^{3/2} - |x_2|^{3/2} = -1\}.$$

By symmetry one easily sees that $\dot{v}$ achieves its maximum on the set C when $x_1 = x_2$. But $\dot{v} < 0$ when $x_1 = x_2$ and so $\dot{v} < 0$ when $x \neq 0$.

We need to show that $\dot{v} < r(v)$ where r is the right-hand side of a finite time differential equation. Letting $r(z) = -z^{4/5}$, we note that $\dot{v} < r(v)$ because $h(v, \dot{v}) = v^{4/5} + \dot{v}$ is largest when $x_1 = x_2$, and there it is negative. Thus by Proposition 1 $\dot{x} = u$ is finite time.

4. Second order systems. Systems of particular interest in many control theoretic situations are second order systems, and one may ask whether it is possible to generate continuous finite time controllers for second order systems.

Second order systems may of course be represented as first order systems with a special structure. One notices immediately that second order systems have at least one Lipschitz component since if

$$\ddot{x} = w(x, \dot{x})$$

then letting $x = x_1$ and $\dot{x} = x_2$

$$\dot{x}_1 = x_2$$

and

$$\dot{x}_2 = w(x_1, x_2).$$

The following theorem describes the behavior of finite time systems which have at least one Lipschitz component.

THEOREM 1. *Suppose that $\dot{x} = g(x)$ is finite time, with x in R^n, $g(0) = 0$, and g in C^1 on $R^n - \{0\}$, and that $g_i(x)$ is Lipschitz at $x = 0$, for some i. If $x(t)$ is a solution which*

reaches zero at $t = T < \infty$ *then*

$$\lim_{t \to T} \frac{x_i(t)}{\|x(t)\|} = 0.$$

Proof. Suppose $x(t, p_0)$ is a solution of $\dot{x} = g$ with $x(0, p_0) = p_0$ and $x(T, p_0) = 0$. By the mean value theorem, there is some q in $[0, T]$ so that

$$0 = x_i(T, p_0) = x_i(0, p_0) + Tg_i(x(q, p_0))$$

or

$$\frac{g_i(x(q, p_0))}{x_i(0, p_0)} = -\frac{1}{T}.$$

T may be considered to be a function of the initial condition p, where $T(p)$ is the time to origin for the trajectory beginning at p. We then have

$$\frac{g_i(x(q(p), p))}{x_i(0, p)} = -\frac{1}{T(p)}.$$

One may take the limit as $p \to 0$ along the trajectory through the point p_0.

Since $x_i(t, p)$ is a smooth function of t and vanishes at $t = T(p)$ then

$$\lim_{p \to 0} \left| \frac{x_i(q, p)}{x_i(0, p)} \right| = 1.$$

Thus

$$\lim_{p \to 0} \frac{g_i(x(q, p))}{x_i(q, p)} \frac{x_i(q, p)}{x_i(0, p)} = \lim_{p \to 0} -\frac{1}{T(p)} = -\infty$$

and so

$$\lim_{p \to 0} \frac{g_i(x(q, p))}{x_i(q, p)} \to -\infty.$$

g_i is Lipschitz so $g_i / \|x\|$ is bounded. Thus

$$\lim_{p \to 0} \frac{g_i(x(q, p))}{\|x(q, p)\|} \frac{\|x(q, p)\|}{x_i(q, p)} = -\infty$$

which implies that

$$\lim_{p \to 0} \frac{x_i}{\|x\|} = 0.$$

This result tells us that the trajectories of a second order finite time system converge in the state space along a hyperplane $x_1 = 0$ (where x_1 denotes the position of the system as above) since such a system must have at least one non-Lipschitz component.

Theorem 1 implies that for the system to reach zero in finite time trajectories must enter the region where the non-Lipschitz terms dominate.

We restrict our search for second order finite time systems to scalar problems. By Theorem 1 we know that trajectories of finite time systems in the $(x, \dot{x})$ plane converge tangent to the line $x = 0$. This tells us (among other things) that finite time trajectories do not spiral around the origin infinitely often as they approach it.

In trying to generate examples of differential equations with certain asymptotic behavior one may frequently exploit the fact that there are contours which may only

be crossed in certain directions, or not at all (as in the case of contours which are invariant with respect to the flow) in order to trap the trajectory into some region. This is of course the heart of Lyapunov theory. If one wishes to show that a second order system is finite time, one could search for a contour that prevented trajectories from spiraling around the origin. It seems natural to search for a contour which is itself invariant. This idea lies at the core of the next two theorems.

THEOREM 2. *Consider the scalar differential equation* $\ddot{x} = g(x, \dot{x})$ *with* $g(0, 0) = 0$. *Let* g *be in* C^1 *except at the origin where it is only assumed to be continuous. Suppose that the origin is asymptotically stable. Then all trajectories which reach zero do so in finite time if and only if*

(i) *there exists a solution* q *to the scalar differential equation*

$$q(z)\frac{dq}{dz} = g(z, q(z)), \qquad q(0) = 0,$$

such that $\dot{z} = q(z)$ *is a finite time scalar differential equation,*

(ii) *every solution* p *to*

$$p(z)\frac{dp}{dz} = g(z, p(z)), \qquad p(0) = 0,$$

is such that $\dot{z} = p(z)$ *is a finite time differential equation.*

Proof. Note that the analysis may be restricted to a sufficiently small neighborhood of the origin, N, such that all solutions with initial conditions in N converge to zero.

To prove sufficiency, two Lemmas are required. The structure of the proof is

$$\text{Lemma 1} \rightarrow \text{Lemma 2} \rightarrow \text{sufficiency.}$$

LEMMA 1. *Suppose* $(x(t), \dot{x}(t))$ *is a solution of* $\ddot{x} = g(x, \dot{x})$ *with* $x(T) = \dot{x}(T) = 0$ *for* $T \leqq \infty$. *Then there is an* S, *with* $0 < S < T$ *such that for* $S < t < T$ $x(t)\dot{x}(t) < 0$.

Proof. If $x(t_1)\dot{x}(t_1) > 0$, for some $t_1 \geqq 0$, then there is a $t_2 > t_1$ such that $x(t_2)\dot{x}(t_2) < 0$. Otherwise $|x(t)|$ is always increasing for $t > t_1$ and thus $x(t)$ cannot converge to zero.

Suppose there is no S such that $x(t)\dot{x}(t) < 0$ for $S < t < T$. This implies that if $x(t)\dot{x}(t) < 0$ for $t \geqq 0$ there is an $s > t$ so that $x(s)\dot{x}(s) = 0$. This in turn implies that $x(s) = 0$ for the following reason: One may show fairly easily that $xg(x, 0) < 0$ for all nonzero x, since the origin is the unique equilibrium solution (in N) of $\ddot{x} = g$, and since the origin is asymptotically stable. Thus the vector field $(\dot{x}, g(x, \dot{x}))$ points into the region in $(x, \dot{x})$ space, $x\dot{x} < 0$, along the line $\dot{x} = 0$, and so trajectories leaving this region must exit through the line $x = 0$.

There is a sequence of times $\{t_i\}$, (with $i = 1, \cdots, \infty$ and $\lim t_i = T$) with $x(t_i) = 0$ and $\dot{x}(t_i) \neq 0$. Note that $\dot{x}(t_i)\dot{x}(t_{i+1}) < 0$. Thus

$$(\dot{x}(t_i))(\dot{x}(t_{i+1})) = (\dot{x}(t_i) - q(x(t_i))(\dot{x}(t_{i+1}) - q(x(t_{i+1})))) < 0.$$

(The equality holds since $q(0) = 0$ and $x(t_i) = 0$ for all i.)

This shows that the function on R^2, $H(x, \dot{x}) = \dot{x} - q(x)$ (here we are regarding x and $\dot{x}$ as variables in R^2 rather than as functions of t) passes through zero along the trajectory $x(t), \dot{x}(t)$. By assumption (i), however, the contour in R^2, $H(x, \dot{x}) = 0$, is invariant with respect to the flow. Thus $H(x, \dot{x})$ cannot change sign when it is evaluated along the trajectory $x(t), \dot{x}(t)$. This contradiction shows that Lemma 1 holds.

LEMMA 2. *Suppose* $(x(t), \dot{x}(t))$ *is a solution of* $\ddot{x} = g(x, \dot{x})$ *with* $x(T) = \dot{x}(T) = 0$ *for* $T \leqq \infty$. *There is an* $r \in R^+$, *and a function* h *which is continuous and smooth away from* 0, *such that for* $T \geqq t > r$

$$\dot{x}(t) = h(x(t)), \qquad h(0) = 0.$$

Proof. By Lemma 1 there is an S satisfying $0<S<T$ so that for $S<t<T$ $x(t)\dot{x}(t)<0$. Clearly x and $\dot{x}$ do not change sign for such t. If $x(t)$ is positive, then it decreases as a function of t, and increases if it is negative. Thus no value of x is reached twice along the trajectory $(x(t), \dot{x}(t))$ for $t>S$. This implies that $\dot{x}(t)$ may be expressed as a function of $x(t)$: $\dot{x}(t)=h(x(t))$ for $t>S$. $h(x)$ is smooth, except perhaps at $x=0$ because $\dot{x}=h(x)$, so $\ddot{x}=(dh/dx)\dot{x}$ or $dh/dx=\ddot{x}/\dot{x}$, and $\ddot{x}(t)/\dot{x}(t)$ is continuous (as a function of x and $\dot{x}$; not t) except perhaps at $\dot{x}=0$. $\dot{x}=0$ on the trajectory only when $x=0$.

h is continuous at zero: $\lim_{x\to 0} h(x)=0$. Lemma 2 is proved. We show that it implies sufficiency.

Let $(x(t), \dot{x}(t))$ be a solution of $\ddot{x}=g(x,\dot{x})$ with $x(T)=\dot{x}(T)=0$ for $T\leqq\infty$. By Lemma 2 there is a function h such that $\dot{x}(t)=h(x(t))$ for sufficiently large t. We thus have

$$\ddot{x}=g(x,h(x))=h(x)\frac{dh}{dx} \tag{1}$$

except at $x=0$. But by L'Hospital's rule

$$\lim_{x\to 0}\frac{h(x)}{x}=\lim_{x\to 0}\frac{g(x,h(x))}{h(x)}=\frac{dh}{dx} \quad \text{at } x=0,$$

showing that (1) is satisfied for $x=0$. Thus h satisfies the conditions of assumption (ii) implying that $\dot{x}=h(x)$ must be finite time. Under these conditions x reaches zero in finite time, and since $h(0)=0$ so does $\dot{x}$.

As we started with an arbitrary trajectory all solutions must reach zero in finite time.

We prove necessity. Suppose $\ddot{x}=g(x,\dot{x})$ is finite time. We know by Theorem 1 that $\lim_{t\to T}(\dot{x}/x)=\pm\infty$ for a solution $(x,\dot{x})$ with $x(T)=\dot{x}(T)=0$.

Consider the function $x\dot{x}$. If there is a sequence of times $\{t_i\}$, $i=1,\cdots,\infty$, with $\lim t_i=T$, where $(x\dot{x})(t_i)=0$ for all i, then there is a sequence of times $\{s_j\}$, $j=1,\cdots,\infty$, so that $\dot{x}(s_j)=0$, for all j. Suppose this were not true. Then there would be some I so that for $i>I$

$$x\dot{x}(t_i)=0\Rightarrow x(t_i)=0.$$

If $x(t_i)=0$ and $x(t_{i+1})=0$ then

$$\dot{x}(t_i)\dot{x}(t_{i+1})<0,$$

which indicates that $\dot{x}=0$ at some point between t_i and t_{i+1}. This contradicts our assumption.

Such a sequence, $\{s_j\}$, contradicts the fact that $\lim(\dot{x}/x)=\pm\infty$ as $t\to T$. Thus there is an $S<T$ so that for $S<t<T$ $x\dot{x}(t)<0$. If for instance $x<0$ then x is strictly increasing and so there is a functional relationship between x and $\dot{x}$:

$$\dot{x}(t)=q(x(t)), \quad \text{with } q(0)=0.$$

One may show that q is smooth except perhaps at $x=0$, so q clearly satisfies the relationship

$$q(x)\frac{dq}{dx}=g(x,q(x))$$

and also satisfies it in the limit as x goes to zero.

By assumption, the components of the solution to $\ddot{x} = g(x, \dot{x})$, x and $\dot{x}$, reach zero in finite time. Thus $\dot{x} = q(x)$ must be a finite time first order differential equation and so condition (i) is satisfied. Condition (ii) also holds since for every solution $p(z)$ to

$$p(z)\frac{dp}{dz} = g(z, p(z)), \qquad p(0) = 0,$$

one obtains a trajectory $(x, \dot{x})$ which satisfies $\dot{x} = p(x)$. Once again, since x and $\dot{x}$ reach zero in finite time then $\dot{x} = p$ must be a finite time differential equation.

One should note that associated to each second order scalar finite time equation, $\ddot{x} = g$, there is a first order discontinuous differential equation (described in condition (i) of Theorem 2). This equation has nonunique solutions through the initial condition $q(0) = 0$, and these solutions are nonzero for $z \neq 0$. This distinguishes them from the classical examples of nonunique solutions of such equations as $\dot{x} = \operatorname{sig}^{1/2} x$, with $x(0) = 0$.

Theorem 2 provides a qualitative description of the phase portraits of second order finite time scalar systems. It can also be used to generate examples as we shall see shortly.

The problem of determining the stability properties of an equilibrium point of a non-Lipschitz differential equation can be difficult. In Theorem 2, asymptotic stability of the origin is assumed. In the next theorem a special structure of g gives a Lyapunov function associated with g.

THEOREM 3. *Let $\ddot{x} = g(x, \dot{x}) = f(x) + d(\dot{x})$, where $f(0) = d(0) = 0$, g is C^1 except at zero where it is continuous, $(\dot{x}, g) = (0, 0)$ only at $(x, \dot{x}) = (0, 0)$, and f is monotone decreasing. Then $\ddot{x} = u$ is finite time if and only if*

(i) *there exists a solution $q(z)$ to the first order differential equation on R*

$$q(z)\frac{dq}{dz} = g(z, q(z)), \qquad q(0) = 0 \tag{2}$$

such that $\dot{z} = q(z)$ is a finite time differential equation, and

(ii) *any solution, h, to (2) is such that $\dot{z} = h(z)$ is a finite time equation.*

Proof. (Sufficiency). We need only prove that the origin is asymptotically stable, and then apply Theorem 2. This will be accomplished by using a Lyapunov function.

Consider the function defined in a neighborhood of the origin

$$v(x, \dot{x}) = (\tfrac{1}{2})\dot{x}^2 - \int_0^x f(z)\, dz.$$

This is positive definite since f is decreasing and $f(0) = 0$.

$$\dot{v} = \dot{x}(\ddot{x} - f(x)) = \dot{x}d(\dot{x}).$$

We will show that this is negative for sufficiently small $(x, \dot{x})$ except when $\dot{x} = 0$.

By condition (i) there is a solution $q(z)$ to

$$q(z)\frac{dq}{dz} = f(z) + d(q(z)),$$

with $q(0) = 0$. $\dot{z} = q$ is finite time, so by Fact 1, dq/dz is negative. If $z > 0$ then $q(z)(dq/dz) > 0$ and so $f(z) + d(q(z)) > 0$. This implies that $d(q(z)) > 0$. (The last inequality holds because $f(z) < 0$ for $z > 0$.) Thus $d(y) > 0$ for $y < 0$. Similarly one may show that $d(y) < 0$ if $y > 0$.

We note that all trajectories are bounded as $t \to \infty$, because if $v(x_0) = c (c \text{ in } R^+)$ then $v(x(t, x_0)) \leqq c$. Thus we may use LaSalle's theorem [4] to note that all trajectories converge to the origin.

Necessity is a tautology because the assumption that $\ddot{x} = g$ is finite time includes the asymptotic stability of the origin. Thus we are back in the situation of Theorem 2.

5. Examples of second order systems. With Theorem 3 we may generate a class of finite time second order systems. These are presented in the following corollary.

COROLLARY 1. *Let* $\ddot{x} = -\text{sig}^a x - \text{sig}^b \dot{x} = g$ *with* $a > 0, b > 0$. *If* (A) $b < 1$ *and* (B) $a > b/(2-b)$, *then* $\ddot{x} = g$ *is finite time.*

Proof. $-\text{sig}^a x$ is monotone decreasing. In order to apply Theorem 3, we must prove that

(i) there exists a finite time solution q to

$$q\frac{dq}{dx} = -\text{sig}^a x - \text{sig}^b q(x) \quad \text{with } q(0) = 0,$$

and

(ii) that all such solutions are finite time.

Suppose that

$$q\frac{dq}{dx} = -\text{sig}^a x - \text{sig}^b q(x), \qquad q(0) = 0. \tag{3}$$

We would like to find a solution to this differential equation. The first thing that one notices about it, however, is that it is discontinuous when $q = 0$. Thus the existence theorems for solutions of continuous differential equations are not applicable. We will generate some continuous differential equations from (3).

Consider $p(x) = (1/2)q(x)^2$. This gives

$$\frac{dp}{dx} = q(x)\frac{dq}{dx} = -\text{sig}^a x - \text{sig}^b q(x).$$

Suppose we knew that $xq(x) < 0$ for $x \neq 0$.

If $x \geqq 0$ and $p \geqq 0$ then

$$\frac{dp}{dx} = -\text{sig}^a x + rp(x)^{b/2}, \qquad p(0) = 0 \tag{4}$$

(where $r = 2^{1/2}$), and if $x \leqq 0, p \geqq 0$ then

$$\frac{dp}{dx} = -\text{sig}^a x - rp(x)^{b/2}, \qquad p(0) = 0. \tag{5}$$

Clearly if each of these equations has a positive solution, p, for $x \geqq 0, x \leqq 0$ respectively, and if $xq(x) < 0$ for nonzero x, then there is a solution to (3) for all x.

We first prove that $xq(x) < 0$ for nonzero x. Suppose that $x > 0$ and $q > 0$. Then

$$q\frac{dq}{dx} < 0 \quad \text{so} \quad \frac{dq}{dx} < 0.$$

If $q(0) = 0$, however, this means we have a function which vanishes at zero, is positive for $x > 0$, where it has a negative derivative. This contradiction leads us to conclude that any solution q of (3) is such that $xq(x) < 0$ for $x > 0$. One can show similarly that $xq(x) \leqq 0$ for all x and only vanishes at the origin.

Both differential equations (4) and (5) are defined on closed sets. In order to use existence theorems for continuous differential equations on open sets to show that these equations have solutions, we must extend their domains of definition. This may be done as follows.

Consider the differential equations

$$\frac{dp}{dx} = -\operatorname{sig}^a x + r|p(x)|^{b/2}, \qquad p(0)=0 \tag{6}$$

and

$$\frac{dp}{dx} = -\operatorname{sig}^a x - r|p(x)|^{b/2}, \qquad p(0)=0. \tag{7}$$

These differential equations are defined for all real x and p, and reduce to (4) and (5) when $p>0$. If (6) has a solution p which is positive for $x>0$, and if (7) has a solution p which is positive for $x<0$ then (4) and (5) have the required solutions.

Consider (6). If there is a positive solution p for $x>0$, with $p(0)=0$ then $dp/dx>0$ for $x>0$ if x is sufficiently small. This implies that either

(i) $\lim_{x\to 0} \frac{dp}{dx} = \lim_{x\to 0} p(x)^{b/2}$ (limits from $x>0$)

or that

(ii) $\lim_{x\to 0} x^a = \lim_{x\to 0} p(x)^{b/2}$ (limits from $x>0$).

In case (i) we have that

$$\lim_{x\to 0} \frac{dp}{dx} = \lim_{x\to 0} p/x = \lim_{x\to 0} p^{b/2},$$

which implies that

$$\lim_{x\to 0} p(x) = \lim_{x\to 0} x^{2/(2-b)}.$$

Since the $p^{b/2}$ term dominates the $-x^a$ term, then we must have $a>b/(2-b)$.

In case (ii) we have that

$$\lim_{x\to 0} x^a = \lim_{x\to 0} p(x)^{b/2} \quad \text{or} \quad \lim_{x\to 0} p(x) = \lim_{x\to 0} x^{2a/b}.$$

In this case we have $2a/b-1 \geqq a$ since the x^a term dominates near 0. But $2a/b-1 \geqq a$ implies $a \geqq b/(2-b)$. In order for there to be a negative solution $p(x)$ to (6) for $x>0$ one may show similarly that it is necessary that $a \leqq b/(2-b)$.

We know that there is a solution $p(x)$, with $p(0)=0$ to (6), for $x>0$, by the existence theorem for solutions of continuous differential equations on open sets. Thus if $a>b/(2-b)$ this solution must be positive.

One may show by the same method that (7) may have a positive solution for $x<0$ if and only if $a \geqq b/(2-b)$ and must have a positive solution if the inequality is strict. Thus if assumption (B) holds, then (3) has a solution q. It is also clear from the above analysis that if assumption (B) holds then

$$\lim_{x\to 0} \frac{q(x)}{-\operatorname{sig}^z x} = 1 \quad \text{for } z = 1/(2-b). \tag{8}$$

We claim that $\dot{x} = q(x)$ is finite time if $b<1$. Define $h(x)$ by the equation $q(x) = h(x) - \operatorname{sig}^z x$. Then

$$\lim_{x\to 0} \frac{h(x)}{\operatorname{sig}^z x} = 0.$$

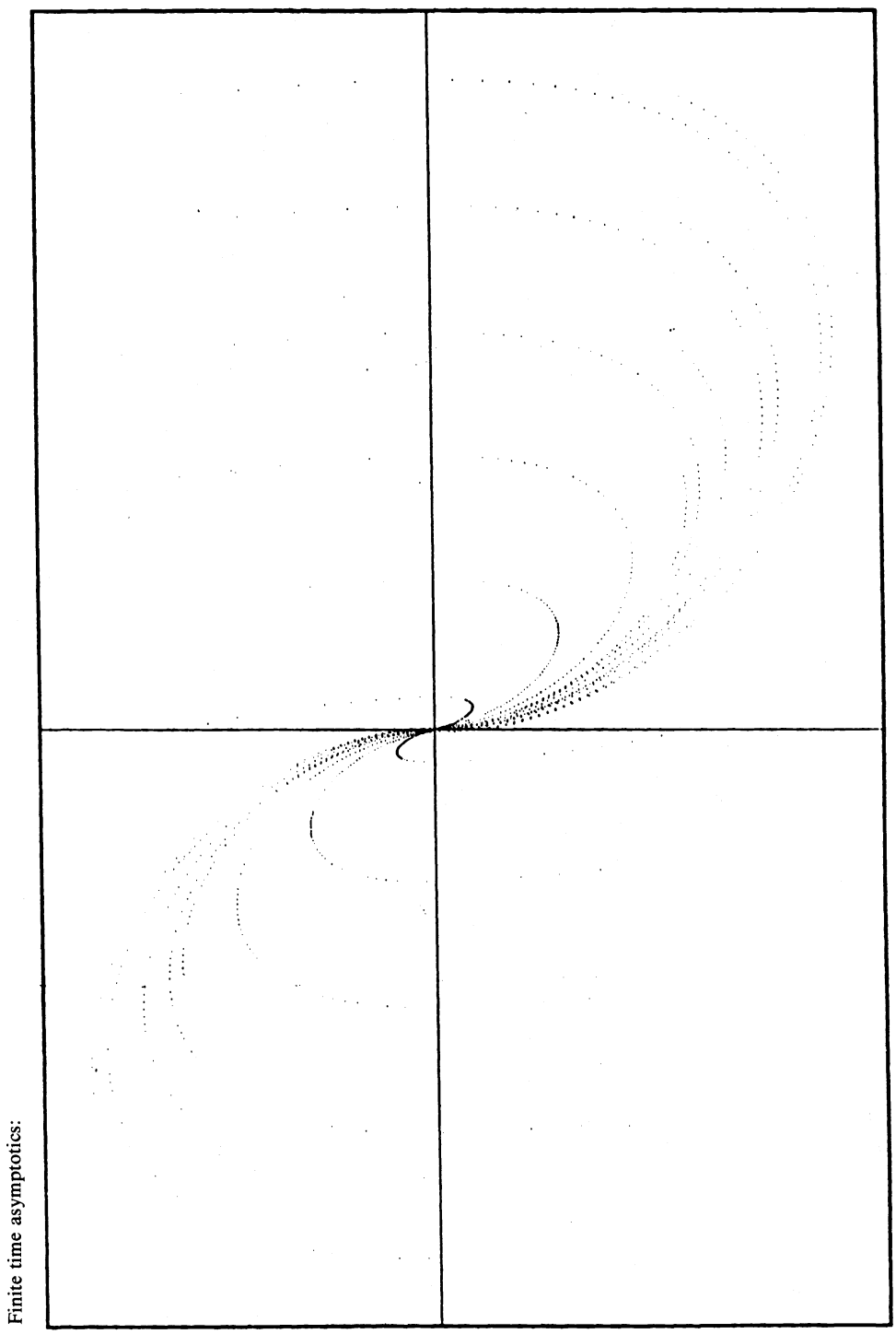


FIG. 1. *Finite time phase portrait.*

Consider the function $v(x) = |x|^{1-z}/1-z$, which is positive definite if $z<1$. Then

$$\dot{v} = \langle \operatorname{grad} v, q\rangle = (\operatorname{sig}^{-z} x)(h(x) - \operatorname{sig}^{z} x)$$

and

$$\lim_{x\to 0} \dot{v} = -1 \leqq -v^{1/2} \quad \text{near zero.}$$

Thus by Proposition 1 $\dot{x} = q$ is finite time if $z<1$, and $z<1$ implies that $b<1$. (Since the time to origin function v, for $\dot{x} = q$, is unbounded at zero if $b \geqq 1$ it follows that the equation $\dot{x} = q$ is finite time iff $b<1$.)

It only remains to show that all solutions of (3) are finite time. By the above analysis any solution of (3), q, satisfies (8). If $b<1$ then all such solutions are finite time.

Note that we have also shown that if $a < b/(2-b)$ or if $b \geqq 1$ then $\ddot{x} = g$ is not finite time. We have for instance

Example 3. $\ddot{x} = -\operatorname{sig}^{.4} x - \operatorname{sig}^{.5}\dot{x}$ is a finite time differential equation.

The phase portrait of this differential equation is displayed in Fig. 1.

It should be noted that this figure also displays the nonunique solutions of the scalar first order differential equation $q(x)(dq/dx) = -\operatorname{sig}^{.4} x - \operatorname{sig}^{.5} q(x)$, with $q(0) = 0$, where the x-axis is horizontal and the $q(x)$ axis is vertical.

A more cumbersome proof along the lines of the proof of Corollary 1 may be used to show that the following holds.

COROLLARY 2. *Let* $\ddot{x} = g = -\operatorname{sig}^{a} x - \operatorname{sig}^{b}\dot{x} + f(x) + d(\dot{x})$ *where* $a>0$, $1>b>0$, $a > b/(2-b)$, $f(0) = d(0) = 0, O(f) > O(|x|^{a})$ *and* $O(d) > O(|\dot{x}|^{b})$; *then* $\ddot{x} = g$ *is a finite time equation.*

The class of examples described in Corollary 2 is quite large. Theorem 2 describes the phase portrait of second order, finite time scalar systems, and also enables us to identify a class of these systems. These may in turn be used as finite time controllers.

For instance, the block design implementation for controlling a link of robot arm is displayed in Fig. 2. In this scheme I is the moment of inertia of the link.

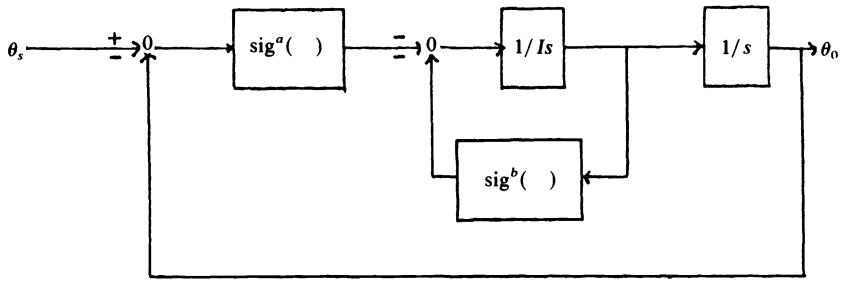


FIG. 2. *Second order servo-system with finite feedback.*

Acknowledgments. I would like to thank Roger Brockett for many helpful discussions on this material. John Baillieul also offered some interesting insights.

REFERENCES

[1] R. W. BROCKETT AND C. I. BYRNES, *Multivariable Nyquist criterion, root locus, and pole placement by output feedback*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 271-284.

[2] S. R. PECK, *Combinatorics of Schubert calculus and inverse eigenvalue problems*, Ph.D. Thesis, Harvard Univ., Cambridge, MA, 1984.

[3] A. E. BRYSON AND Y. C. HO, *Applied Optimal Control*, John Wiley, New York, 1975.

[4] J. P. LASALLE, *Some extensions of Lyapunov's second method*, IEEE Trans. Circuit Theory, CT-7 (1960), pp. 520-527.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 4, July 1986

014

A MULTI-RESPONSE QUADRATIC CONTROL PROBLEM*

SUNG J. LEE†

Abstract. We consider a new class of regular quadratic control problem in a compact interval, which is associated with a first order system of ordinary differential equations and very general boundary operators. In particular the state is chosen to minimize a functional and there may be infinitely many minimizers. The theory of least-squares solutions of multi-valued operator equations is used in the investigation.

Key words. first-order differential equation, adjoint, generalized inverse, multi-valued operator, optimal controller

AMS(MOS) subject classification. 49A10

1. Introduction. The linear quadratic control problem with single responses generated by an ordinary differential equation has had an extensive study. The main objective of the problem is to establish the existence of an optimal controller and describe it by a system of differential equations. See a recent survey article [2]. Single responses are generated when an initial condition or a certain two-point condition is imposed in the state equation. But if an insufficient initial condition or a general or generalized boundary condition is imposed on the state equation, then a controller may generate multiple responses and so the associated cost functional becomes a multi-valued map, making the situation more complicated. This is a mathematically interesting case with a potential application, and has not yet been investigated in full in the literature. Thus we will consider a new class of quadratic control problem where the state is subject to a system of ordinary linear differential equations in a compact interval and is chosen to minimize a functional with possibly infinitely many minimizers. Because of the nature of multiple responses we will introduce a new method for the problem. The idea of the investigation is to change the control problem into a least-squares problem for a multi-valued operator equation in Hilbert space, and then characterize the result by Fredholm integro-differential equations and some side conditions. This process involves the recent results on least-squares solutions and generalized inverses for multi-valued linear operators developed in [7], [8], [9]. The main results of this paper are contained in Theorem 3.7 for the existence and Theorem 3.12 for the characterization.

2. Adjoint, least-squares solutions and generalized inverses of linear manifolds. In this section we will state some recent results on multi-valued linear operators which will be needed in the next section. In this paper we will not distinguish an operator (single-valued) from its graph. A vector space will be called a linear manifold. Let H_1, H_2, H_3 be Hilbert spaces over the real field $\mathbb{R}$ or the complex field $\mathbb{C}$. For notational convenience, the inner products of these spaces are denoted by $\langle \cdot, \cdot \rangle$ while their norms are denoted by $\| \cdot \|$. $H_1 \oplus H_2$ will denote the Hilbert space of all ordered pairs $\{x_1, x_2\} (x_1 \in H_1, x_2 \in H_2)$ with its inner product $\langle \cdot, \cdot \rangle$ defined by

$$\langle \{x_1, x_2\}, \{y_1, y_2\} \rangle := \langle x_1, y_1 \rangle + \langle x_2, y_2 \rangle,$$

and its norm $\| \cdot \|$ defined by

$$\| \{x_1, x_2\} \| := (\|x_1\|^2 + \|x_2\|^2)^{1/2}.$$

* Received by the editors January 24, 1984, and in revised form March 19, 1985.
† Department of Mathematics, University of South Florida, Tampa, Florida 33620.

Let M be a linear manifold (also called a linear relation) in $H_1 \oplus H_2$. Then

$$\begin{aligned}
&\text{Dom } M := \{x: \{x, y\} \in M \text{ for some } y\},\\
&\text{Range } M := \{y: \{x, y\} \in M \text{ for some } x\},\\
&\text{Null } M := \{x: \{x, 0\} \in M\},\\
&M(x) := \{y: \{x, y\} \in M\},\\
&M^{-1} := \{\{x, y\}: \{y, x\} \in M\} \ (x \in \text{Dom } M),\\
&M^{\perp} \text{ (The orthogonal complement of } M)\\
&\quad := \{\{x, y\} \in H_1 \oplus H_2: \langle \{x, y\}, m\rangle = 0 \text{ for all } m \in M\},\\
&M^* \text{ (the adjoint of } M) := \{\{x, y\}: \{y, -x\} \in M^{\perp}\}.
\end{aligned}$$

One can show easily that M^* is a closed linear manifold, and M^* is an operator if Dom M is dense. Moreover, if M is a densely defined closed operator, then M^* is the graph of the usual adjoint operator of M. For this result and other related results, see [6] where related references may be found.

Suppose that $S \subset H_2 \oplus H_3$ is a linear manifold. Then

$$SM := \{\{x, y\} \in H_1 \oplus H_3: \{x, z\} \in M, \{z, y\} \in S \text{ for some } z\}.$$

Let P be the orthogonal projector from H_1 onto Null M, i.e., P is the bounded linear operator from H_1 onto Null M such that $P^2 = P$ and $P^* = P$. Let P^+ be the orthogonal projector from H_2 onto Null M^*. Define a linear manifold $M^\#$ in $H_2 \oplus H_1$ by

$$M^\# := [\text{graph } (I - P)]M^{-1}[\text{graph } (I - P^+)]$$

where I denotes the identity operator. $M^\#$ is called the orthogonal generalized inverse of M.

The following theorem can be found in [8] (see also [7]).

THEOREM 2.1. *Let M be a closed linear manifold in $H_1 \oplus H_2$. Then*

(i) *$M^\#$ is a closed linear operator with*

$$\text{Dom } M^\# = \text{Range } M \dotplus \text{Null } M^*,$$

$$\text{Range } M^\# = \text{Dom } M \cap (\text{Null } M)^{\perp},$$

where $\dotplus$ denotes an algebraic direct sum. In particular, Range M *is closed if and only if* Dom $M^\# = H_2$.

(ii) $M^\# = ((I - P)M^{-1}) \dotplus (\text{Null } M^* \oplus \{0\})$, *direct sum, where P is the orthogonal projector of H_1 onto* Null M. *In particular, $\{y, g\} \in M$ if and only if $g \in$* Range M *and $y = M^\#(g) + k$ for some $k \in$* Null M.

We remark here that if M is a densely defined closed linear operator, then $M^\#$ is precisely the Moore-Penrose generalized inverse of M. Moreover, $R \equiv (I - P)M^-$ is the operator such that $M^{-1} = R \dotplus (\{0\} \oplus \text{Null } M)$, orthogonal direct sum.

The conjugate transpose and the transpose of a matrix N are denoted by N^* and $N^{\mathbf{t}}$ (bold faced $\mathbf{t}$ to distinguish it from time t), respectively. $\mathbb{C}^k$ will denote the k-dimensional complex Euclidean space with its inner product $\langle \cdot, \cdot \rangle$ and its norm $\| \cdot \|$ defined by

$$\langle x, y\rangle = x^{\mathbf{t}}\bar{y}, \qquad \|x\| = (x^*x)^{1/2},$$

where $\bar{y}$ denotes the complex conjugate of y (where we have identified any element of $\mathbb{C}^k$ as a *column* vector).

Suppose T is a $m \times n$ constant complex matrix. Then the Moore-Penrose inverse matrix of T (which is a $n \times m$ matrix) is also denoted by $T^\#$. We can (and will) identify

T as a bounded operator from $\mathbb{C}^n$ into $\mathbb{C}^m$. Thus

$$\text{Null } T = \{x \in \mathbb{C}^n : Tx = 0\},$$

$$\text{Range } T = \{Tx : x \in \mathbb{C}^n\}.$$

When T is identified as an operator, its generalized inverse is an operator from $\mathbb{C}^m$ into $\mathbb{C}^n$, and this inverse can be identified as a $n \times m$ matrix with respect to the standard orthonormal bases of $\mathbb{C}^n$ and $\mathbb{C}^m$. This is precisely T^*. Hence the notation T^* is consistent with the general one given earlier.

We say that $u \in H_1$ is a least-squares solution, briefly LSS, of an inclusion $g \in M(x)$ (g is a given element in H_2) if $u \in \text{Dom } M$ and there exists $y \in M(u)$ such that

$$\|g - y\| = \text{Min}\{\|g - z\| : z \in \text{Range } M\}.$$

Notice that when M is an operator, this reduces to the usual definition of a least-squares solution.

The following can be found in [9].

THEOREM 2.2. *Let M be a linear manifold in $H_1 \oplus H_2$ and $g \in H_2$. Assume that M is closed. Then*

(i) *$g \in M(x)$ has a* LSS *if and only if $g \in \text{Dom } M^*$. In particular, if* Range *M is closed, then a* LSS *always exists.*

(ii) *u is a* LSS *of $g \in M(x)$ if and only if $u \in \text{Dom } M$ and $g \in M(u) \dotplus \text{Null } M^*$.*

(iii) *Assume that $g \in \text{Dom } M^*$. Then the set of all* LSS *are given by the coset*

$$M^*(g) \dotplus \text{Null } M.$$

In particular, $M^*(g)$ is the unique LSS *of smallest H_1-norm.*

3. Multi-response control problem. Let $[t_0, t_1]$ be a compact interval. For a natural number q, let X_q denote the Hilbert space of $x : [t_0, t_1] \to \mathbb{C}^q$ such that $\|x\| := (\int_{t_0}^{t_1} x^*(t)x(t)\,dt)^{1/2} < \infty$. The inner product $\langle \cdot, \cdot \rangle$ of X_q is defined by $\langle x, y \rangle := \int_{t_0}^{t_1} x^t(t)\overline{y(t)}\,dt$. In this paper all the analysis will take place within the field $\mathbb{C}$, but our results remain valid if $\mathbb{C}$ is replaced by the real field $\mathbb{R}$.

Let $A(t)(t \in [t_0, t_1])$ be an $n \times n$ continuous complex matrix-valued function. Let T_1 be the (graph of the) maximal differential operator contained in $X_n \oplus X_n$ defined by

$$T_1 x := \dot{x} - Ax \qquad (\dot{x} \equiv dx/dt)$$

for $x \in \text{Dom } T_1$, where $x \in \text{Dom } T_1$ if and only if $x \in X_n$, x is absolutely continuous entrywise in $[t_0, t_1]$ and $\dot{x} \in X_n$.

In this paper we will consider the problem (which will be referred to later as "the control problem"): Find $u^+ \in X_m$ (called an optimal controller), minimizing

$$\mathscr{C}(u, x) := \int_{t_0}^{t_1} (u^* Uu + (x - x_0)^* W(x - x_0))(t)\,dt + \|F_\Lambda(x)\|^2, \tag{3.1}$$

subject to

$$u \in X_m,\ x \in \text{Dom } T_1, \tag{3.2}$$

$$\dot{x}(t) = A(t)x(t) + B(t)u(t), \quad \text{almost all } t \in [t_0, t_1], \tag{3.3}$$

$$\|F_\Omega(x) - \gamma\| = \text{Min}\{\|F_\Omega(y) - \gamma\| : y \in \text{Dom } T_1,\ T_1 y = Bu\}. \tag{3.4}$$

Here

(i) $U(t)$, $W(t)$, $B(t)$ are $m \times m$, $n \times n$, $n \times m$ continuous complex matrix-valued functions of t. For each t, $U^*(t) = U(t)$, $W^*(t) = W(t)$, and $U(t)$ is positive definite and $W(t)$ is nonnegative definite.

(ii) x_0 is a given element of X_n and γ is a given element of $\mathbb{C}^d$.

(iii) F_Ω and F_Λ are the (boundary) operators defined on Dom T_1 by

$$F_\Omega(x) := \int_{t_0}^{t_1} (\Omega_1^*(t)x(t) + \Omega_2^*(t)\dot{x}(t))\, dt,$$

$$F_\Lambda(x) := \int_{t_0}^{t_1} (\Lambda_1^*(t)x(t) + \Lambda_2^*(t)\dot{x}(t))\, dt,$$

where $\Omega_1(t)(n\times d)$, $\Omega_2(t)(n\times d)$, $\Lambda_1(t)(n\times d_1)$, $\Lambda_2(t)(n\times d_1)$ are complex matrix-valued functions of $t\in[t_0, t_1]$ whose columns are X_n.

When $\{u^+, x^+\}$ minimizes $\mathscr{C}$, u^+ will be called an optimal controller and x^+ an optimal response corresponding to u^+.

We can say that the operator F_Ω (and hence F_Λ) is the most general boundary operator of finite-dimensional range generated by the formal expression $\dot{x} - Ax$ in the sense that F_Ω is a continuous linear operator from the Hilbert space Dom T_1 equipped with the T_1-norm defined by

$$\|x\|_{T_1} := (\|x\|^2 + \|T_1x\|^2)^{1/2}$$

into the finite-dimensional space $\mathbb{C}^d$ if and only if F_Ω has the representation as in the above. This can be proved easily using the Riesz–Fischer representation theorem. This operator can be a familiar general point-evaluation operator (see Lemma 3.4 below).

The main feature of this paper is: (i) the "most" general boundary operators are involved, (ii) the state is chosen to minimize $\|F_\Omega(x) - \gamma\|$ and there may be infinitely many minimizers. Consequently, the functional $\mathscr{C}$ becomes a multi-valued map of u. In particular, for a given control u, a response x is chosen so that $F_\Omega(x)$ is as close to the ideal target γ as possible. This idea becomes clearer if $F_\Omega(x)$ becomes a point-evaluation operator such as $x(t_0)$ or $x(t_1)$.

There are two obvious ways of treating $\mathscr{C}$: One is to treat it as a single-valued map of u and x and the other is to view it as a multi-valued map of u. We will present a new method of treating the problem by viewing $\mathscr{C}$ as a multi-valued map. This will be based on the recent theories [7], [8], [9] of the least-squares solutions and generalized inverses of a multi-valued linear operator.

Let us define our dynamical system $\mathscr{D}\subset X_m\oplus X_n$ by

$$\mathscr{D} = \{\{u, x\}: u \text{ and } x \text{ satisfy (3.2), (3.3), (3.4)}\}.$$

For $\{u, x\}\in\mathscr{D}$, u is called a control and x a response corresponding to u. First we will describe $\mathscr{D}$. To do this let Φ be the $n\times n$ fundamental matrix solution satisfying

$$\dot{\Phi}(t) = A(t)\Phi(t), \qquad t_0\leqq t\leqq t_1,$$

$$\Phi(t_0) = I_{n\times n}.$$

Define operators $\mathscr{H}: X_m\to X_n$ and $H_\Omega: X_m\to\mathbb{C}^d$ by

$$\mathscr{H}(u) := \Phi(t)\int_{t_0}^{t} (\Phi^{-1}Bu)(s)\, ds, \qquad t_0\leqq t\leqq t_1, \tag{3.5}$$

$$H_\Omega(u) := \int_{t_0}^{t_1} (\Omega_1^*\mathscr{H}(u) + \Omega_2^*(A\mathscr{H}(u) + Bu))(t)\, dt. \tag{3.6}$$

Let Q_Ω be the $d\times n$ matrix defined by

$$Q_\Omega := \int_{t_0}^{t_1} ((\Omega_1^* + \Omega_2^*A)\Phi)(t)\, dt. \tag{3.7}$$

We have the following description of $\mathscr{D}$.

LEMMA 3.1.

[I] *The following are equivalent*:

(i) $\{u, x\} \in \mathscr{D}$.

(ii) $u \in X_m$, $x = \mathscr{H}(u) + \Phi(\alpha - Q_\Omega^\# H_\Omega(u)) + \Phi Q_\Omega^\# \gamma$ *for some* $\alpha \in \text{Null } Q_\Omega$.

(iii) $u \in X_m$, $x = \mathscr{H}(u) + \Phi\alpha$ *for some* $\alpha \in \mathbb{C}^n$ *which is a least-squares solution of the matrix equation* (*for u fixed*)

$$Q_\Omega(\beta) = \gamma - H_\Omega(u).$$

[II] $\mathscr{D} = \{0, \Phi Q_\Omega^\# \gamma\} \dot{+} \mathscr{S}$, *where*

$$\mathscr{S} := \{\{u, \mathscr{H}(u) + \Phi(\alpha - Q_\Omega^\# H_\Omega(u))\} : u \in X_m, \alpha \in \text{Null } Q_\Omega\}.$$

In particular, Dom $\mathscr{D} = X_m$ *and* $\mathscr{D}$ *is a closed convex set in* $X_m \oplus X_n$.

Proof. Let $u \in X_m$, $y \in \text{Dom } T_1$ satisfy (3.3). Then $y = \mathscr{H}(u) + \Phi\beta$ for some $\beta \in \mathbb{C}^n$. Now $\mathscr{H}(u)$ is the solution of (3.3), subject to $x(t_0) = 0$. Thus using the definition of F_Ω,

$$F_\Omega(y) = Q_\Omega(\beta) + H_\Omega(u).$$

Thus $\{u, x\} \in \mathscr{D}$ if and only if $x = \mathscr{H}(u) + \Phi\alpha$ for some $\alpha \in \mathbb{C}^n$ such that $\|Q_\Omega(\beta) + H_\Omega(u) - \gamma\| = \text{Minimum } \{\|Q_\Omega(\beta) + H_\Omega(u) - \gamma\| : \beta \in \mathbb{C}^n\}$, or equivalently α is a least-squares solution of the matrix equation (for u fixed) $Q_\Omega(\beta) = \gamma - H_\Omega(u)$. Since Range Q_Ω is closed, the set of all such α is given by the coset

$$Q_\Omega^\#(\gamma - H_\Omega(u)) \dot{+} \text{Null } Q_\Omega.$$

This proves [I]. The first part of [II] follows from [I]. Clearly Dom $\mathscr{D} = X_m$. Since Null Q_Ω is finite-dimensional and the map $u \to \mathscr{H}(u) - \Phi Q_\Omega^\# H_\Omega(u)$ defines a bounded linear operator from X_m into X_n, we see that $\mathscr{S}$ is closed in $X_m \oplus X_n$. □

Remark. Let $\{u, x\} \in \mathscr{D}$ be as (I.ii) of the above lemma. Then

$$x(t_0) = \alpha - Q_\Omega^\# H_\Omega(u) + Q_\Omega^\# \gamma.$$

Thus the initial state of x depends on the control u. But if $H_\Omega(u) = 0$ (which is the case, for example, if $F_\Omega(x) = x(t_0)$), then $x(t_0)$ does not depend on u.

Remark. Minamide and Nakamura [10] considered an optimal control problem whose state is somewhat similar to ours in appearance: Minimize $\mathscr{C}$ (when $F_\Lambda \equiv 0$) over $u \in X_m$ such that

(1) $\dot{x} = Ax + Bu$ almost everywhere,

(2) $x(t_0) = \gamma$,

(3) $\|x(t_1) - x_0\| \leqq \|y(t_1) - x_0\|$ for *all* $v \in X_m$ and y such that $\dot{y} = Ay + Bv$, $y(t_0) = \gamma$.

Thus if we define an operator $\hat{\mathscr{H}} : X_m \to \mathbb{C}^n$ by $\hat{\mathscr{H}}(v) = (\mathscr{H}(v)(t_1)$, then this problem becomes: Minimize $\mathscr{C}$ over $u \in X_m$ such that u is a least-squares solution of $\hat{\mathscr{H}}(v) = x_0 - \Phi(t_1)\gamma$ and $\dot{x} = Ax + Bu$, $x(t_0) = \gamma$. This means that the state is subject to the usual initial condition (and hence generates a unique response), but the control space is restricted.

Thus there is no direct connection between the problems in [10] and ours. The condition (2) in the above corresponds to (3.4). Notice that the control space in [10] is the coset $\hat{\mathscr{H}}^\#(x_0 - \Phi(t_1)\gamma) \dot{+} \text{Null } \hat{\mathscr{H}}$ while our control space is the whole space X_m.

In [11] Minamide and Nakamura considered a problem which is somewhat similar in spirit to ours:

Find $u^+ \in X_m$ minimizing $\sum_{j=1}^m \int_{t_0}^{t_1} |u_j(t)|\, dt$, subject to $\dot{x} = Ax + Bu$, $x(t_0) = x_0$ and $\|x(t_1) - x_1\| < \varepsilon$, where x_0, x_1 are given constant vectors and $\varepsilon > 0$.

In [10], the closed forms of generalized inverses of (single-valued) operators are used to derive the differential equations for an optimal controller. This idea cannot be used in this paper as the "closed" forms of generalized inverses of multi-valued operators are not known in the literature. Instead, we use an abstract result for least-squares solutions to derive equations for the optimal controller for our problem.

In the following we will describe $\mathscr{S}$ and $\mathscr{D}$ in terms of equality constraints.

PROPOSITION 3.2.

(i) *Let $\mathscr{S}$ be as in the above lemma. Then $\{u, y\} \in \mathscr{S}$ if and only if $y \in \text{Dom } T_1$, $u \in X_m$ such that*

$$\dot{y} = Ay + Bu \quad \textit{almost everywhere},$$

$$F_\Omega(y) = (I - Q_\Omega Q_\Omega^\#) H_\Omega(u).$$

(ii) *$\{u, x\} \in \mathscr{D}$ if and only if $u \in X_m$, $x \in \text{Dom } T_1$ such that*

$$\dot{x} = Ax + Bu \quad \textit{almost everywhere},$$

$$F_\Omega(x) - \gamma = (Q_\Omega Q_\Omega^\# - I)(\gamma - H_\Omega(u)).$$

Proof. (i) The "only if" part is easy to check. Assume now that $y \in \text{Dom } T_1$ and $u \in X_m$ satisfy the above two equations. Write $y = \mathscr{H}(u) + \Phi\alpha$, $\alpha \in \mathbb{C}^n$. Then the second condition becomes

$$0 = Q_\Omega(\alpha + Q_\Omega^\# H_\Omega(u)).$$

Thus $\alpha = -Q_\Omega^\# Q_\Omega Q_\Omega^\# H_\Omega(u) + k = -Q_\Omega^\# H_\Omega(u) + k$ for some $k \in \text{Null } Q_\Omega$. Hence $y = \mathscr{H}(u) + \Phi[k - Q_\Omega^\# H_\Omega(u)]$, and so $\{u, y\} \in \mathscr{S}$. We now prove (ii). Take any $\{u, x\} \in \mathscr{D}$. Then by Lemma 3.1, $\{u, x\} = \{u, \Phi Q_\Omega^\# \gamma + y\}$ for some y such that $\{u, y\} \in \mathscr{S}$. Then using the definition of Q_Ω and part (i) above, $\dot{x} = Ax + Bu$ almost everywhere and $F_\Omega(x) - \gamma = (Q_\Omega Q_\Omega^\# - I)(\gamma - H_\Omega(u))$. This argument can be traced back. □

Remark. By the above proposition we can say that for $\{u, x\} \in \mathscr{D}$, the response x "misses" the target γ by $(Q_\Omega Q_\Omega^\# - I)(\gamma - H_\Omega(u))$.

Remark. Assume that $n = m$, $B = W = U = I_{n\times n}$ and Rank $Q_\Omega = d$. Then our problem can be stated equivalently (by taking $u^+ = T_1 x^+$ when x^+ is a minimizer):

Minimize $(\|x\|^2 + \|T_1 x\|^2)$, subject to $x \in \text{Dom } T_1 \cap F_\Omega^{-1}(\gamma)$.

Clearly $F_\Omega^{-1}(\gamma)$ is a closed convex set. Thus this particular case is a special case of the abstract problem considered in [1, Thm. 4.4]. However, the vector ξ appearing in this theorem vanishes in our case, and so his theorem does not give any new information to the above case. We thank J. Burns for drawing our attention to reference [1].

Since $Q_\Omega Q_\Omega^\#$ is the orthogonal projector on $\mathbb{C}^n$ onto Range Q_Ω, from Lemma 3.1 and Proposition 3.2 we have the following.

COROLLARY 3.3.

(i) *For each $u \in X_m$, u has a unique response if and only if* Dim Null $Q_\Omega = 0$.

(ii) *If* Rank $Q_\Omega = d$ *or* $\gamma - H_\Omega(u) \in \text{Range } Q_\Omega$ *for all* $u \in X_m$, *then* $\mathscr{D} = \{\{u, x\}: u \in X_m, x \in \text{Dom } T_1, T_1 x = Bu, F_\Omega(x) = \gamma\}$.

The following notation will be needed later. If x, y are $n \times p$ and $n \times q$ matrix-valued functions of t, then

$$\langle x, y \rangle := \int_{t_0}^{t_1} x^t(t) \bar{y}(t)\, dt, \qquad (p \times q).$$

In the following we find a necessary and sufficient condition for F_Ω to be a generalized two-point operator.

LEMMA 3.4. *Let M, N be $d \times n$ constant matrices and let f be a $d \times n$ matrix-valued function of t such that its rows are square-integrable over $[t_0, t_1]$. Then*:

(i) $$F_\Omega(x) = Mx(t_0) + Nx(t_1) + \int_{t_0}^{t_1} f(t)x(t)\,dt \quad \textit{for all } x \in \mathrm{Dom}\ T_1$$

if and only if $\Omega_2 \in \mathrm{Dom}\ T_1$ columnwise, $f^* + \dot{\Omega}_2 = \Omega_1$ almost everywhere, and

$$M = -\Omega_2^*(t_0), \qquad N = \Omega_2^*(t_1).$$

(ii) *If F_Ω has the form as the above, then*

$$Q_\Omega = \int_{t_0}^{t_1} f\Phi\,dt + M + N\Phi(t_1),$$

$$H_\Omega(u) = N^*(\mathscr{H}(u))(t_1) + \int_{t_0}^{t_1} f\mathscr{H}(u)\,dt, \qquad u \in X_m,$$

$$H_\Omega^*(\beta) = B^*(t)(\Phi^*)^{-1}(t)\left(\Phi^*(t_1)N^* + \int_t^{t_0} (\Phi^* f^*)(s)\,ds\right)\beta, \qquad \beta \in \mathbb{C}^d.$$

Proof. (i) Assume that F_Ω has the form as in the theorem. Then for all $x \in \mathrm{Dom}\ T_1$ with $x(t_0) = x(t_1) = 0$,

(*) $$\int_{t_0}^{t_1} (\Omega_1^* - f)x\,dt + \int_{t_0}^{t_1} \Omega_2^* \dot{x}\,dt = 0.$$

We now use the idea first used in [4]. For the above x, put $\dot{x} = g$. Then $g \in X_n$ and

(**) $$x(t) = \int_{t_0}^{t} g(s)\,ds.$$

Since $x(t_1) = 0$, this implies that g is orthogonal to $\mathbb{C}^n$ (after we have identified $\mathbb{C}^n$ as a subspace of X_n in the obvious manner). Conversely, if $g \in X_n$ and is orthogonal to $\mathbb{C}^n$, then the x defined by (**) belongs to Dom T_1 with $x(t_0) = x(t_1) = 0$. Now take any $g \in X_n$ orthogonal to $\mathbb{C}^n$ and define x as (**). Then (*) can be written (after interchanging the order of the integration) as

(*)′ $$\int_{t_0}^{t_1} \left(\int_t^{t_1} (\Omega_1^* - f)(s)\,ds + \Omega_2^*(t) \right) g(t)\,dt = 0.$$

Therefore

$$\int_t^{t_1} (\Omega_1^* - f)(s)\,ds + \Omega_2^*(t) = k \quad \text{a.e. } t \in [t_0, t_1]$$

for some $k \in \mathbb{C}^n$. Redefining Ω_2 in the set of measure zero if necessary, we see that the above is true for *all* t and hence Ω_2 is absolutely continuous on $[t_0, t_1]$. Differentiating the equality, we see that

$$\Omega_1^*(t) - f(t) = \dot{\Omega}_2^*(t), \quad \text{a.e. } t \in [t_0, t_1].$$

Returning to the representation of F_Ω, we have

$$\begin{aligned} Mx(t_0) + Nx(t_1) + \int_{t_0}^{t_1} f(t)x(t)\,dt &= \int_{t_0}^{t_1} [(f + \dot{\Omega}_2^*)x + \Omega_2^* \dot{x}]\,dx \\ &= \Omega_2^*(t_1)x(t_1) - \Omega_2^*(t_0)x(t_0) + \int_{t_0}^{t_1} f(t)x(t)\,dt \end{aligned}$$

for all $x \in \mathrm{Dom}\ T_1$. Thus $M = \Omega_2^*(t_0)$, $N = \Omega_2^*(t_1)$.

The above argument can be traced back. This proves (i). We now prove (ii). The first two parts follow directly from the definitions of Q_Ω and H_Ω together with part (i). The second part is easy to check as the general H_Ω^* is given by

$$H_\Omega^*(\beta) = B^*\Omega_2\beta + \mathscr{H}^*((\Omega_1 + A^*\Omega_2)\beta) \quad \text{all } \beta \in \mathbb{C}^d,$$

$$\mathscr{H}^*(x) = B^*(t)\Phi^{*-1}(t)\int_t^{t_1} \Phi^*(s)x(s)\,ds \quad \text{for all } x \in X_m.$$

This together with (i) gives the description of H_Ω^* as claimed. □

We will now describe $\mathscr{C}(u, x)$ in terms of the norms in X_m, X_n and $\mathbb{C}^{d_1}$. To do this it is convenient to introduce the following:

$$Q_\Lambda := \int_{t_0}^{t_1} ((\Lambda_1^* + \Lambda_2^* A)\Phi)(t)\,dt \qquad (d_1 \times n), \tag{3.8}$$

$$H_\Lambda(u) := \int_{t_0}^{t_1} (\Lambda_1^*\mathscr{H}(u) + \Lambda_2^*(A\mathscr{H}(u) + Bu))(t)\,dt, \qquad u \in X_m, \tag{3.9}$$

where $\mathscr{H}$ is defined in (3.5).

For each $t \in [t_0, t_1]$, let $U^{1/2}(t)$ denote the positive square root of $U(t)$, and $W^{1/2}(t)$ the nonnegative square root of $W(t)$.

The following will be needed later. We state it without a proof because it is easy to check.

LEMMA 3.5. *For $\{u, x\} \in \mathscr{D}$, write $x = \mathscr{H}(u) + \Phi q$, where*

$$q = \alpha - Q_\Omega^\# H_\Omega(u) + Q_\Omega^\# \gamma, \qquad \alpha \in \text{Null } Q_\Omega.$$

Then

$$\int_{t_0}^{t_1} (\Lambda_1^* x + \Lambda_2^* \dot{x})(t)\,dt = Q_\Lambda q + H_\Lambda(u),$$

$$\mathscr{C}(u, x) = \|\{U^{1/2}u,\ W^{1/2}(\mathscr{H}(u) + \Phi q - x_0),\ H_\Lambda(u) + Q_\Lambda q\}\|^2,$$

where $\|\{\cdot, \cdot, \cdot\}\|$ denotes the standard norm of $X_m \oplus X_n \oplus \mathbb{C}^{d_1}$.

We will now show that the control problem can be written as a least-squares problem for a multi-valued closed operator equation. We need the following definitions.

DEFINITION. ζ is the vector in $X_m \oplus X_n \oplus \mathbb{C}^{d_1}$ defined by

$$\zeta = \{0,\ W^{1/2}(x_0 - \Phi Q_\Omega^\# \gamma), -Q_\Lambda Q_\Omega^\# \gamma\}.$$

DEFINITION. $\mathscr{M}$ is a vector subspace of $X_m \oplus (X_m \oplus X_n \oplus \mathbb{C}^{d_1})$ defined by $\{u, g\} \in \mathscr{M}$ if and only if $u \in X_m$ and

$$g = \{U^{1/2}u,\ W^{1/2}(\mathscr{H}(u) + \Phi(\alpha - Q_\Omega^\# H_\Omega(u))),\ H_\Lambda(u) + Q_\Lambda(\alpha - Q_\Omega^\# H_\Omega(u))\},$$

for some $\alpha \in \text{Null } Q_\Omega$.

Since $U^{1/2}$ is invertible on $[t_0, t_1]$ and $\mathscr{H}$, H_Ω and H_Λ are bounded linear operators, we see easily that $\mathscr{M}$, Range $\mathscr{M}$ are closed and Null $\mathscr{M} = \{0\}$.

In the following theorem our control problem is changed to a least-squares problem.

THEOREM 3.6. *Let $u^+ \in X_m$. Then u^+ is an optimal controller if and only if u^+ is a least-squares solution of $\zeta \in \mathscr{M}(u)$.*

Proof. Note first that $\{u^+, x^+\}$ minimizes $\mathscr{C}$ subject to (3.2)-(3.4) if and only if it minimizes $\mathscr{C}$ over $\mathscr{D}$. Let $\{u^+, x^+\}$ be a minimizer of $\mathscr{C}$ and write it as

$$x^+ = \mathscr{H}(u^+) + \Phi q^+ \tag{1}$$

where

$$q^+ = \alpha^+ - Q_\Omega^\# H_\Omega(u^+) + Q_\Omega^\# \gamma, \qquad \alpha^+ \in \text{Null } Q_\Omega.$$

Take any $\{u, x\} \in \mathscr{D}$ and write it also as

$$x = \mathscr{H}(u) + \Phi q \tag{2}$$

where

$$q = \alpha - Q_\Omega^\# H_\Omega(u) + Q_\Omega^\# \gamma, \qquad \alpha \in \text{Null } Q_\Omega.$$

Then using Lemma 3.5, $\{u^+, x^+\}$ is a minimizer of $\mathscr{C}$ over $\mathscr{D}$ if and only if

$$\begin{aligned} \mathscr{C}(u^+, x^+) = \|\{U^{1/2}u^+,\ & W^{1/2}(\mathscr{H}(u^+) + \Phi\alpha^+ - \Phi Q_\Omega^\# H_\Omega(u^+)), \\ & H_\Lambda(u^+) + Q_\Lambda \alpha^+ - Q_\Lambda Q_\Omega^\# H_\Omega(u^+)\} - \zeta\|^2 \\ \leqq \mathscr{C}(u, x) = \|\{U^{1/2}u,\ & W^{1/2}(\mathscr{H}(u) + \Phi\alpha - \Phi Q_\Omega^\# H_\Omega(u)), \\ & H_\Lambda(u) + Q_\Lambda \alpha - Q_\Lambda Q_\Omega^\# H_\Omega(u)\} - \zeta\|^2 \end{aligned} \tag{3}$$

for all $u \in X_m$, $\alpha \in \text{Null } Q_\Omega$. Let us put

$$\begin{aligned} y^+ = \{U^{1/2}u^+,\ & W^{1/2}(\mathscr{H}(u^+) + \Phi\alpha^+ - \Phi Q_\Omega^\# H_\Omega(u^+)), \\ & H_\Lambda(u^+) + Q_\Lambda \alpha^+ - Q_\Lambda Q_\Omega^\# H_\Omega(u^+)\}. \end{aligned} \tag{4}$$

Then $y^+ \in \mathscr{M}(u^+)$, and so (3) becomes:

$$\|y^+ - \zeta\| = \text{Min}\,\{\|y - \zeta\| : y \in \text{Range } \mathscr{M}\}. \tag{5}$$

Thus if u^+ is an optimal controller for the control problem, then by definition there exists x^+ such that $\{u^+, x^+\} \in \mathscr{D}$ which minimizes $\mathscr{C}$ over $\mathscr{D}$. But then by (5), u^+ is a least-squares solution of $\zeta \in \mathscr{M}(u)$. Conversely, let u^+ be a least-squares solution of $\zeta \in \mathscr{M}(u)$. Then there exists $y^+ \in \mathscr{M}(u^+)$ satisfying (5). We now write y^+ as in (4) for some $\alpha^+ \in \text{Null } Q_\Omega$.

Define x^+ as in (1) and x as in (2) for some $u \in X_m$ and $\alpha \in \text{Null } Q_\Omega$. Then it follows from (5) that

$$\mathscr{C}(u^+, x^+) \leqq \mathscr{C}(u, x)$$

for all $\{u, x\} \in \mathscr{D}$. Thus $\{u^+, x^+\}$ minimizes $\mathscr{C}$ over $\mathscr{D}$, and hence u^+ is an optimal controller. □

Having identified an optimal controller as a least-squares solution, a general theorem, Theorem 2.2, is applicable.

THEOREM 3.7. *There always exists a unique optimal controller for the control problem. This is given by* $u^+ := \mathscr{M}^*(\zeta)$. *Moreover, the set of the responses corresponding to* u^+ *is given by*

$$\{\mathscr{H}(u^+) + \Phi(\alpha - Q_\Omega^\# H_\Omega(u^+) + Q_\Omega^\# \gamma) : \alpha \in \text{Null } Q_\Omega\}.$$

Proof. It was noted earlier that Null $\mathscr{M} = \{0\}$, and $\mathscr{M}$, Range $\mathscr{M}$ are closed. Thus the first part of the theorem follows from Theorem 3.6 together with (1), (3) of Theorem 2.2. By definition, x is a response corresponding to u^+ if and only if $\{u^+, x\} \in \mathscr{D}$. Thus the second part follows from (I.ii) of Lemma 3.1. □

A response x corresponding to $\mathscr{M}^*(\zeta)$ may not be an optimal response, i.e., $\{\mathscr{M}^*(\zeta), x\}$ may not minimize $\mathscr{C}$ over $\mathscr{D}$. Recall that when this element minimizes $\mathscr{C}$, x is called an optimal response (corresponding $\mathscr{M}^*(\zeta)$). Later we will characterize all

optimal responses. To do this, we introduce a linear operator R: Null $Q_\Omega \to X_m \oplus X_n \oplus \mathbb{C}^{d_1}$, and a function $p: X_m \to X_m \oplus X_n \oplus \mathbb{C}^{d_1}$ by

$$R(\beta) = \{0,\ W^{1/2}\Phi\beta,\ Q_\Lambda \beta\}, \qquad \beta \in \text{Null } Q_\Omega,$$

$$p(u) = \zeta - \{U^{1/2}u,\ W^{1/2}(\mathscr{H}(u) - \Phi Q_\Omega^{\#} H_\Omega(u)),\ H_\Lambda(u) - Q_\Lambda Q_\Omega^{\#} H_\Omega(u)\}.$$

Evidently, R is a bounded linear operator with closed range, and so Dom $R^{\#} = X_m \oplus X_n \oplus \mathbb{C}^{d_1}$.

First we have the following lemma.

LEMMA 3.8. $\|(RR^{\#} - I)(p(\mathscr{M}^{\#}(\zeta)))\| \leqq \|R(\beta) - p(u)\|$ *for all* $\beta \in$ Null Q_Ω, $u \in X_m$.

Proof. Put $u^+ = \mathscr{M}^{\#}(\zeta)$ and let x^+ be an optimal response corresponding to u^+. Then since $\{u^+, x^+\} \in \mathscr{D}$,

$$(1) \qquad x^+ = \mathscr{H}(u^+) + \Phi[\alpha^+ - Q_\Omega^{\#} H_\Omega(u^+) + Q_\Omega^{\#}\gamma]$$

for some $\alpha^+ \in$ Null Q_Ω, and since $\{u^+, x^+\}$ is a minimizer of $\mathscr{C}$,

$$(2) \qquad \mathscr{C}(u^+, x^+) \leqq \mathscr{C}(u, x) \quad \text{for all } \{u, x\} \in \mathscr{D}.$$

Take any $\{u, x\} \in \mathscr{D}$ and write x as

$$(3) \qquad x = \mathscr{H}(u) + \Phi[\alpha - Q_\Omega^{\#} H_\Omega(u^+) + Q_\Omega^{\#}\gamma]$$

for some $\alpha \in$ Null Q_Ω. Then using the definitions of R and p and Lemma 3.5 we have

$$(4) \qquad \mathscr{C}(u, x) = \|R(\alpha) - p(u)\|.$$

Thus (2) can be written equivalently as

$$(5) \qquad \|R(\alpha^+) - p(u^+)\| \leqq \|R(\alpha) - p(u)\|$$

for all $\alpha \in$ Null Q_Ω and $u \in X_m$.

This is true, in particular, for all $\alpha \in$ Null Q_Ω and $u = u^+$. Thus it follows from (4) that α^+ is a least-squares solution of $R(\beta) = p(u^+)$, and hence $\alpha^+ = R^{\#}(p(u^+)) + k$ for some $k \in$ Null R. Returning to (4), we have

$$\|R[R^{\#}(p(u^+)) + k] - p(u^+)\| = \|(RR^{\#} - I)(p(u^+))\| \leqq \|R(\alpha) - p(u)\|$$

for all $\alpha \in$ Null Q_Ω, $u \in X_m$. □

In the following theorem we find all optimal responses.

THEOREM 3.9. x^+ *is an optimal response corresponding to the optimal controller* $u^+ = \mathscr{M}^{\#}(\zeta)$ *if and only if*

$$x^+ = \mathscr{H}(u^+) + \Phi(\alpha^+ - Q_\Omega^{\#} H_\Omega(u^+) + Q_\Omega^{\#}\gamma)$$

for some $\alpha^+ \in$ Null Q_Ω *belonging to the coset*

$$R^{\#}(p(u^+)) \dotplus \text{Null } R,$$

or equivalently, for some $\alpha^+ \in$ Null Q_Ω *which is a least-squares solution of* $R(\beta) = p(u^+)$.

In particular, an optimal response is unique if and only if Dim Null $R = 0$.

Proof. Let x^+ be an optimal response corresponding to u^+. Then $\{u^+, x^+\}$ is a minimizer of $\mathscr{C}$ over $\mathscr{D}$. Thus

$$(1) \qquad x^+ = \mathscr{H}(u^+) + \Phi[\alpha^+ - Q_\Omega^{\#} H_\Omega(u^+) + Q_\Omega^{\#}\gamma]$$

for some $\alpha^+ \in$ Null Q_Ω, and

$$(2) \qquad \mathscr{C}(u^+, x^+) \leqq \mathscr{C}(u, x) \quad \text{for all } \{u, x\} \in \mathscr{D}.$$

Arguing similarly as in the proof of the above lemma, we see that α^+ must belong to the coset in the theorem. Suppose now that x^+ is defined as in (1) for some $\alpha^+ \in \text{Null } Q_\Omega$ such that

$$\alpha^+ = R^\#(p(u^+)) + k, \qquad k \in \text{Null } R. \tag{3}$$

Then by (4) of the above lemma,

$$\mathscr{C}(u^+, x^+) = \|R(\alpha^+) - p(u^+)\| = \|(RR^\# - I)(p(u^+))\|. \tag{4}$$

Take any $\{u, x\} \in \mathscr{D}$ and write x as

$$x = \mathscr{H}(u) + \Phi[\alpha - Q_\Omega^\# H_\Omega(u) + Q_\Omega^\# \gamma]$$

for some $\alpha \in \text{Null } Q_\Omega$. Then using (4) of the Lemma 3.8 again,

$$\mathscr{C}(u, x) = \|R(\alpha) - p(u)\|. \tag{5}$$

Thus by Lemma 3.8,

$$\mathscr{C}(u^+, x^+) = \|(RR^\# - I)(p(u^+))\| \leqq \|R(\alpha) - p(u)\| = \mathscr{C}(u, x).$$

Thus x^+ defined as (1) for some $\alpha^+ \in \text{Null } Q_\Omega$ defined as (3) is an optimal response. The second part of the theorem is clear. □

We have identified the optimal controller to the control problem as $\mathscr{M}^\#(\zeta)$. In a practical point of view this is not satisfactory as the "closed" form of $T^\#$ for a general multi-valued operator T is not known in the literature. Thus in the sequel of this paper we will characterize the optimal controller by integro-differential equations. The following is (ii) of Theorem 2.2 adopted to the present situation.

LEMMA 3.10. *u^+ is the optimal controller to the control problem if and only if*

$$\zeta \in \mathscr{M}(u^+) \dotplus \text{Null } \mathscr{M}^*.$$

We will replace the above inclusion by integro-differential equations. To do this, first we describe $\mathscr{M}^*$ which is a subspace of $(X_m \oplus X_n \oplus \mathbb{C}^{d_1}) \oplus X_m$.

LEMMA 3.11. *$\{\{v, f, \beta\}, u\} \in \mathscr{M}^*$ if and only if*

(i) $u \in X_m$, $\{v, f, \beta\} \in X_m \oplus X_n \oplus \mathbb{C}^{d_1}$,

(ii) $\overline{\langle W^{1/2}\Phi, f\rangle} + Q_\Lambda^* \beta \in (\text{Null } Q_\Omega)^\perp$,

(iii) $u = (U^{1/2})^* v + B^*(\Lambda_2 \beta - \Omega_2 l) + \mathscr{H}^*\big((W^{1/2})^* f + (\Lambda_1 + A^*\Lambda_2)\beta - (\Omega_1 + A^*\Omega_2) l\big)$

where

$$l := (Q_\Omega^\#)^*\big(Q_\Lambda^* \beta + \overline{\langle W^{1/2}\Phi, f\rangle}\big).$$

Proof. Take any $\{\{v, f, \beta\}, u\} \in \mathscr{M}^*$. Then for any $\{u_1, \{v_1, f_1, \beta_1\}\} \in \mathscr{M}$,

$$-\langle u, u_1\rangle + \langle v, v_1\rangle + \langle f, f_1\rangle + \langle \beta, \beta_1\rangle = 0. \tag{1}$$

By the definition $\mathscr{M}$,

$$v_1 = U^{1/2} u_1, \qquad f_1 = W^{1/2}[\mathscr{H}(u_1) + \Phi(\alpha - Q_\Omega^\# H_\Omega(u_1))],$$

$$\beta_1 = H_\Lambda(u_1) + Q_\Lambda[\alpha - Q_\Omega^\# H_\Omega(u_1)]$$

for $u_1 \in X_m$, $\alpha \in \text{Null } Q_\Omega$. Taking $u_1 = 0$ in (1), we see that

$$\langle f, W^{1/2}\Phi\alpha\rangle + \langle \beta, Q_\Lambda \alpha\rangle = 0 \tag{2}$$

for all $\alpha \in \text{Null } Q_\Omega$. Thus

$$\langle W^{1/2}\Phi, f\rangle + Q_\Lambda^* \beta \in (\text{Null } Q_\Omega)^\perp. \tag{3}$$

Returning to (1), we see that

$$(4)\qquad -\langle u, u_1\rangle+\langle v, U^{1/2}u_1\rangle+\langle f, W^{1/2}[\mathscr{H}(u_1)-\Phi Q_\Omega^{\#}H_\Omega(u_1)]\rangle$$
$$+\langle\beta, H_\Lambda(u_1)-Q_\Lambda Q_\Omega^{\#}H_\Omega(u_1)\rangle=0 \quad \text{for all } u_1\in X_m.$$

This then yields that

$$(5)\qquad u=(U^{1/2})^*v+\mathscr{H}^*((W^{1/2})^*f)-H_\Omega^*((Q_\Omega^{\#})^*\overline{\langle W^{1/2}\Phi, f\rangle})$$
$$+H_\Lambda^*(\beta)-H_\Omega^*((Q_\Lambda Q_\Omega^{\#})^*\beta).$$

Thus v, f, β, u satisfy (3) and (5). Conversely if these satisfy (3) and (5), then we see easily that (1) holds for all $\{u_1, \{v_1, f_1, \beta_1\}\}\in\mathscr{M}$. Now as was given in the proof of Lemma 3.4,

$$(6)\qquad H_\Omega^*(\delta)=B^*\Omega_2\delta+\mathscr{H}^*[(\Omega_1+A^*\Omega_2)\delta], \qquad \delta\in\mathbb{C}^d,$$

$$(7)\qquad H_\Lambda^*(\delta)=B^*\Lambda_2\delta+\mathscr{H}^*[(\Lambda_1+A^*\Lambda_2)\delta], \qquad \delta\in\mathbb{C}^{d_1}.$$

Thus since $\{\{v, f, \beta\}, u\}\in\mathscr{M}^*$ if and only if (3) and (5) hold, we see that using (6), (7) the result follows immediately. □

The following is a main result of this paper. Recall that $Q_\Omega(d\times n)$, $Q_\Lambda(d_1\times n)$ are defined in (3.7), (3.8), respectively, while $H_\Omega: X_m\to\mathbb{C}^d$, $H_\Lambda: X_m\to\mathbb{C}^{d_1}$ are defined in (3.6), (3.9), respectively.

THEOREM 3.12. *$u^+\in X_m$ is the optimal controller for the control problem if and only if*

$$u^+=U^{-1}B^*\eta \quad \textit{almost everywhere}$$

where η is an element of X_m for which there exists $z\in$ Dom T_1 such that

(i) $\dot z=Az+BU^{-1}B^*\eta$ *almost everywhere,*
(ii) $Q_\Omega(z(t_0)+Q_\Omega^{\#}H_\Omega(U^{-1}B^*\eta)-Q_\Omega^{\#}\gamma)=0$,
(iii) $\int_{t_0}^{t_1}\Phi^*W(z-x_0)\,dt+Q_\Lambda^*(H_\Lambda(U^{-1}B^*\eta)+Q_\Lambda z(t_0))\in$ Range (Q_Ω^*),
(iv) $\eta-\delta\in$ Dom T_1,
(v) $(\eta-\delta)(t_1)=0$,
(vi) $d(\eta-\delta)/dt=-A^*(\eta-\delta)+\psi$ *almost everywhere.*

Here δ and ψ are vector-valued functions of t (depending on η and z) defined by:

$$\delta=\Omega_2(Q_\Omega^{\#})^*\Big(Q_\Lambda^*H_\Lambda(U^{-1}B^*\eta)+Q_\Lambda^*Q_\Lambda z(t_0)+\int_{t_0}^{t_1}\Phi^*W(z-z_0)\,dt\Big)$$
$$-\Lambda_2(H_\Lambda(U^{-1}B^*\eta)+Q_\Lambda z(t_0)),$$
$$\psi=W(z-x_0)+(\Lambda_1+A^*\Lambda_2)(H_\Lambda(U^{-1}B^*\eta)+Q_\Lambda z(t_0))$$
$$-(\Omega_1+A^*\Omega_2)(Q_\Omega^{\#})^*\Big(Q_\Lambda^*H_\Lambda(U^{-1}B^*\eta)+Q_\Lambda^*Q_\Lambda z(t_0)+\int_{t_0}^{t_1}\Phi^*W(z-x_0)\,dt\Big).$$

Proof. By Lemma 3.10, u is an optimal controller if and only if $u\in X_m$ and

$$(1)\qquad \zeta\in\mathscr{M}(u)\dot+\text{Null }\mathscr{M}^*.$$

Equivalently, there exists $g\in\mathscr{M}(u)$ such that

$$(2)\qquad g-\zeta\in\text{Null }\mathscr{M}^*.$$

Using the definition of ζ and $\mathscr{M}$, we can write

$$(3)\qquad g-\zeta=\{U^{1/2}u, W^{1/2}[\mathscr{H}(u)+\Phi(\alpha-Q_\Omega^{\#}H_\Omega(u)+Q_\Omega^{\#}\gamma)-x_0],$$
$$H_\Lambda(u)+Q_\Lambda[\alpha-Q_\Omega^{\#}H_\Omega(u)+Q_\Omega^{\#}\gamma]\},$$

for some $\alpha \in \text{Null } Q_\Omega$. Let

$$q = \alpha - Q_\Omega^* H_\Omega(u) + Q_\Omega^* \gamma, \tag{4}$$

$$z = \mathscr{H}(u) + \Phi q. \tag{5}$$

Then it is clear that $z \in \text{Dom } T_1$ and is the unique solution such that

(i) $\dot{z} = Az + Bu$ almost everywhere

and

$$z(t_0) = q. \tag{6}$$

Using (4), this initial condition is equivalent to

(ii) $Q_\Omega[z(t_0) + Q_\Omega^* H_\Omega(u) - Q_\Omega^* \gamma] = 0.$

Using (3), (4), (5) we can rewrite (2) as

$$\{U^{1/2}u,\ W^{1/2}(z - x_0),\ H_\Lambda(u) + Q_\Lambda q\} \in \text{Null } \mathscr{M}^*. \tag{7}$$

Since $(W^{1/2})^* W^{1/2} = W$, $(U^{1/2})^* U^{1/2} = U$, $z(t_0) = q$, by Lemma 3.11 the above inclusion is equivalent to (iii) and (8) below:

(iii) $\overline{\langle W\Phi,\ z - x_0\rangle} + Q_\Lambda^*(H_\Lambda(u) + Q_\Lambda z(t_0)) \in \text{Range } (Q_\Omega^*)$,

$$\begin{aligned} 0 = {} & Uu + B^*[\Lambda_2(H_\Lambda(u) + Q_\Lambda z(t_0)) - \Omega_2 l] \\ & + \mathscr{H}^*[W(z - x_0) + (\Lambda_1 + A^*\Lambda_2)(H_\Lambda(u) + Q_\Lambda z(t_0)) - (\Omega_1 + A^*\Omega_2) l], \end{aligned} \tag{8}$$

where

$$l = (Q_\Omega^*)^*[Q_\Lambda^*(H_\Lambda(u) + Q_\Lambda z(t_0)) + \overline{\langle W^{1/2}\Phi,\ W^{1/2}(z - x_0)\rangle}].$$

Let

$$\tilde{\delta} = \Omega_2 l - \Lambda_2(H_\Lambda(u) + Q_\Lambda z(t_0)), \tag{9}$$

$$\tilde{\psi} = W(z - x_0) + (\Lambda_1 + A^*\Lambda_2)(H_\Lambda(u) + Q_\Lambda z(t_0)) - (\Omega_1 + A^*\Omega_2) l. \tag{10}$$

Then using the characterization of $\mathscr{H}^*$ given in the proof of Lemma 3.4, we can rewrite (8) equivalently as

$$0 = U(t)u(t) - B^*(t)\left[\tilde{\delta}(t) - \int_t^{t_1} \Phi^{*-1}(t)\Phi^*(s)\tilde{\psi}(s)\, ds\right] \tag{11}$$

for almost all $t \in [t_0, t_1]$. Let

$$\eta(t) := \tilde{\delta}(t) - \int_t^{t_1} \Phi^{*-1}(t)\Phi^*(s)\tilde{\psi}(s)\, ds, \qquad t_0 \leqq t \leqq t_1. \tag{12}$$

Then, since $(d/dt)\Phi^{*-1} = -A^*\Phi^{*-1}$, we can rewrite (12) equivalently as the three statements below:

$$\eta - \tilde{\delta} \in \text{Dom } T_1, \tag{13}$$

$$\frac{d}{dt}(\eta - \tilde{\delta}) = -A^*(\eta - \tilde{\delta}) + \tilde{\psi}, \quad \text{almost all } t, \tag{14}$$

$$(\eta - \bar{\delta})(t_1) = 0. \tag{15}$$

Now (11) can be written as

$$u(t) = U^{-1}(t)B^*(t)\eta(t), \quad \text{almost all } t. \tag{16}$$

Thus $\tilde{\delta} = \delta$, $\tilde{\psi} = \psi$ and the statements in (13), (14) and (15) are equivalent to the statements (iv), (vi) and (v) in the theorem. □

Remark. In the above theorem η may not even be continuous. Thus we *cannot* write the condition (v) as $\eta(t_1)=\zeta(t_1)$ as they are meaningless. But we can if the columns of Ω_2 and Λ_2 are in Dom T_1. In fact, $\eta\in$ Dom T_1 if and only if $\zeta\in$ Dom T_1.

Remark. There always exists $\eta\in X_m$, and $z\in$ Dom T_1 satisfying (i)-(vi) of Theorem 3.12 as an optimal controller for the control problem always exists.

If Ω_2 and Λ_2 are sufficiently smooth, then Theorem 3.12 can be simplified further.

COROLLARY 3.13. *Assume that the columns of* Ω_2 *and* Λ_2 *are in* Dom T_1. *Then the optimal controller* u^+ *for the control problem is given by*

$$u^+ = U^{-1}B^*\eta \quad \textit{almost everywhere,}$$

where η *is an element of* Dom T_1 *for which there exists an element* $z\in$ Dom T_1 *satisfying* (i), (ii), (iii) *of Theorem* 3.12 *in addition to*

$$\text{(v)}\quad \eta(t_1)=\Omega_2(t_1)(Q_\Omega^\#)^*\Big(Q_\Lambda^*H_\Lambda(U^{-1}B^*\eta)+Q_\Lambda^*Q_\Lambda z(t_0)+\int_{t_0}^{t_1}\Phi^*W(z-x_0)\,dt\Big)$$
$$-\Lambda_2(t_1)\big(H_\Lambda(U^{-1}B^*\eta)+Q_\Lambda z(t_0)\big),$$

$$\text{(vi)}\quad \dot\eta=-A^*\eta+W(z-x_0)+(\Lambda_1-\dot\Lambda_2)\big(H_\Lambda(U^{-1}B^*\eta)+Q_\Lambda z(t_0)\big)$$
$$+(\dot\Omega_2-\Omega_1)(Q_\Omega^\#)^*\Big(Q_\Lambda^*H_\Lambda(U^{-1}B^*\eta)+Q_\Lambda^*Q_\Lambda z(t_0)+\int_{t_0}^{t_1}\Phi^*W(z-x_0)\,dt\Big)$$

almost everywhere.

Proof. Because of the assumption on Ω_2 and Λ_2, the condition (iv) of Theorem 3.12 is equivalent to $\eta\in$ Dom T_1. Thus (v) and (vi) of Theorem 3.12 can be written as (v) and (vi) of this corollary. □

If F_Ω is the usual two-point evaluation operator, then Corollary 3.13 becomes

COROLLARY 3.14. *Assume that* $\Lambda_2\in$ Dom T_1 (*columnwise*) *and there exist constant matrices* $M(d\times n)$, $N(d\times n)$ *such that*

$$F_\Omega(x)=Mx(t_0)+Nx(t_1)\quad \textit{for all } x\in \text{Dom } T_1.$$

Let u^+ *be the unique optimal controller for the control problem. Then it is given by*

$$u^+ = U^{-1}B^*\eta \quad \textit{almost everywhere}$$

where η *is an element of* Dom T_1 *for which there exists an element* $z\in$ Dom T_1 *satisfying the following*:

(i) $\dot z = Az+BU^{-1}B^*\eta$ *almost everywhere*;

(ii) $(M+N\Phi(t_1))(M+N\Phi(t_1))^\#(\gamma-N(\mathcal{H}(U^{-1}B^*\eta))(t_1))$
$=(M+N\Phi(t_1))z(t_0)$;

(iii) $\displaystyle\int_{t_0}^{t_1}\Phi^*W(z-x_0)\,dt+Q_\Lambda^*\big(H_\Lambda(U^{-1}B^*\eta)+Q_\Lambda z(t_0)\big)$

belongs to Range $((M+N\Phi(t_1))^*)$;

(iv) $\eta(t_1)=N^*\big((M+N\Phi(t_1))^\#\big)^*$
$$\cdot\Big(Q_\Lambda^*H_\Lambda(U^{-1}B^*\eta)+Q_\Lambda^*Q_\Lambda z(t_0)+\int_{t_0}^{t_1}\Phi^*W(z-x_0)\,dt\Big)$$
$$-\Lambda_2(t_1)\big(H_\Lambda(U^{-1}B^*\eta)+Q_\Lambda z(t_0)\big);$$

(v) $\dot{\eta} = -A^*\eta + W(z - x_0) + (\Lambda_1 - \dot{\Lambda}_2)\big(H_\Lambda(U^{-1}B^*\eta) + Q_\Lambda z(t_0)\big)$

almost everywhere.

Proof. By Lemma 3.4, $\Omega_2 \in \text{Dom } T_1$ and $N^* = \Omega_2(t_1)$, $Q_\Omega = M + N\Phi(t_1)$, $\dot{\Omega}_2 = \Omega_1$, $H_\Omega(U^{-1}B^*\eta) = \mathcal{N}(\mathcal{H}(U^{-1}B^*\eta))(t_1)$. Thus the result follows from Corollary 3.14. □

If F_Λ is the usual two-point evaluation operator, then Corollary 3.13 (using Lemma 3.4) becomes.

COROLLARY 3.15. *Assume that* $\Omega_2 \in \text{Dom } T_1$ *and*

$$F_\Lambda = M_1 x(t_0) + N_1 x(t_1) \quad \textit{for all } x \in \text{Dom } T_1$$

for some constant matrices $M_1(d_1 \times n)$ *and* $N_1(d_1 \times n)$. *Let* u^+ *be the unique optimal controller for the control problem. Then it is given by*

$$u^+ = U^{-1}B^*\eta \quad \textit{almost everywhere,}$$

where η *is an element of* Dom T_1 *for which there exists an element* $z \in \text{Dom } T_1$ *satisfying the following*:

(i) $\dot{z} = Az + BU^{-1}B^*\eta$ *almost everywhere,*

(ii) $Q_\Omega\big(z(t_0) + Q_\Omega^* H_\Omega(U^{-1}B^*\eta) - Q_\Omega^*\gamma\big) = 0,$

(iii) $\displaystyle\int_{t_0}^{t_1} \Phi^* W(z - x_0)\, dt$

$$+ \big(M_1 + N_1\Phi(t_1)\big)^*\big(N_1(\mathcal{H}(U^{-1}B^*\eta))(t_1) + (M_1 + N_1\Phi(t_1))z(t_0)\big)$$

belongs to Range (Q_Ω^*),

(iv) $\eta(t_1) = -N_1^*\big(N_1(\mathcal{H}(U^{-1}B^*\eta))(t_1) + (M_1 + N_1\Phi(t_1))z(t_0)\big)$

$$+ \Omega_2(t_1)(Q_\Omega^*)^*\big((M_1 + N_1\Phi(t_1))^* N_1(\mathcal{H}(U^{-1}B^*\eta))(t_1)$$

$$+ (M_1 + N_1\Phi(t_1))^*(M_1 + N_1\Phi(t_1))z(t_0)\big),$$

(v) $\dot{\eta} = -A^*\eta + W(z - x_0) + (\dot{\Omega}_2 - \Omega_1)(Q_\Omega^*)^*$

$$\cdot \Big((M_1 + N_1\Phi(t_1))^*(\mathcal{H}(U^{-1}B^*\eta))(t_1)$$

$$+ (M_1 + N_1\Phi(t_1))^*(M_1 + N_1\Phi(t_1))z(t_0)$$

$$+ \int_{t_0}^{t_1} \Phi^* W(z - x_0)\, dt \Big) \quad \textit{almost everywhere.}$$

We have so far seen that if one of F_Ω and F_Λ is not a two-point evaluation operator, then the optimal controller is described by a Fredholm integro-differential equation. In the following, we will show that the optimal controller is described by a usual differential equation if F_Ω and F_Λ are point-evaluation operators.

COROLLARY 3.16. *Assume that*

$$F_\Omega(x) = Mx(t_0) + Nx(t_1),$$

$$F_\Lambda(x) = M_1 x(t_0) + N_1 x(t_1)$$

for all $x \in \text{Dom } T_1$, *where* $M(d \times n)$, $N(d \times n)$, $M_1(d_1 \times n)$, $N_1(d_1 \times n)$ *are constant matrices. Let* u^+ *be the optimal controller. Then it is described by*

$$u^+ = U^{-1}B^*\eta \quad \textit{almost everywhere,}$$

where $\eta \in \text{Dom } T_1$ *is such that there exists* $z \in \text{Dom } T_1$ *satisfying the following*:

(i) $\dot{z} = Az + BU^{-1}B^*\eta$ *almost everywhere,*

(ii) $(M + N\Phi(t_1))(M + N\Phi(t_1))^*\big(\gamma - N(\mathscr{H}(U^{-1}B^*\eta))(t_1)\big)$
$= (M + N\Phi(t_1))z(t_0),$

(iii) $\int_{t_0}^{t_1} \Phi^* W(z - x_0)\, dt + \big(M_1 + N_1\Phi(t_1)\big)^*\big(N_1(\mathscr{H}(U^{-1}B^*\eta))(t_1)$
$+ (M_1 + N_1\Phi(t_1))z(t_0)\big) \in \text{Range } ((M + N\Phi(t_1))^*),$

(iv) $\eta(t_1) = -N_1^*\big(N_1(\mathscr{H}(U^{-1}B^*\eta))(t_1) + (M_1 + N_1\Phi(t_1)z(t_0))\big)$
$+ N^*\big((M + N\Phi(t_1))^*\big)^*\big((M_1 + N_1\Phi(t_1))^* N_1(\mathscr{H}(U^{-1}B^*\eta))(t_1)$
$+ (M_1 + N_1\Phi(t_1))^*(M_1 + N_1\Phi(t_1))z(t_0)\big),$

(v) $\dot{\eta} = -A^*\eta + W(z - x_0)$ *almost everywhere.*

Proof. This follows from Corollaries 3.14 and 3.15. □

If $N = 0$ and $M_1 = 0$ then the above result reduces to a rather familiar result below.

COROLLARY 3.17. *Assume that*

$$F_\Omega(x) = Mx(t_0), \qquad F_\Lambda(x) = N_1 x(t_1)$$

for all $x \in \text{Dom } T_1$ *where* M *and* N_1 *are* $d \times n$ *and* $d_1 \times n$ *constant matrices, respectively. Then the unique optimal controller* u^+ *for the control problem is given by*

$$u^+ = U^{-1}B^*\eta \quad \textit{almost everywhere,}$$

where η *is an element of* Dom T_1 *for which there exists* $z \in \text{Dom } T_1$ *satisfying*:

(i) $\dot{z} = Az + BU^{-1}B^*\eta$ *almost everywhere,*

(ii) $MM^{\#}\gamma = Mz(t_0),$

(iii) $\int_{t_0}^{t_1} \Phi^* W(z - x_0)\, dt + (N_1\Phi(t_1))^*\big(N_1(\mathscr{H}(U^{-1}B^*\eta))(t_1) + N_1\Phi(t_1)z(t_0)\big)$
$\in \text{Range } (M^*\cdot),$

(iv) $\eta(t_1) = -N_1^*\big(N_1(\mathscr{H}(U^{-1}B^*\eta))(t_1) + N_1\Phi(t_1)z(t_0)\big),$

(v) $\dot{\eta} = -A^*\eta + W(z - x_0).$

Remark. When $M = I_{n\times n}$, the above result is well known in the literature. Notice that in this case the condition (iii) becomes redundant.

Example. In the control problem, take $[t_0, t_1] = [0, 1]$, $A = 0_{n\times n}$, $U = I_{m\times m}$, $W = I_{n\times n}$, $x_0 = 0_{n\times 1}$, $F_\Lambda = 0$, and

$$F_\Omega(x) = Mx(0) + Nx(1) + \int_0^1 f(t)x(t)\, dt, \qquad x \in \text{Dom } T_1,$$

where $M(d \times n)$, $N(d \times n)$ are constant matrices and $f(t)(d \times n)$ is a matrix-valued function of t such that f^*f is integrable in $[0, 1]$ entrywise.

Then the problem becomes:

Minimize $\int_0^1 (x^*x + u^*u)\, dt$, subject to $u \in X_m$, $x \in \text{Dom } T_1$, $\dot{x} = Bu$ almost everywhere, and the least-squares condition:

$$\left\| Mx(0) + Nx(1) + \int_0^1 fx\, dt - \gamma \right\|$$
$$= \text{Min}\left\{ \left\| My(0) + Ny(1) + \int_0^1 fy\, dt - \gamma \right\| : y \in \text{Dom } T_1, \dot{y} = Bu \text{ almost everywhere} \right\}.$$

Let $\mathcal{D}$ be the corresponding dynamical system, i.e., the set of all ordered pairs $\{u, x\}$ satisfying the above constraints. Then we see that the following are equivalent:

(1) $\{u, x\} \in \mathcal{D}$.

(2) $u \in X_m$, $x \in \text{Dom } T_1$, $\dot{x} = Bu$ almost everywhere,

and

$$\int_0^1 fx\,dt + Mx(0) + Nx(1) - \gamma = (QQ^* - I)\left(\gamma - \int_0^1 \left(N^* + \int_t^1 f(s)\,ds\right)B(t)u(t)\,dt\right).$$

(3) $u \in X_m$, $x = \int_0^t B(s)u(s)\,ds - \alpha$

$$- Q^* \int_0^1 \left(N^* + \int_t^1 f(s)\,ds\right)B(t)u(t)\,dt + Q^*\gamma$$

for some $\alpha \in \text{Null } Q$,

where Q (corresponds to Q_Ω) is the $d \times n$ matrix $M + N + \int_0^1 f\,dt$. Notice that for $\{u, x\} \in \mathcal{D}$,

$$x(0) = Q^*\gamma - \alpha - Q^* \int_0^1 \left(N^* + \int_t^1 f(s)\,ds\right)B(t)u(t)\,dt$$

for some $\alpha \in \text{Null } Q$.

Thus if $f = 0$, $N = 0$, then the least squares condition becomes a transversality condition:

$$x(0) - M^*\gamma \in \text{Null } M.$$

Since this condition does not depend on controls, even though $\mathcal{D}$ generates multi-responses, one can treat the control problem as a standard single response problem by eliminating "multi-responseness".

However, if $N \neq 0$, $f \neq 0$, and Null $Q \neq \{0\}$, then $x(0)$ depends directly on controls and Null Q, and so the least-squares condition is no longer a transversality condition. Thus it is not clear whether the maximum principle or the Bellman's dynamical programming method can even be useful in handling this case.

By Corollary 3.13 we have: The optimal controller u^+ is given by

$$u^+ = B^*\eta \quad \text{almost everywhere}$$

where η is an element of Dom T_1 for which there exists $z \in \text{Dom } T_1$ such that

(i) $\dot{z} = BB^*\eta$ almost everywhere,

(ii) $z(0) + Q^*\left(N^*(z(1) - z(0)) + \int_0^1 f(z - z(0))\,dt - \gamma\right) \in \text{Null } Q$,

(iii) $\int_0^1 z\,dt \in \text{Range } Q^*$,

(iv) $\eta(1) = N^*(Q^*)^* \int_0^1 z\,dt$,

(v) $\dot{\eta} = z - f^*(Q^*)^* \int_0^1 z\,dt$ almost everywhere.

Let $\mathcal{P}$ be the orthogonal projector of $\mathbb{C}^n$ onto Null Q, and let

$$g = Q^{\#}\gamma - Q^{\#}\int_0^1 \left(N^* + \int_t^1 f(s)\,ds\right)B(t)u(t)\,dt.$$

Then using Theorem 3.9 we see that

$$x^+ := \int_0^t B(s)u^+(s)\,ds + g + \mathcal{P}\left(\int_0^1 (s-1)B(s)u^+(s)\,ds\right)$$

is the unique optimal response corresponding to u^+, while for any $\alpha \in$ Null Q,

$$x := \int_0^t B(s)u^+(s)\,ds + g + \alpha$$

is a nonoptimal response corresponding to u^+.

Acknowledgment. The author thanks the referees for the constructive comments which led to improving this paper.

REFERENCES

[1] J. A. BURNS, *Existence theorems and necessary conditions for a general formulation of the minimum effort problems*, J. Optim. Theory Appl., 15 (1975), pp. 413-440.

[2] J. L. CASTI, *The linear-quadratic control problem: some recent results and outstanding problems*, SIAM Rev., 22 (1980), pp. 459-485.

[3] R. DATKO, *A linear control problem in an abstract Hilbert space*, J. Differential Equations, 9 (1971), pp. 346-359.

[4] I. HALPERIN, *Closures and adjoints of linear differential operators*, Ann. Math., 38 (1937), pp. 880-919.

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

[6] S. J. LEE, *Boundary conditions for linear manifolds* I, J. Math. Anal. Appl., 73 (1980), pp. 138-160.

[7] S. J. LEE AND M. Z. NASHED, *Generalized inverses for linear manifolds and applications to boundary value problems in Banach spaces*, C.R. Math. Rep. Acad. Sci. Canada, 4 (1982), pp. 347-352.

[8] ———, *Operator parts and generalized inverses of multi-valued operators with applications to ordinary differential subspaces*, to appear.

[9] ———, *Least-squares solution of multivalued linear operator equations in Hilbert spaces*, J. Approx. Theory, 38 (1983), pp. 380-391.

[10] N. MINAMIDE AND K. NAKAMURA, *A restricted pseudoinverse and its application to constrained minima*, SIAM J. Appl. Math., 19 (1970), pp. 167-177.

[11] ———, *Linear bounded phase coordinate control problems under certain regularity and normality conditions*, SIAM J. Control, 10 (1) (1972), pp. 82-92.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 4, July 1986

015

OPEN LOOP CONTROL OF WATER WAVES IN AN IRREGULAR DOMAIN*

RUSSELL M. REID†

Abstract. This paper considers open loop control of linear, small amplitude waves on a fluid surface. It extends the controllability results of Reid and Russell [4] to a two-dimensional domain with irregular bottom contour, constructing a null control via its Laplace transform, provided the bottom has finite arclength and no beaches. The domain may be multiply connected, i.e. contain fixed objects, provided they do not touch the surface.

Key words. water waves, small amplitude waves, linear waves

AMS(MOS) subject classification. 93

1. Introduction. This paper builds on results from an earlier paper by Reid and Russell [4], and relies heavily on that paper for background and comprehensibility. In [4], we developed a model for a controlled system of water waves on the surface of a two-dimensional region. The development of the model and the proofs of existence/uniqueness required no special assumptions about the shape of the region. The controllability result in [4], however, depended on knowledge of the eigenvalues of the underlying system, and thus applied only to a simple geometry for which the eigenvalues are known (infinitely deep tank with straight sides). The control result obtained was that an (essentially) arbitrary initial state can be steered to zero (surface is still) in infinite time, but not in any finite time, with the boundary control modeled. The rate of convergence to zero was not examined.

The current paper perturbs the argument in [4] to obtain the same result for variable, finite-depth regions. It constructs on open-loop, boundary null-control, i.e. one that steers an "arbitrary" initial state to zero in infinite time, and shows that no such control is possible in finite time.

This work begins with the evolution equation derived in [4] to describe linear waves controlled by a vibrating wall, and uses the existence/uniqueness results proved there. It also uses the result that the evolution operator A is self-adjoint with compact resolvent, and that an open loop null control must satisfy a certain moment equation involving exponentials of eigenvalues (equation (1.5) here). In [4], we solved the equation (1.5) for a simple geometry for which both eigenvalues and eigenfunctions are known, namely an infinitely deep, straight-sided tank. This work solves (1.5) for much more general geometries, following the same line of attack as [4]. It proceeds by perturbation of the work in [4], first by estimating the eigenvalues for a more general geometry, then by establishing functions which have zeros at those eigenvalues and which are recognizable as Laplace transforms of L^2 functions. Motivation for the perturbation techniques comes from equation (2.4) of this paper, which displays a potential function for the infinite depth tank. That potential decays exponentially with depth, leading one to expect that the effect of a bottom should also decay exponentially with depth. Such a guess is essentially correct, and is the key to eigenvalue estimation: one can estimate the influence of a bottom by looking in the infinite depth tank at the flow through the proposed bottom, then computing the influence of placing a bottom there. Calculating that surface influence is the same as computing the electric field at the surface due to a charge distribution placed on the bottom, where the charge

* Received by the editors April 1, 1984, and in revised form April 15, 1985.

† Department of Mathematical and Computer Sciences, Michigan Technological University, Houghton, Michigan 49931.

distribution equals the flow in the infinite depth tank through the proposed bottom; an integral form for such a potential is known.

In the proofs here, ideas and estimates are borrowed from the proofs in [4], so a reader must either recall key estimates from [4] or accept them without proof. If arguments duplicate those in [4], only necessary changes are supplied.

We begin the current work with a summary of the problem formulation, then give a brief overview of what follows.

Consider a fluid in a two-dimensional region Ω, with nonzero vertical walls at water level, i.e. no beaches. Denote the vertical segments by W_1 and W_2, the remaining fixed boundary (the bottom) by Γ, and the undisturbed surface ($z=0$) by S_0. Assume Γ has finite arc length. See Fig. 1.

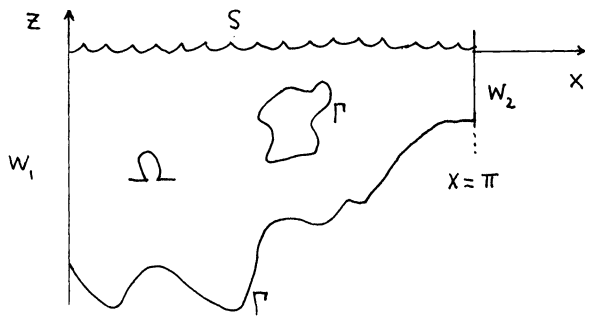


FIG. 1

The surface contour $\zeta(x, t)$ is a solution of

$$\ddot{\zeta} = -A\zeta, \tag{1.1}$$

where the evolution operator A is defined in terms of a harmonic "acceleration potential" ψ_ζ, which satisfies:

$$\begin{aligned} &\psi_\zeta = \zeta \quad \text{on } S_0, \\ &\frac{\partial \psi_\zeta}{\partial n} = 0 \quad \text{on } \Gamma \cup W_1 \cup W_2. \end{aligned} \tag{1.2}$$

(If fluid velocity is $-\nabla\phi$, then $\psi = \partial\phi/\partial t$, hence the name "acceleration potential.")

Then A is defined by

$$A\zeta = \left.\frac{\partial \psi_\zeta}{\partial z}\right|_{S_0}, \tag{1.3}$$

provided $\zeta \in D(A)$, i.e. $\zeta \in H^1 \cap L^2[0, \pi]$ and $\int_0^\pi \zeta(x)\,dx = 0$. (The last condition is needed to ensure volume conservation; hereafter we write it as $L_0^2(0, \pi)$.)

The controlled system, where wall W_2 oscillates with velocity $F(z)U(t)$, can be written

$$\frac{\partial}{\partial t}\begin{pmatrix} \zeta \\ \eta \end{pmatrix} = \begin{bmatrix} 0 & I \\ -A & 0 \end{bmatrix}\begin{pmatrix} \zeta \\ \eta \end{pmatrix} + \begin{pmatrix} 0 \\ d(x) \end{pmatrix} u(t) \tag{1.4}$$

or more briefly $(\partial/\partial t)\xi = A_0\xi + B(x)u(t)$, with $\eta = \partial\zeta/\partial t$, and $d(x)$ an L^2 control distribution function.

Existence, uniqueness, and regularity of solutions to (1.4) were established in [4], and an open loop null control $u(t)$ valid in infinite time was shown to satisfy

$$a_k + b_k \int_0^\infty e^{-w_k t} u(t)\, dt = 0, \qquad k = \pm 1, \pm 2, \pm 3, \cdots. \tag{1.5}$$

Here w_k are the eigenvalues of A_0; a_k and b_k are expansion coefficients of $\xi(x, 0)$ and $B(x)$ in terms of eigenfunctions of A_0. (It was shown in [4] that no such control can be hoped for in finite time, because wave propagation speeds approach zero as wavelength approaches zero.)

Existence of a control $u(t)$ satisfying (1.5) depends on the eigenvalues w_k. For a simple domain $\Omega_{\text{inf}} = \{(x, z): 0 \leqq x \leqq \pi,\ z \leqq 0\}$ (with straight sides and infinite depth), the operator $A \equiv A_{\text{inf}}$, its eigenvalues and eigenfunctions are known; solutions to (1.5) were constructed for this case in [4]. Here we prove similar results for a more general region Ω. In what follows, Ω_{inf}, A_{inf}, ψ_{inf} refer to the infinite depth domain, while Ω, A, ψ refer to the finite domain. The eigenvalues of A_{inf} were established in [4] as $\lambda_k = k$, $k = 0, 1, 2, \cdots$; here we establish and use estimates of the eigenvalues of A. We will use λ_k henceforth to denote eigenvalues of A, and w_k to denote eigenvalues of A_0 as above, specifically

$$w_{\pm k} = \pm i\sqrt{\lambda_k}, \qquad k = 1, 2, 3, \cdots. \tag{1.6}$$

Theorem 2.1 shows that the eigenvalues λ_k of the evolution operator A in the finite domain Ω are close to the eigenvalues of A_{inf} in the simpler domain Ω_{inf}, specifically $|\lambda_k - k| = O(1/k)$. Lemma 3.1 uses this closeness to show that a product analogous to $1/\Gamma(z)$, but with zeros at λ_k instead of k, is asymptotic to $1/\Gamma(z)$ in any sector lying in the right half plane. Using this product and a property similar to the reflection property of the Gamma function, Theorem 3.1 establishes a function $G_p(z)$ with zeros at $\pm i\sqrt{\lambda_k}$ and bounded in the right half plane. Theorem 3.2 shows that a control $u(t)$ can be constructed using $G_p(z)$ in the same way as for the simpler geometry Ω_{inf}. Smoothness conditions on the initial state which ensure convergence of the series defining $u(t)$ are the same as for the infinite depth case.

2. Eigenvalue estimates.

THEOREM 2.1. *The operator A has a discrete spectrum of eigenvalues $\lambda_k = k + \varepsilon_k/k$, where the $|\varepsilon_k|$ are bounded, and an associated complete orthonormal set of eigenfunctions $\phi_k(x)$.*

Proof. We prove three lemmas showing that the effect of a bottom is bounded: Define B and ψ_p by

$$A = A_{\text{inf}} + B, \qquad \psi = \psi_{\text{inf}} + \psi_p, \tag{2.1}$$

where B is a bottom perturbation operator and ψ_p is a harmonic function defined in Ω. The domain of B is that of A and A_{inf}, namely $H^1 \cap L_0^2(0, \pi)$, and B is related to ψ_p by (1.3), $B = \partial\psi_p/\partial z|_{S_0}$. Since both A and A_{inf} are self-adjoint (from [4]) and since

$$A^2 = A_{\text{inf}}^2 + A_{\text{inf}} \circ B + B \circ A_{\text{inf}} + B^2, \tag{2.2}$$

(where $D(A^2) = H^2 \cap L_0^2(0, \pi)$), we can display the eigenvalues of A as $(k^2 + O(1))^{1/2} = k + O(k^{-1})$ provided we can show that the last three terms in (2.2) are bounded. To this end, let

$$\zeta(x) = \sum_1^\infty a_n \cos(nx) \tag{2.3}$$

be the Fourier expansion of ζ ($a_0 = 0$ by conservation of volume). The potential

corresponding to $\zeta(x)$ in Ω_{inf} must satisfy $\psi(z=0)=\zeta(x)$, $\partial\psi/\partial x=0$ at $x=0$, $x=\pi$ and $\partial\psi/\partial z\to 0$ as $z\to-\infty$, and thus can be verified to be

$$\psi_{\text{inf}}(x, z)=\sum_{1}^{\infty} a_n e^{nz}\cos nx. \tag{2.4}$$

Therefore,

$$A_{\text{inf}}(\cos nx)=n\cos nx, \qquad n=1, 2, 3, \cdots. \tag{2.5}$$

Assume $\zeta\in H^1[0, \pi]$, so

$$\sum_{1}^{\infty} n^2 a_n^2<\infty. \tag{2.6}$$

To estimate B, we derive an integral representation for ψ_p as defined in (2.1). Construct an extended domain as shown, (see Fig. 2) so that the new domain is symmetric about $x=0$, $x=\pi$, and $z=0$. If $\Gamma(s)$ is a parametrization of the bottom contour, define a flux density $\sigma(s)$ on Γ by

$$\pi\sigma(s)\equiv\left.\frac{-\partial\psi_{\text{inf}}}{\partial n}\right|_{\Gamma(s)}, \tag{2.7}$$

defined on $\Gamma(s)$ and extended as in Fig. 2 to be symmetric about $x=0$, $x=\pi$, antisymmetric about $z=0$. $(\sigma(s, z)=-\sigma(s, -z))$

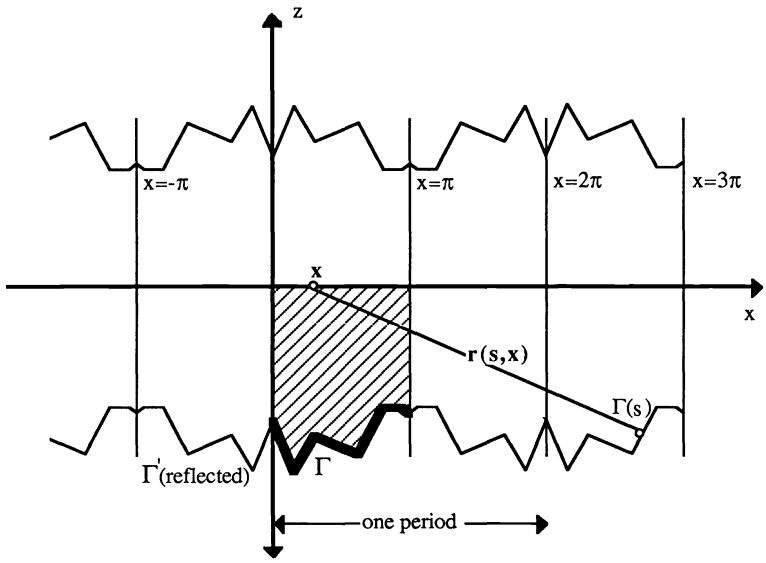


FIG. 2

Let Γ' represent the extension of Γ shown in Fig. 2, and hereafter $\sigma(s)$ refers to the extended domain although the notation is unchanged.

LEMMA 2.2. *The perturbation potential ψ_p and operator B of* (2.1) *are given by*

$$\psi_p(x, z)=\int_{\Gamma'}\sigma(s)\ln r(s, x, z)\,dx, \tag{2.8}$$

$$B\zeta(x)=\int_{\Gamma'}\frac{\sigma(s)h(s)}{[r(s, x, 0)]^2}\,ds \tag{2.9}$$

where $h(s)$ is the z coordinate at the point parametrized by s, and $r(s, x, z)$ is its distance from (x, z). (Dependence of ψ_p and B on $\zeta(x)$ comes from $\sigma(s)$ via (2.3), (2.4), (2.7).) *Note that $B = (\partial/\partial z)(\psi_p)|_{z=0}$, and that B is linear.*

Proof. ψ_p is the integral form of the two-dimensional potential whose normal derivative on $\Gamma'(s)$ is $\pi\sigma(s)$. Symmetry of the domain and the flux density $\sigma(s)$ shows that ψ_p has normal derivative zero on W_1 and W_2 and that $\psi_p = 0$ on $z = 0$, and equation (2.7) shows that $(\partial/\partial n)(\psi_{\text{inf}} + \psi_p) = 0$ on Γ. Thus $\psi_{\text{inf}} + \psi_p$ fulfills the requirements (1.2) for ψ_ζ (recall that $\psi_{\text{inf}} = \zeta$ on S_0), so ψ_p and $B\zeta \equiv (\partial/\partial z)\psi_p|_{z=0}$ satisfy (2.1). □

LEMMA 2.3. *$B \circ A_{\text{inf}}$ is bounded.*

Proof. Using (2.3), (2.5), we have

$$B \circ A_{\text{inf}}(\zeta) = \sum_1^\infty n a_n (B \cos nx). \tag{2.10}$$

Let $\sigma_n(s)$ be the flux density $\sigma(s)$ obtained by taking $\zeta(x) = \cos nx$ (compute ψ_{inf} with value $\cos nx$ on the surface S_0, and denote the result by ψ^n_{inf}. Use (2.7) to calculate $\sigma(s)$, and denote that result by $\sigma_n(s)$). Then $B(\cos nx)$ can be estimated by putting an estimate of $\sigma_n(s)$ into the integral form (2.9) for B. Since the potential ψ^n_{inf} is just that of a single term in (2.4),

$$\psi^n_{\text{inf}} = e^{nz} \cos nx, \tag{2.11}$$

and (2.7) implies

$$|\sigma_n(s)| \leqq \frac{1}{\pi} |\nabla \psi^n_{\text{inf}}| \leqq n\, e^{-nH} \tag{2.12}$$

where H is the minimum depth of the tank. If M is the maximum depth, we see from (2.9) that

$$|B \cos (nx)| \leqq \int_{\Gamma'} \frac{Mn\, e^{-nH}}{[r(s, x, 0)]^2}\, ds \leqq Kn\, e^{-nH} \tag{2.13}$$

for some k independent of n, provided Γ has finite arc length. Combining this with (2.10) we have, recalling (2.3),

$$\| B \circ A_{\text{inf}}(\zeta) \| \leqq K \sum_1^\infty n^2 a_n\, e^{-nH} \leqq C \| \zeta \|, \tag{2.14}$$

so that $B \circ A_{\text{inf}}$ is bounded. □

LEMMA 2.4. *$A_{\text{inf}} \circ B$ is bounded.*

Proof.

$$A_{\text{inf}} \circ B = \sum_1^\infty a_n A_{\text{inf}} \circ B \cos (nx), \tag{2.15}$$

so by the Schwarz inequality we need only a bound for

$$\sum_1^\infty |A_{\text{inf}} \circ B(\cos nx)|^2. \tag{2.16}$$

Use estimate (2.12) to define $|\hat{\sigma}_n(s)| \leqq 1$ by

$$\sigma_n(s) = n\, e^{-nH} \hat{\sigma}_n(s), \tag{2.17}$$

then consider an integral similar to (2.13):

$$A_{\text{inf}} \circ B(\cos nx) = n\, e^{-nH} \left[A_{\text{inf}} \circ \int_{\Gamma'} \frac{\hat{\sigma}_n(s) h(s)}{[r(s, x, 0)]^2}\, ds \right]. \tag{2.18}$$

The term in brackets may be shown to be bounded independent of n by estimating the derivative of the integral, and noting from (2.5) that

$$\|A_{\inf} f(x)\|_{L^2} = \|f_n'(x)\|_{L^2}.$$

The derivative of the integral can be estimated by:

$$\begin{aligned}\int_{\Gamma'} \frac{\hat{\sigma}_n(s)h(s)(\partial/\partial x)r(s,x,0)}{-2r(s,x,0)^3}\,dx &\leqq \frac{M}{2}\int_{\Gamma'} \frac{(\partial/\partial x)r(s,x,0)}{((x+H)/\sqrt{2})^3}\,ds \\ &\leqq \frac{M\sqrt{2}}{4}\int_{\Gamma'} \frac{1}{(x+H)^3}\,ds < \infty.\end{aligned} \tag{2.19}$$

Note that $|(\partial/\partial x)r(s,x,0)| \leqq 1$ because $r(s,x,0)$ is the distance from $(x,0)$ to the point parametrized by s, and that the arclength of Γ is finite. □

Finally, we note that $B \circ B$ is bounded, from (2.13), establishing Theorem 2.1.

We know from [14] that A is self-adjoint with compact resolvent, hence has a complete system of eigenfunctions. □

3. Construction of an open-loop control. Under some assumptions on a_k and b_k, (1.5) can be solved for a control $u(t) \in L^1 \cap L^2[0,\infty)$. Rewrite (1.5), recalling the w_k are eigenvalues of $A_0 = \left[\begin{smallmatrix} 0 & I \\ -A & 0 \end{smallmatrix}\right]$, and the λ_k the eigenvalues of A, as

$$\int_0^\infty e^{-w_k t} u(t)\,dt = c_k \equiv \frac{-a_k}{b_k}, \qquad k = \pm 1, \pm 2, \cdots, \tag{3.1}$$

with the convention $c_0 = 0$, and the approximate controllability assumption $b_k \neq 0$. Solve this moment problem for $u(t)$ in a way parallelling [4]: establish a function $G_p(z)$ which is uniformly bounded in the right half-plane and which has simple zeros at the $w_{\pm k} = \pm i\sqrt{\lambda_k}$, $k = 1, 2, 3, \cdots$. Then factor out zeros one at a time to establish a set of functions $p_k(z)$, $k = \pm 1, \pm 2, \cdots$, satisfying

$$\int_0^\infty e^{-w_k t} p_j(t)\,dt = \delta_{kj}, \qquad k, j = \pm 1, \pm 2, \pm 3, \cdots \tag{3.2}$$

so that $u(t)$ satisfying (3.1) is

$$u(t) = \sum_{-\infty}^{\infty} c_k p_k(t). \tag{3.3}$$

To construct $G_p(z)$ as noted above, let $\Gamma_p(z)$ be a "perturbed" gamma function with poles at $z = -\lambda_n$ instead of $z = -n$:

$$\frac{1}{\Gamma_p(z)} = z\, e^{\gamma z} \prod_{n=1}^{\infty} \left(1 + \frac{z}{\lambda_n}\right) e^{-z/n}. \tag{3.4}$$

The product converges absolutely since $\{\lambda_n - n\}$ is bounded (Theorem 2.1).

THEOREM 3.1. *The function*

$$G_p(z) = \frac{\Gamma[(z+1)^2]}{\Gamma_p(z^2)[e^{z+1}\Gamma(z+1)]^4} \tag{3.5}$$

is analytic for $\operatorname{Re}(z) > -1$, *uniformly bounded for* $\operatorname{Re}(z) \geqq 0$, *has bounded derivatives on the imaginary axis, and simple zeros at* $w_{\pm k} = \pm i\sqrt{\lambda_k}$.

Proof. Zeros and analyticity of $G_p(z)$ are clear; we establish boundedness by comparing $G_p(z)$ with the $G(z)$ discussed in [4], using the result from [4] that $G(z)$ is uniformly bounded in $\operatorname{Re}(z) \geqq 0$ with bounded derivatives on $\operatorname{Re}(z) = 0$.

LEMMA 3.1. *Let*

$$H(z)=\frac{\Gamma(z^2)}{\Gamma_p(z^2)}. \tag{3.6}$$

Then $H(z)$ and $1/H(z)$ are uniformly bounded for

$$|\arg z|\leqq\frac{\pi}{2}-\theta \quad \textit{for any } \theta>0. \tag{3.7}$$

Proof.

$$H(z)=\prod_{n=1}^{\infty}\frac{(1+z^2/(n+\varepsilon_n))}{(1+z^2/n)}=\prod_{n=1}^{\infty}\left(1-\frac{\varepsilon_n}{n+\varepsilon_n}+\frac{\varepsilon_n}{n+z^2}-\frac{\varepsilon_n^2}{(n+\varepsilon_n)(n+z^2)}\right) \tag{3.8}$$

where $\lambda_n=n+\varepsilon_n$ as in Theorem 2.1. Since $\varepsilon_n=\delta_n/n$ where $|\delta_n|\leqq C$,

$$\left|\frac{\varepsilon_n}{n+z^2}\right|\leqq\frac{C}{n^{3/2}}\csc(2\theta). \tag{3.9}$$

The product in (3.8) converges provided the sum of the last three terms converges, and

$$\sum_{n=1}^{\infty}\left|\frac{\varepsilon_n}{n+\varepsilon_n}\right|+\left|\frac{\varepsilon_n}{n+z^2}\right|+\left|\frac{\varepsilon_n}{(n+\varepsilon_n)(n+z^2)}\right|$$
$$\leqq\sum_{1}^{\infty}\frac{C}{(n+\varepsilon_n)\sqrt{n}}+\frac{C\csc(2\theta)}{n\sqrt{n}}+\frac{C\csc 2\theta}{(n+\varepsilon_n)n\sqrt{n}}.$$

The series on the right converges absolutely, and is bounded independent of z in the argument range (3.7). A similar analysis shows uniform boundedness of $1/H(z)$ in the same sector. □

Lemma 3.1 implies that $G_p(z)=H(z)G(z)$ is bounded in the argument range (3.7). In establishing boundedness of $G(z)$ when $|\arg z|=\pi/2$ in [4], we used the reflection principle

$$\frac{1}{\Gamma(z)}=\frac{-z\sin\pi z}{\pi}\Gamma(-z). \tag{3.10}$$

We need a similar reflection equation for $\Gamma_p(z)$; from (3.4) we see that

$$\frac{1}{\Gamma_p(z)}=-z^2\prod_{1}^{\infty}\left(1-\frac{z^2}{\lambda_n^2}\right)\Gamma_p(-z). \tag{3.11}$$

Let

$$p(z)=z\prod_{1}^{\infty}\left(1-\frac{z^2}{\lambda_n^2}\right). \tag{3.12}$$

Then we expect $p(z)$ to be related to $\sin\pi z$; in fact we can show:

LEMMA 3.2. *For some $\beta(s)\in L^2[-\pi,\pi]$, and some constant C,*

$$p(z)=z\prod_{1}^{\infty}\left(1-\frac{z^2}{\lambda_n^2}\right)=C\left[\sin\pi z-\int_{-\pi}^{\pi}\beta(s)\,e^{izs}\,ds\right]. \tag{3.13}$$

Proof. The exponentials $e^{\pm i\lambda_k s}$, together with 1, form a Riesz basis for $L^2[-\pi,\pi]$ because $\{\lambda_k-k\}$ is square summable (see [2]). In addition, $\{\sin\lambda_k\pi\}$ is square summable, so that there is a unique $\beta\in L^2[-\pi,\pi]$ with

$$\int_{-\pi}^{\pi}\beta(s)\,e^{i\lambda_k s}\,ds=\sin\lambda_k\pi, \qquad k=0,\pm1,\pm2,\pm3. \tag{3.14}$$

Therefore

$$g(z) = \sin \pi z - \int_{-\pi}^{\pi} \beta(s)\, e^{izs}\, ds \tag{3.15}$$

is entire, and has zeros precisely at $z = 0, \pm i\lambda_k$. The product (3.12) shows $p(z)$ to be entire with the same zeros. Choose α a zero of neither p nor g, define

$$p(\alpha) = Cg(\alpha),$$

then

$$\hat{\xi}(z) \equiv \frac{p(z) - Cg(z)}{z - \alpha}$$

is entire, L^2 on the real axis, and of exponential type π (see [3]). Therefore it is the Fourier transform of some $\xi(t) \in L^2[-\pi, \pi]$, which is orthogonal to 1, $e^{\pm\lambda_k t}$ and thus is zero. □

We complete the proof of Theorem 3.1 by noting that for $\pi/2 - \theta \leqq |\arg z| \leqq \pi/2$ (using estimates of $G(z)$ from [4],

$$G_p(z) = \frac{1}{4\pi^2 e} \cdot \left[1 - e^{-2\pi i z^2} - \int_{-\pi}^{\pi} \beta(s)\, e^{iz^2(s-\pi)}\, ds\right] C \frac{\Gamma_p(-z^2)}{\Gamma(-z^2)} \tag{3.16}$$

and thus is bounded and has bounded derivatives there.

THEOREM 3.2. *The functions $\hat{P}_k(z)$, defined by*

$$\hat{p}_k(z) = \frac{w_k G(z)}{z(z - w_k) G'(w_k)}, \qquad k = \pm 1, \pm 2, \pm 3, \cdots \tag{3.17}$$

are Laplace transforms of functions $p_k(t) \in L^1 \cap L^2[0, \infty)$ and satisfy biorthogonality relations

$$\hat{p}_k(w_j) = \delta_{kj}. \tag{3.18}$$

Proof. Estimates of $G(z)$ from [4] and the fact that $|w_k - i\sqrt{k}| = O(1/k^{3/2})$ allow the proof to follow the infinite depth case: $\hat{p}_k(z)$ is in $L^2(-\infty, \infty)$ as a function of y, uniformly in x and with estimates independent of k. Since y derivatives of $\hat{p}_k(iy)$ are in L^2, $tp_k(t)$ must be in L^1 and therefore $p_k(t) \in L^1 \cap L^2[0, \infty)$. □

The series (3.3) for a control $a(t)$ will converge provided the initial state is sufficiently smooth.

The a_k are the expansion coefficients of the initial state ξ_0 in terms of eigenfunctions of A_0 (see (1.5)). The b_k are coefficients of a control distribution function $B(x)$, and are design parameters of the control mechanism. Assume they were chosen to be slowly decaying, say $|b_k| \geqq 1/k^p$, where $p > \frac{1}{2}$ ensures that $\{b_k\}$ is square summable. Then (3.3) will converge to an admissible control provided $|a_k/b_k|$ is summable, for example if $|a_k| < 1/k^q$ for some $q > p + 1$.

REFERENCES

[1] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.

[2] N. LEVINSON, *Gap and Density Theorems*, American Mathematical Society Colloquium Publications 26, 1940.

[3] R. E. A. C. PALEY AND N. WIENER, *The Fourier transform in the complex domain*, Amer. Math. Soc. Colloq. Publ. 19, 1934.

[4] R. M. REID AND D. L. RUSSELL, *Boundary control and stability of linear water waves*, this Journal, 23 (1985), pp. 111-121.

[5] D. L. RUSSELL, *Nonharmonic Fourier series in the control of distributed parameter systems*, J. Math. Anal. Appl., 18 (1977), pp. 542-559.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 4, July 1986

016

FINITE DIMENSIONAL COMPENSATORS FOR INFINITE DIMENSIONAL SYSTEMS WITH UNBOUNDED INPUT OPERATORS*

R. F. CURTAIN† AND D. SALAMON†‡

Abstract. This paper contains a design procedure for constructing finite dimensional compensators for a class of infinite dimensional systems with unbounded input operators. Applications to retarded functional differential systems with delays in the input or the output variable and to partial differential equations with boundary input operators are discussed.

Key words. infinite dimensional systems, unbounded input operators, compensator, retarded functional differential equations, partial differential equations, boundary control

AMS(MOS) subject classifications. 34K20, 93C15, 93C25, 93D15

1. Introduction. In [15], [16], [17] Schumacher presented a design procedure for constructing stabilizing compensators for a class of infinite dimensional systems. The novel feature was that the compensators were finite dimensional and that they could be readily numerically calculated from finitely many system parameters. The class included those systems described by retarded functional differential and partial differential equations provided that the eigenvectors of the system operators were complete and provided that the input and output operators were bounded. In [3] Curtain presented an alternative compensator design which applied to essentially the same class of systems, except that for the special case of parabolic systems unbounded input and output operators were allowed. By means of enlarging the state space of the given distributed boundary control system, Curtain in [2] essentially transformed the original problem with unbounded control into one with bounded control action so that the techniques of either [16] or [3] could be applied. The resulting control, however, was of integral type. Neither of these two compensator designs are applicable to retarded systems with delays in the control or the observation.

In the present paper we make use of the abstract approach developed by Salamon [14] to extend the results of Schumacher [16] to allow for unbounded control action. This is done in a direct way without reformulating the original problem into one with a bounded input operator. In § 2 we outline the abstract formulation and prove a theorem on the existence of a finite dimensional compensator paralleling the development in [16]. In § 3 the general approach is then applied to retarded functional differential systems with delays in either the input or the output variables. The conditions are easy to check and they are quite reasonable, except for the assumption of completeness of the eigenfunctions which seems to be too strong. In a special case we are able to weaken this assumption.

Finally, in § 4, we show how boundary control systems fit into the abstract framework of § 2 and give an example. The results are compared with the approach in [3].

* Received by the editors May 30, 1984, and in revised form February 8, 1985.

† Mathematisch Instituut, Rijksuniversiteit Groningen, the Netherlands. This work was sponsored by the U.S. Army under contract DAAG29-80-C-0041.

‡ Present address, Forschungsinstitut für Mathematik, ETH-Zentrum, CH-8092 Zürich, Switzerland. This material is based upon work of this author supported by the National Science Foundation under grant MCS-8210950.

2. A general result. We consider the abstract Cauchy problem

$$\dot{x}(t) = Ax(t) + Bu(t), \qquad x(0) = x_0 \in X, \tag{2.1.1}$$

$$y(t) = Cx(t), \tag{2.1.2}$$

on the real, reflexive Banach space X where $A: \mathcal{D}(A) \to X$ is the infinitesimal generator of a strongly continuous semigroup $S(t)$ on X and $C \in \mathcal{L}(X, \mathbb{R}^m)$.

In order to give a precise definition of what we mean by an unbounded input operator we need an extended state space $Z \supset X$. For this purpose let us first introduce the subspace

$$Z^* = \mathcal{D}_{X^*}(A^*) \subset X^*$$

endowed with the graph norm of A^*. Then Z^* becomes a real, reflexive Banach space and the injection of Z^* into X^* is continuous and dense. Defining Z to be the dual space of Z^*, we obtain by duality that

$$X \subset Z$$

with a continuous dense injection.

Remarks 2.1. (i) A^* can be regarded as a bounded operator from Z^* into X^* and $S^*(t)$ restricts to a strongly continuous semigroup on Z^*. By duality, A extends to a bounded operator from X into Z. This extension, regarded as an unbounded operator on Z, is the infinitesimal generator of the extended semigroup $S(t) \in \mathcal{L}(Z)$ (see [14, Lemma 1.3.2]).

(ii) If $\mu \notin \sigma(A) = \sigma(A^*)$, then the operator $\mu I - A: X \to Z$ is bijective. Furthermore, this operator commutes with the semigroup $S(t)$ so that it provides a similarity action between the semigroups $S(t) \in \mathcal{L}(X)$ and $S(t) \in \mathcal{L}(Z)$.

(iii) It follows from (ii) that the exponential growth rate of the semigroup $S(t)$ is the same on the state spaces X and Z, i.e.

$$\omega_0 = \lim_{t\to\infty} t^{-1} \log \|S(t)\|_{\mathcal{L}(X)} = \lim_{t\to\infty} t^{-1} \log \|S(t)\|_{\mathcal{L}(Z)}.$$

(iv) It also follows from (ii) that the spectrum of A on the state space X coincides with the spectrum of A on Z (see [14, Lemma 1.3.2]). Furthermore, the generalized eigenvectors for both operators are the same, since the eigenvectors of A and Z are contained in $\mathcal{D}_Z(A) = X$. Finally, the (generalized) eigenvectors of A are complete in X if and only if they are complete in Z.

We will always assume that B is a bounded, linear operator from R^l into Z. However, we want the solutions of (2.1.1) to be in the smaller space X on which the output operator is defined. Therefore we need the following hypothesis.

(H1) For every $T > 0$ there exists a constant $b_T > 0$ such that $\int_0^T S(T-s)Bu(s)\,ds \in X$ and

$$\left\| \int_0^T S(T-s)Bu(s)\,ds \right\|_X \leqq b_T \|u(\cdot)\|_{L^p[0,T;R^l]}$$

for every $u(\cdot) \in L^p[0, T; R^l]$ where $1 \leqq p < \infty$. In the following we collect some important consequences of (H1) which have been established in [14, § 1.3].

Remarks 2.2. (i) (H1) is satisfied if and only if the inequality

$$\|B^* S^*(\cdot)x^*\|_{L^q[0,T;R^l]} \leqq b_T \|x^*\|_{X^*}$$

holds for every $x^* \in Z^*$ and every $T > 0$ where $1/p + 1/q = 1$.

(ii) If (H1) is satisfied, then

$$x(t) = S(t)x_0 + \int_0^t S(t-s)Bu(s)\,ds \in X \tag{2.2}$$

is the unique strong solution of (2.1.1) for every $x_0 \in X$ and every $u(\cdot) \in L^p[0, T; R^l]$. More precisely $x(t)$ is continuous in X on the interval $[0, T]$ and satisfies

$$x(t) = x_0 + \int_0^t [Ax(s) + Bu(s)]\,ds, \qquad 0 \leqq t \leqq T,$$

where the integral has to be understood is the state space Z. Thus (2.1.1) is satisfied in the space Z for almost every $t \in [0, T]$.

(iii) If (H1) is satisfied, then for every $w(\cdot) \in \mathscr{C}[0, T; X]$ there exists a unique $x(\cdot) \in \mathscr{C}[0, T; X]$ satisfying the equation

$$x(t) = w(t) + \int_0^t S(t-s)BFx(s)\,ds, \qquad t \geqq 0.$$

This solution $x(\cdot)$ depends continuously on $w(\cdot)$.

Moreover hypothesis (H1) implies the following important perturbation result.

THEOREM 2.3. *Let $F \in \mathscr{L}(X; \mathbb{R}^l)$ be given. Then the following statements hold.*

(i) *There exists a unique strongly continuous semigroup $S_F(t)$ on X satisfying*

$$S_F(t)x = S(t)x + \int_0^t S(t-s)BFS_F(s)\,x\,ds \tag{2.3}$$

for every $x \in X$ and every $t \geqq 0$. Its infinitesimal generator is given by

$$\mathscr{D}(A_F) = \{x \in X \mid Ax + BFx \in X\},$$

$$A_F x = Ax + BFx.$$

(ii) $A_F^* = A^* + F^*B^*: \mathscr{D}(A_F^*) = Z^* \to X^*$.

(iii) *$S_F(t)$ extends to a strongly continuous semigroup on Z and the infinitesimal generator of the extended semigroup is given by $A + BF: X \to Z$.*

(iv) *Let $x_0 \in X$ and $v(\cdot) \in L^p[0, T; \mathbb{R}^l]$ be given and let $x(\cdot) \in \mathscr{C}[0, T; X]$ be the unique solution of*

$$x(t) = S(t)x_0 + \int_0^t S(t-s)B[Fx(s) + v(s)]\,ds, \qquad 0 \leqq t \leqq T. \tag{2.4}$$

Then

$$x(t) = S_F(t)x_0 + \int_0^t S_F(t-s)Bv(s)\,ds, \qquad 0 \leqq t \leqq T. \tag{2.5}$$

(v) *Hypothesis* (H1) *is satisfied with $S(t)$ replaced by $S_F(t)$ and $S_F(t)$ satisfies*

$$S_F(t)x = S(t)x + \int_0^t S_F(t-s)BFS(s)x\,ds \tag{2.6}$$

for every $x \in X$ and every $t \geqq 0$.

(vi) *Let $x_0 \in X$ and $f(\cdot) \in L^p[0, T; X]$ be given and define*

$$x(t) = S_F(t)x_0 + \int_0^t S_F(t-s)f(s)\,ds, \qquad 0 \leqq t \leqq T. \tag{2.7}$$

Then

$$x(t) = S(t)x_0 + \int_0^t S(t-s)[BFx(s) + f(s)]\,ds, \qquad 0 \leqq t \leqq T. \tag{2.8}$$

Proof. Statement (i) has been shown in [14, Thm. 1.3.7] and (ii) follows from [14, Thm. 1.3.9] since the input space is finite-dimensional. By (ii), $S_F^*(t)$ restricts to a strongly continuous semigroup on $Z^* = \mathcal{D}(A_F^*)$ and hence $S_F(t)$ extends to a semigroup on Z. By Remark 2.1(i), the extended semigroup is generated by the adjoint operator of A_F^*, where A_F^* is regarded as a bounded operator from Z^* into X^*. This proves statement (iii).

In order to establish statement (iv), let us first assume that $v(\cdot) \in \mathscr{C}^1[0, T; \mathbb{R}^l]$ and let $x(\cdot) \in \mathscr{C}[0, T; X]$ be the unique solution of (2.4). Then it follows from Remark 2.2 (ii) that $x(\cdot) \in \mathscr{C}^1[0, T; Z]$ and

$$\frac{d}{dt} x(t) = (A + BF)x(t) + Bv(t), \qquad 0 \leqq t \leqq T.$$

Hence it follows from (iii) and a classical result in semigroup theory that $x(\cdot)$ is given by (2.5). In general, statement (iv) follows from the fact that the unique solutions of both (2.4) and (2.5), regarded as continuous functions with values in Z, depend continuously on $v(\cdot) \in L^p[0, T; \mathbb{R}^l]$.

It follows immediately from (iv) that (H1) is satisfied with $S(t)$ replaced by $S_F(t)$. Now let $x(t)$, $t \geqq 0$, be defined by the RHS of (2.6). Then it follows from (iv) that

$$\begin{aligned} x(t) &= S(t)x + \int_0^t S(t-s)B\left[F\int_0^s S_F(s-\tau)BFS(\tau)x\,d\tau + FS(s)x\right] ds \\ &= S(t)x + \int_0^t S(t-s)BFx(s)\,ds \end{aligned}$$

for $t \geqq 0$, and hence $x(t) = S_F(t)x$, by the definition of $S_F(t)$. This proves statement (v).

Statement (vi) can be established straightforwardly by inserting (2.3) into (2.7) and interchanging integrals. □

The aim of this section is to give sufficient conditions under which system (2.1) can be stabilized by a finite-dimensional compensator of the form

$$\dot{w}(t) = Mw(t) - Hy(t), \qquad w(0) = w_0, \tag{2.9.1}$$

$$u(t) = Kw(t) + v(t), \tag{2.9.2}$$

where $M \in \mathbb{R}^{N\times N}$, $H \in \mathbb{R}^{N\times m}$, $K \in \mathbb{R}^{l\times N}$ are suitably chosen matrices. To this end we need the following well-posedness result for the connected system (2.1), (2.9).

PROPOSITION 2.4. *Let* (H1) *be satisfied. Then for all* $x_0 \in X$, $w_0 \in \mathbb{R}^N$, $v(\cdot) \in L^p_{\mathrm{loc}}[0, \infty; \mathbb{R}^l]$ *there exists a unique solution pair* $x(t)$, $w(t)$ *of* (2.1) *and* (2.9). *This means that* $x(t)$ *is continuous in* X *and absolutely continuous in* Z, *that* (2.1.1) *is satisfied for almost every* $t \geqq 0$ *where* $u(t)$ *is given by* (2.9.2) *and that* $w(t) \in \mathbb{R}^N$ *is continuously differentiable and satisfies* (2.9.1) *where* $y(t)$ *is given by* (2.1.2).

Proof. Let us introduce the spaces $X_e = X \times \mathbb{R}^N$, $Z_e = Z \times \mathbb{R}^N$, $U_e = \mathbb{R}^l \times \mathbb{R}^m$ and the operators $S_e(t) \in \mathscr{L}(X_e)$, $B_e \in \mathscr{L}(U_e, Z_e)$, $F_e \in \mathscr{L}(X_e, U_e)$ by

$$S_e(t) = \begin{bmatrix} S(t) & 0 \\ 0 & e^{Mt} \end{bmatrix}, \quad B_e = \begin{bmatrix} B & 0 \\ 0 & -H \end{bmatrix}, \quad F_e = \begin{bmatrix} 0 & K \\ C & 0 \end{bmatrix}.$$

Then hypothesis (H1) is satisfied with X, Z, $S(t)$, B replaced by X_e, Z_e, $S_e(t)$, B_e, respectively. Moreover $x(t) \in X$ and $w(t) \in \mathbb{R}^N$ satisfy (2.1) and (2.9) in the above sense if and only if the following equation holds for every $t \geqq 0$

$$\begin{pmatrix} x(t) \\ w(t) \end{pmatrix} = S_e(t)\begin{pmatrix} x_0 \\ w_0 \end{pmatrix} + \int_0^t S_e(t-s)B_e\left[F_e\begin{pmatrix} x(s) \\ w(s) \end{pmatrix} + \begin{pmatrix} v(s) \\ 0 \end{pmatrix}\right] ds.$$

This proves the statement of the proposition. □

The following hypothesis together with (H1) will turn out to be sufficient for the existence of a stabilizing, finite dimensional compensator for system (2.1). It generalizes the approach of Schumacher [15], [16], [17] to systems with unbounded input operators.

(H2) Suppose that there exist operators $F \in \mathscr{L}(X, \mathbb{R}^l)$, $G \in \mathscr{L}(\mathbb{R}^m, X)$ and a finite dimensional subspace $W \subset X$ such that the following conditions are satisfied.

1. The feedback semigroup $S_F(t) \in \mathscr{L}(X)$, defined by (2.3), is exponentially stable.
2. The observer semigroup $S^G(t) \in \mathscr{L}(X)$, generated by $A+GC$ is exponentially stable.
3. $S_F(t)W \subset W$ for all $t \geqq 0$.
4. Range $G \subset W$.

If (H2) is satisfied and $N = \dim W$, then there exist linear maps $\iota: \mathbb{R}^N \to X$, $\pi: X \to \mathbb{R}^N$ satisfying

$$\pi\iota = \mathrm{id}, \quad \iota\pi x = x, \quad x \in W. \tag{2.10}$$

Moreover, $W \subset \mathscr{D}(A_F)$ and hence $\pi A_F \iota$ is a well defined linear map on $\mathbb{R}^N$. We will show that the system

$$\begin{aligned} \dot{w}(t) &= \pi(A_F + GC)\iota w(t) - \pi G y(t), \quad w(0) = w_0, \\ u(t) &= F\iota w(t), \end{aligned} \tag{2.11}$$

defines a stabilizing compensator for the Cauchy problem (2.1).

THEOREM 2.5. *If* (H1), (H2) *and* (2.10) *are satisfied, then the closed loop system* (2.1), (2.11) *is exponentially stable.*

Proof. By Proposition 2.4, the system (2.1), (2.11) is a well-posed Cauchy problem. Now let $x(t) \in X$, $w(t) \in \mathbb{R}^N$ be any solution pair of (2.1), (2.11) and define

$$z(t) = \iota w(t) - x(t) \in X, \qquad t \geqq 0. \tag{2.12}$$

Then

$$\dot{w}(t) = \pi A_F \iota w(t) + \pi G C z(t), \tag{2.13}$$

and hence, using $\pi S_F(t)\iota = e^{\pi A_F \iota t}$ and Theorem 2.3(vi) with $f(t) = GCz(t)$, we get

$$\begin{aligned} z(t) &= \iota\pi S_F(t)\iota w_0 + \int_0^t \iota\pi S_F(t-s)\iota\pi GCz(s)\, ds - x(t) \\ &= S_F(t)\iota w_0 + \int_0^t S_F(t-s)GCz(s)\, ds - x(t) \\ &= S(t)\iota w_0 + \int_0^t S(t-s)[BF\iota w(s) + GCz(s)]\, ds \\ &\quad - S(t)x_0 - \int_0^t S(t-s)Bu(s)\, ds \\ &= S(t)z(0) + \int_0^t S(t-s)GCz(s)\, ds, \qquad t \geqq 0. \end{aligned}$$

This implies $z(t) = S^G(t)z(0)$ and hence, by (2.13), stability of the pair $z(t)$, $w(t)$. Now the stability of the pair $x(t)$, $w(t)$ follows from (2.12). □

Clearly, the hypothesis (H2) is not very useful in the present form since it is rather difficult to check in concrete examples. Following the ideas of Schumacher [16], we transform (H2) into an easily verifiable criterion. The basic idea is to approximate G by generalized eigenvectors of A_F and to show that, if A has a complete set of

generalized eigenvectors and is stabilizable through B, then there exists a stabilizing feedback operator F which does not destroy the completeness property of A.

More precisely, we need the following assumptions on A.

(H3) The resolvent operator of A is compact and the set $\Lambda = \{\lambda \in P\sigma(A) | \operatorname{Re} \lambda \geqq -\omega\}$ is finite for some $\omega > 0$.

If (H3) is satisfied, then we may introduce the projection operator

$$P_\Lambda = \frac{1}{2\pi i} \int_\Gamma (\mu I - A)^{-1} \, d\mu$$

where Γ is a simple rectifiable curve surrounding Λ but no other eigenvalue of A. Clearly, P_Λ is a projection operator on both X and Z. Correspondingly we obtain the decomposition

$$X = X_\Lambda \oplus X^\Lambda, \qquad Z = X_\Lambda \oplus Z^\Lambda,$$

where $X_\Lambda = \text{range } P_\Lambda$, $Z^\Lambda = \ker P_\Lambda$, $X^\Lambda = Z^\Lambda \cap X$ are invariant subsapces under $S(t)$. If $N_\Lambda = \dim X_\Lambda$, we may identify X_Λ with $\mathbb{R}^{N_\Lambda}$ and obtain two maps

$$\iota_\Lambda : \mathbb{R}^{N_\Lambda} \to X_\Lambda, \qquad \pi_\Lambda : Z \to \mathbb{R}^{N_\Lambda}$$

with the properties

$$\pi_\Lambda \iota_\Lambda = \text{id}, \qquad \iota_\Lambda \pi_\Lambda = P_\Lambda. \tag{2.14}$$

Then the projection $x_\Lambda(t) = \pi_\Lambda x(t)$ of a solution to (2.1) satisfies the finite dimensional ODE

$$\begin{aligned} \dot{x}_\Lambda(t) &= A_\Lambda x_\Lambda(t) + B_\Lambda u(t), \qquad x_\Lambda(0) = \pi_\Lambda x_0, \\ y_\Lambda(t) &= C_\Lambda x_\Lambda(t) \end{aligned} \tag{2.15}$$

where

$$A_\Lambda = \pi_\Lambda A \iota_\Lambda, \quad B_\Lambda = \pi_\Lambda B, \quad C_\Lambda = C\iota_\Lambda. \tag{2.16}$$

Now we can replace (H2) by the following stronger conditions which can in many cases be easily verified. The result has been proved by Schumacher [16] for the case that range $B \subset X$ (bounded input operator).

PROPOSITION 2.6. *Let the operator A satisfy* (H3), *assume that the exponential estimate*

$$\| S(t)|_{X^\Lambda} \|_{\mathcal{L}(X^\Lambda)} \leqq M e^{-\omega t}, \qquad t \geqq 0, \tag{2.17}$$

holds for some $M \geqq 1$ and that the reduced finite dimensional system (2.15) *is controllable and observable. Furthermore, assume that the generalized eigenvectors of A are complete in X. Then* (H2) *is satisfied.*

Proof. Since (2.15) is controllable, there exists a matrix $F_\Lambda \in \mathbb{R}^{l \times N_\Lambda}$ such that the matrix $A_\Lambda + B_\Lambda F_\Lambda$ is stable. Furthermore, the estimate (2.17) implies that

$$\| S(t)|_{Z^\Lambda} \|_{\mathcal{L}(Z^\Lambda)} \leqq M e^{-\omega t}, \qquad t \geqq 0,$$

(Remark 2.1(iii)). It is a well-known result in infinite dimensional systems theory (see e.g. [5] or [16]) that under these assumptions the closed loop semigroup $S_F(t) \in \mathcal{L}(Z)$, generated by $A + BF : X \to Z$ with $F = F_\Lambda \pi_\Lambda : Z \to \mathbb{R}^l$ is exponentially stable. By Theorem 2.3, $S_F(t)$ restricts to a strongly continuous semigroup on X and the operator $\mu I - A - BF : X \to Z$ provides a similarity action between both semigroups (Remark 2.1(ii)). Hence the restricted semigroup $S_F(t) \in \mathcal{L}(X)$ is still exponentially stable (Remark 2.1(iii)).

By assumption, the generalized eigenvectors of A are complete in X and hence they are complete in Z (Remark 2.1(iv)). Therefore it is possible to choose the feedback matrix F_Λ such that $S_F(t)$ is exponentially stable and the generalized eigenvectors of $A+BF: X \to Z$ are complete in Z (Schumacher [16]). It follows again from the above similarity argument, that this completeness property carries over to the restricted operator $A_F: \mathcal{D}(A_F) \to X$ introduced in Theorem 2.3(i).

Now choose $G_\Lambda \in \mathbb{R}^{N_\Lambda \times m}$ such that $A_\Lambda + G_\Lambda C_\Lambda$ is stable and define $G = \iota_\Lambda G_\Lambda: \mathbb{R}^m \to X$. Then it is again a well-known fact from infinite-dimensional linear systems theory that $A+GC: \mathcal{D}(A) \to X$ generates an exponentially stable semigroup on X (see [5] or [16]). It is also well known that $A+\hat{G}C$ still generates an exponentially stable semigroup on X whenever $\|\hat{G}-G\|_{\mathcal{L}(\mathbb{R}^m,X)}$ is sufficiently small. Now we make use of the fact that the generalized eigenvectors of A_F are complete in X. This implies that $G: R^m \to X$ can be approximated arbitrarily close by an operator $\hat{G}: \mathbb{R}^m \to X$ whose range is spanned by finitely many generalized eigenvectors of A_F. We choose $\hat{G}$ in such a way that $A+\hat{G}C$ generates a stable semigroup and denote by W the finite dimensional subspace of X which is invariant under A_F and generated by those generalized eigenvectors which span the range of $\hat{G}$. Since W is a finite dimensional subspace contained in $\mathcal{D}(A_F)$ and invariant under A_F, the restriction of A_F to W is a bounded, linear operator generating a semigroup $S_W(t)$ on W. Since $d/dt\, S_W(t)x = A_F S_W(t)x$, the semigroup $S_W(t)$ coincides with $S_F(t)$ on W. Hence W is also invariant under the semigroup $S_F(t)$. We conclude that the operators $F: Z \to \mathbb{R}^l$ and $\hat{G}: \mathbb{R}^m \to X$ satisfy hypothesis (H2). □

Combining Proposition 2.6 with Theorem 2.5, we obtain a constructive procedure for designing a finite dimensional compensator for the Cauchy problem (2.1). The construction is based on the knowledge of the finite dimensional reduced system (2.15) and on the knowledge of sufficiently many eigenvalues and eigenvectors of the operator A_F. For the case of bounded input operators (range $B \subset X$) the procedure has been described in detail by Schumacher [16]. Precisely the same algorithm applies to the case where range $B \not\subset X$.

3. Retarded systems. In this section we apply the abstract result of the previous section to retarded functional differential equations (RFDE) with delays either in the input or in the output variable. If delays occur in the input and output variables at the same time, the RFDE can still be reformulated as an abstract Cauchy problem (see e.g. Pritchard-Salamon [12]) however, the completeness assumption will no longer be satisfied.

3.1. Retarded systems with output delays. We consider the linear RFDE

$$\begin{aligned} \dot{x}(t) &= Lx_t + B_0 u(t), \\ y(t) &= Cx_t, \end{aligned} \tag{3.1}$$

where $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^l$, $y(t) \in \mathbb{R}^m$ and x_t is defined by $x_t(\tau) = x(t+\tau)$ for $-h \leqq \tau \leqq 0$, $h > 0$. Correspondingly B_0 is a real $n \times l$-matrix and L and C are bounded linear functionals on $\mathcal{C} = \mathcal{C}[-h, 0; \mathbb{R}^n]$ with values in $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively. These can be written in the form

$$L\phi = \int_{-h}^{0} d\eta(\tau)\phi(\tau), \quad C\phi = \int_{-h}^{0} d\gamma(\tau)\phi(\tau), \quad \phi \in \mathcal{C},$$

where $\eta(\tau)$ and $\gamma(\tau)$ are normalized functions of bounded variation, i.e. they vanish for $\tau \geqq 0$, are constant for $\tau \leqq -h$ and left continuous for $-h < \tau < 0$.

It is well known that equation (3.1) admits a unique solution $x(\cdot) \in L^p_{\text{loc}}[-h, \infty; \mathbb{R}^n] \cap W^{1,p}_{\text{loc}}[0, \infty; \mathbb{R}^n]$ for every input $u(\cdot) \in L^p_{\text{loc}}[0, \infty; \mathbb{R}^l]$ and every initial condition of the form

$$x(0) = \phi^0, \quad x(\tau) = \phi^1(\tau), \quad -h \leqq \tau \leqq 0, \tag{3.2}$$

where $\phi = (\phi^0, \phi^1) \in \mathbb{R}^n \times L^p[-h, 0; \mathbb{R}^n] = M^p$. Moreover, in these spaces the solution $x(\cdot)$ of (3.1) and (3.2) depends continuously on ϕ and $u(\cdot)$. This has motivated the definition of the state of system (3.1) at time $t \geqq 0$ to be the pair

$$\hat{x}(t) = (x(t), x_t) \in M^p. \tag{3.3}$$

The evolution of $\hat{x}(t)$ can be described by the variation-of-constants formula

$$\hat{x}(t) = S(t)\phi + \int_0^t S(t-s)Bu(s)\, ds, \qquad t \geqq 0, \tag{3.4}$$

where $B \in \mathcal{L}(\mathbb{R}^l, M^p)$ maps $u \in \mathbb{R}^l$ into the pair $Bu = (B_0 u, 0)$ and $S(t) \in \mathcal{L}(M^p)$ is the strongly continuous semigroup generated by

$$\mathcal{D}(A) = \{\phi \in M^p \mid \phi^1 \in W^{1,p},\ \phi^1(0) = \phi^0\},$$
$$A\phi = (L\phi^1, \dot{\phi}^1).$$

Here $W^{1,p}$ denotes the Sobolev space $W^{1,p}[-h, 0, \mathbb{R}^n]$.

We will consider the evolution of the state (3.3) of system (3.1) in the dense subspace $\{(\phi(0), \phi) \mid \phi \in W^{1,p}\} \subset M^p$ which we shall identify with $W^{1,p}$. Then B becomes an "unbounded" operator ranging in the larger space M^p. However, it follows from the existence, uniqueness and continuous dependence result for the solutions of (3.1) and (3.2) that the state $\hat{x}(t)$ of (3.1), (3.2) defines a continuous function in $W^{1,p}$ provided that $\phi \in W^{1,p}$ and $u(\cdot) \in L^p_{\text{loc}}[0, \infty; \mathbb{R}^l]$. Hence, the operators A and B satisfy hypothesis (H1) with $Z = M^p$ and $X = W^{1,p}$. This implies that the state $\hat{x}(t) \in W^{1,p}$ of (3.1), (3.2) with $\phi \in W^{1,p}$ satisfies the Cauchy problem

$$\begin{aligned} &\frac{d}{dt}\hat{x}(t) = A\hat{x}(t) + Bu(t), \qquad \hat{x}(0) = \phi \in W^{1,p}, \\ &v(t) = C\hat{x}(t), \end{aligned} \tag{3.5}$$

in the sense of Remark 2.2(ii). Of course, the output operator $C: W^{1,p} \to \mathbb{R}^m$ is given by

$$C\phi = \int_{-h}^0 d\gamma(\tau)\phi^1(\tau), \qquad \phi \in W^{1,p}.$$

On the state space $W^{1,p}$ this operator is bounded.

Remarks 3.1. (i) If the equation

$$\eta(\tau) = \eta(-h) + A_1, \qquad -h < \tau \leqq \varepsilon - h, \tag{3.6}$$

holds for some $\varepsilon > 0$, then the generalized eigenfunctions of A are complete in M^p and in $W^{1,p}$ if and only if

$$\det A_1 \neq 0 \tag{3.7}$$

(see Manitius [9], Salamon [14, Chap. 3]).

(ii) It is well known that the operator A satisfies (H3).

(iii) The exponential growth of the semigroup $S(t)$ on the complementary subspace X^Λ corresponding to $\Lambda = \{\lambda \in \sigma(A) \mid \operatorname{Re} \lambda \geqq 0\}$ is determined by $\sup\{\operatorname{Re} \lambda \mid \lambda \in \sigma(A), \operatorname{Re} \lambda < 0\} < 0$.

(iv) Let (2.15) denote the reduced finite-dimensional system obtained by spectral projection of the solutions of (3.5) on the generalized eigenspace X_Λ. Then (2.15) is controllable iff

$$\operatorname{rank}[\lambda I - L(e^{\lambda \cdot}), B_0] = n \quad \forall \lambda \in \Lambda \tag{3.8}$$

(Pandolfi [10]) and observable iff

$$\operatorname{rank}\begin{bmatrix} \lambda I - L(e^{\lambda \cdot}) \\ C(e^{\lambda \cdot}) \end{bmatrix} = n \quad \forall \lambda \in \Lambda \tag{3.9}$$

(Bhat-Koivo [1], Salamon [13], [14]).

Combining these facts with Proposition 2.6 and Theorem 2.5, we obtain the following existence result for a finite dimensional compensator for system (3.1).

THEOREM 3.2. *If* (3.6)-(3.9) *are satisfied, then there exists a finite dimensional compensator of the form* (2.9) *such that the closed loop system* (3.1), (2.9) *is exponentially stable.*

3.2. Retarded systems with input delays. In this section we consider the RFDE

$$\begin{aligned} \dot{x}(t) &= Lx_t + Bu_t, \\ y(t) &= C_0 x(t), \end{aligned} \tag{3.10}$$

with general delays in the state and input and no delays in the output variable. This time C_0 is a real $m \times n$-matrix and B a bounded linear functional on $\mathscr{C}[-h, 0, \mathbb{R}^l]$ with values in $\mathbb{R}^n$ given by

$$B\xi = \int_{-h}^{0} d\beta(\tau)\xi(\tau), \qquad \xi \in \mathscr{C}[-h, 0, \mathbb{R}^l],$$

where $\beta(\tau)$ is an $n \times l$-matrix valued, normalized function of bounded variation. Of course, we can immediately get an existence result for a finite dimensional compensator for system (3.10) by dualizing Theorem 3.2. However, for reasons to become clear later, we make use of a direct approach for system (3.10), following the ideas of Vinter and Kwong [18] (see also Delfour [6], Salamon [14]).

First note that (3.10) admits a unique solution $x(\cdot) \in L^p_{\mathrm{loc}}[-h, \infty, \mathbb{R}^n] \cap W^{1,p}_{\mathrm{loc}}[0, \infty, \mathbb{R}^n]$ for every input $u(\cdot) \in L^p_{\mathrm{loc}}[0, \infty; \mathbb{R}^l]$ and every initial condition of the form

$$\begin{aligned} x(0) &= \phi^0, \qquad x(\tau) = \phi^1(\tau), \\ u(\tau) &= \xi(\tau), \qquad -h \leqq \tau < 0, \end{aligned} \tag{3.11}$$

where $\phi \in M^p$ and $\xi \in L^p[-h, 0; \mathbb{R}^l]$. In order to reformulate system (3.10) as an evolution equation in a product space, we rewrite (3.10)-(3.11) as

$$\begin{aligned} \dot{x}(t) &= \int_{-t}^{0} d\eta(\tau)x(t+\tau) + \int_{-t}^{0} d\beta(\tau)u(t+\tau) + f^1(-t), \\ y(t) &= C_0 x(t), \qquad x(0) = f^0, \end{aligned} \tag{3.12}$$

where the pair $f = (f^0, f^1) \in M^p$, given by

$$\begin{aligned} f^0 &= \phi^0, \\ f^1(\sigma) &= \int_{-h}^{\sigma} d\eta(\tau)\phi^1(\tau - \sigma) + \int_{-h}^{\sigma} d\beta(\tau)\xi(\tau - \sigma), \qquad -h \leqq \sigma \leqq 0, \end{aligned} \tag{3.13}$$

is regarded as the initial state of system (3.12). The corresponding state at time $t \geqq 0$ is given by

$$\begin{aligned} &\hat{x}(t) = (x(t), x^t) \in M^p, \\ &x^t(\sigma) = \int_{\sigma-t}^{\sigma} d\eta(\tau) x(t+\tau-\sigma) + \int_{\sigma-t}^{\sigma} d\beta(\tau) u(t+\tau-\sigma) + f^1(\sigma-t). \end{aligned} \tag{3.14}$$

It has been shown in [6], [14], [18] that the evolution of this state can be described by the variation-of-constants formula

$$\hat{x}(t) = S^{T^*}(t) f + \int_0^t S^{T^*}(t-s) B^{T^*} u(s)\, ds, \qquad t \geqq 0. \tag{3.15}$$

Here $S^{T^*}(t) \in \mathscr{L}(M^p)$ is the adjoint semigroup of $S^T(t) \in \mathscr{L}(M^q)$, $1/p + 1/q = 1$, which corresponds to the transposed equation $\dot{x}(t) = L^T x_t$ in the sense of § 3.1. Since $S^T(t)$ restricts to a semigroup on the dense subspace $W^{1,q} \subset M^q$, the adjoint semigroup $S^{T^*}(t)$ extends to the dual space $W^{-1,p} = (W^{1,q})^*$ which contains M^p as a dense subspace in a natural way. The input operator $B^{T^*} \in \mathscr{L}(\mathbb{R}^l, W^{-1,p})$ is the adjoint operator of $B^T \in \mathscr{L}(W^{1,q}, \mathbb{R}^l)$ given by

$$B^T \psi = \int_{-h}^{0} d\beta(\tau) \psi^1(\tau) \in \mathbb{R}^l, \qquad \psi \in W^{1,q} \subset M^q.$$

Since the infinitesimal generator A^{T^*} of $S^{T^*}(t)$ and the input operator B^{T^*} satisfy the hypothesis (H1) of § 2 with $X = M^p$ and $Z = W^{-1,p}$ (see Salamon [14]), the state $\hat{x}(t) \in M^p$ of system (3.12), given by (3.14), defines the unique solution of the abstract Cauchy problem

$$\begin{aligned} &\frac{d}{dt}\hat{x}(t) = A^{T^*}\hat{x}(t) + B^{T^*}u(t), \qquad \hat{x}(0) = f \in M^p, \\ &y(t) = C\hat{x}(t), \end{aligned} \tag{3.16}$$

in the sense of Remark 2.2(ii). Of course, the output operator $C: M^p \to \mathbb{R}^m$ is now given by

$$Cf = C_0 f^0 \in \mathbb{R}^m, \qquad f \in M^p.$$

In order to make the results of this section more precise, we briefly outline the construction of the reduced system (2.15). For this purpose let $X_\Lambda \subset W^{1,p}$ and $X_\Lambda^T \subset W^{1,q}$ denote the generalized eigenspaces of A and A^T, respectively, correspondig to $\Lambda = \{\lambda \in \sigma(A) | \operatorname{Re} \lambda \geqq 0\}$. Since Λ is a symmetric set, we can choose real bases $\{\phi_1, \cdots, \phi_{N_\Lambda}\}$ or X_Λ and $\{\psi_1, \cdots, \psi_{N_\Lambda}\}$ of X_Λ^T such that the matrices

$$\begin{aligned} \Phi &= [\phi_1 \cdots \phi_{N_\Lambda}] \in W^{1,p}[-h, 0; \mathbb{R}^{n \times N_\Lambda}], \\ \Psi &= [\psi_1 \cdots \psi_{N_\Lambda}] \in W^{1,q}[-h, 0; \mathbb{R}^{n \times N_\Lambda}], \end{aligned}$$

satisfy

$$\Psi^T(0)\Phi(0) + \int_{-h}^{0} \int_{\tau}^{0} \Psi^T(\sigma)\, d\eta(\tau) \Phi(\tau-\sigma)\, d\sigma = I \in \mathbb{R}^{N_\Lambda \times N_\Lambda}.$$

Then $\iota_\Lambda: \mathbb{R}^{N_\Lambda} \to M^p$ and $\pi_\Lambda: M^p \to \mathbb{R}^{N_\Lambda}$ may be defined by

$$[\iota_\Lambda x_\Lambda]^0 = \Phi(0) x_\Lambda, \quad [\iota_\Lambda x_\Lambda]^1(\sigma) = \int_{-h}^{\sigma} d\eta(\tau) \Phi(\tau-\sigma) x_\Lambda, \quad -h \leqq \sigma \leqq 0,$$

$$\pi_\Lambda f = \Psi^T(0) f^0 + \int_{-h}^{0} \Psi^T(\sigma) f^1(\sigma)\, d\sigma$$

for $x_\Lambda \in \mathbb{R}^{N_\Lambda}$ and $f \in M^p$, and the matrices $A_\Lambda \in \mathbb{R}^{N_\Lambda \times N_\Lambda}$, $B_\Lambda \in \mathbb{R}^{N_\Lambda \times l}$, $C_\Lambda \in \mathbb{R}^{m \times N_\Lambda}$ are given by

$$A(\Phi(0),\Phi) = (\Phi(0), \Phi)A_\Lambda, \quad B_\Lambda = \int_{-h}^{0} \Psi^T(\tau)\, d\beta(\tau), \quad C_\Lambda = C_0\Phi(0)$$

(see Salamon [14, § 2.4]).

Remarks 3.3. (i) If (3.6) is satisfied for some $\varepsilon > 0$, then the eigenfunctions of A^{T^*} are complete in M^P and in $W^{-1,p}$ if and only if (3.7) holds (see Manitius [9], Salamon [14, Chapter 3]).

(ii) If A_Λ, B_Λ, C_Λ are defined as above, then system (2.15) is controllable iff

$$\operatorname{rank}[\lambda I - L(e^{\lambda \cdot}), B(e^{\lambda \cdot})] = n \quad \forall \lambda \in \Lambda \tag{3.17}$$

and observable iff

$$\operatorname{rank}\begin{bmatrix} \lambda I - L(e^{\lambda \cdot}) \\ C_0 \end{bmatrix} = n \quad \forall \lambda \in \Lambda \tag{3.18}$$

(see Salamon [14]).

THEOREM 3.4. *If* (3.6)-(3.7) *and* (3.17)-(3.18) *are satisfied, then there exists a finite-dimensional compensator of the form* (2.9), *such that the closed loop system* (3.10), (2.9) *is exponentially stable.*

Remark 3.1(i) and Remark 3.3(i) show that the completeness property of A and A^{T^*} can be destroyed by arbitrarily small perturbations in the delays (compare Manitius [9]). However such perturbations would not affect the stability of the closed loop system (3.1), (2.9) respectively (3.10), (2.9). This indicates that the completeness assumption is somewhat artifical for the purpose of stabilization by a finite-dimensional compensator. This assumption can be weakened slightly in the special case of the RFDE

$$\begin{aligned} \dot{x}(t) &= A_0x(t) + A_1x(t-h) + B_0u(t), \\ y(t) &= C_0x(t), \end{aligned} \tag{3.19}$$

with a single point delay in the state variable if the state space is chosen to be

$$X = (f \in M^p \mid f^1(\tau) \in \text{range } A_1, -h \leqq \tau \leqq 0\}.$$

It has been shown by Manitius [9] that the completeness property for the operator A^{T^*} in this space is equivalent to the rank condition

$$\operatorname{rank}\begin{bmatrix} A_0 - \lambda I & A_1 \\ A_1 & 0 \end{bmatrix} = n + \operatorname{rank} A_1 \tag{3.20}$$

for some $\lambda \in \mathbb{C}$. Therefore we have the following result.

COROLLARY 3.5. *If* (3.8), (3.18) *and* (3.20) *are satisfied then there exists a finite dimensional compensator of the form* (2.9) *such that the closed loop system* (3.19), (2.9) *is exponentially stable.*

Remark 3.6. This result suggests that it should be possible to weaken the completeness assumption for the general RFDE which would be an important improvement. Another extension would be an existence result for RFDEs with delays in simultaneously the control and the observation. However, it is not obvious how this can be achieved with the present approach, the main difficulty being the completeness property.

Remark 3.7. Although the main results of this section, Theorems 3.2 and 3.4 are stated as existence results, we emphasize that the stabilizing compensator can in fact be constructed using exactly the same procedure as it is explained in detail in [16] for

RFDEs without delays in the external variables. The construction, as outlined in the proof of Proposition 2.6, involves calculating finitely many eigenvalues and eigenvectors of A and hence the projected system A_Λ, B_Λ, C_Λ as it is described in § 3.2. Then the matrices F_Λ and G_Λ can be calculated by standard finite dimensional procedures. The most difficult part of the design lies in finding eigenvectors of $A+BF$ to generate the subspace W. The approximation of $G=\iota_\Lambda G_\Lambda$ by an operator $\hat{G}$ with range in W then reduces to a finite-dimensional linear optimization procedure. This procedure has to be repeated—while increasing W—until $\hat{G}$ is close enough to G. The numerical example for a retarded system examined in [16] gives insight into the details of the design procedure.

4. Boundary control systems. The purpose of this section is to show how abstract boundary control systems in Hilbert spaces fit into the framework of § 2. When these results are applied to obtain finite-dimensional compensators for particular classes of partial differential equations (PDE), there is a considerable overlap with results of Curtain in [2], [3], [4]. The relation between both approaches will be discussed in detail at the end of this section.

Let W, X, U, Y be Hilbert spaces and suppose that

$$W \subset X$$

with a continuous, dense injection. Furthermore, let $\Delta \in \mathscr{L}(W, X)$, $\Gamma \in \mathscr{L}(W, U)$, $C \in \mathscr{L}(X, Y)$ be given. Then we consider the boundary control system

$$\begin{aligned} &\frac{d}{dt}x(t)=\Delta x(t), \qquad x(0)=x_0 \in W, \\ &\Gamma x(t)=u(t), \qquad t \geqq 0, \end{aligned} \tag{4.1}$$

with the output

$$y(t)=Cx(t), \qquad t \geqq 0. \tag{4.2}$$

DEFINITION 4.1 (*strong solution, well-posedness*).

(i) Let $u(\cdot) \in \mathscr{C}[0, T; U]$ and $x_0 \in W$ satisfy $\Gamma x_0 = u(0)$. Then a function $x(\cdot) \in \mathscr{C}[0, T; W]$ is said to be a (strong) solution of (4.1) if $x(\cdot) \in \mathscr{C}^1[0, T; X]$ and if (4.1) is satisfied for every $t \in [0, T]$.

(ii) The boundary control system (4.1) is said to be well-posed if the subspace $\{x \in W | \Gamma x = 0\}$ is dense in X, if the restriction of Λ to this subspace is a closed operator on X, and if for all $x_0 \in W$ and $u(\cdot) \in W^{1,2}[0, T; U]$ with $\Gamma x_0 = u(0)$ there exists a unique solution $x(\cdot) \in \mathscr{C}[0, t; W] \cap \mathscr{C}^1[0, T; X]$ of (4.1) depending continuously on x_0 and $u(\cdot)$. This means that there exists a constant $K > 0$ such that the inequality

$$\sup_{0 \leqq t \leqq T} \|x(t)\|_W + \sup_{0 \leqq t \leqq T} \|\dot{x}(t)\|_X \leqq K \left\{ \|x_0\|_W + \left[\int_0^T \|\dot{u}(t)\|_U^2 dt \right]^{1/2} \right\}$$

holds for every solution $x(t)$ of (4.1).

Remarks 4.2. Let system (4.1) be well posed.

(i) Taking $u(t) \equiv 0$, it follows from a classical result in semigroup theory (Phillips [11]) that the operator

$$Ax = \Delta x, \qquad \mathscr{D}(A) = \{x \in W | \Gamma x = 0\} \tag{4.3}$$

is the infinitesimal generator of a strongly continuous semigroup $S(t)$ on X and that, for every $x_0 \in \mathscr{D}(A)$, the function $x(t) = S(t)x_0$ is the solution of (4.1) with $u(t) \equiv 0$.

(ii) As in § 2 we introduce the dense subspace $Z^* = \mathcal{D}_{X^*}(A^*) \subset X^*$. Then

$$X \subset Z$$

with a continuous, dense injection, A extends to a bounded operator from X to Z and $S(t)$ to a strongly continuous semigroup on Z.

(iii) It follows from [14, Lemma 1.3.2(i)] that

$$\{x \in W | \Gamma x = 0\} = \mathcal{D}_X(A) = \{x \in X = \mathcal{D}_Z(A) | Ax \in X\}$$

or in other words, if $x \in X$ and $Ax \in X$, then $x \in W$ and $\Gamma x = 0$. Furthermore, the W-norm on $\mathcal{D}_X(A)$ is equivalent to the graph norm of A [14, Remark 1.3.1(iii)]. This means that there exists a constant $K_1 > 0$ such that the inequality

$$\|x\|_W \leqq K_1[\|x\|_X + \|Ax\|_X]$$

holds for all $x \in W$ with $\Gamma x = 0$.

(iv) Γ is onto. Hence there exists a constant $K_0 > 0$ such that for every $u \in U$ there exists a $w \in W$ such that

$$\Gamma w = u, \qquad \|w\|_W \leqq K_0 \|u\|_U. \tag{4.4}$$

Let us now construct the input operator $B \in \mathcal{L}(U, Z)$.

1. Given $u \in U$ we may choose $w \in W$ such that $\Gamma w = u$ since Γ is onto (Remark 4.2(iv)). For this $w \in W$ we define $Bu := \Delta w - Aw \in Z$. This expression is well defined since $\Gamma w = 0$ if and only if $Aw = \Delta w$ (Remark 4.2(iii)). Hence the map $B: U \to Z$ satisfies, by definition, the equation

$$B\Gamma x = \Delta x - Ax, \qquad x \in W. \tag{4.5}$$

2. It is easy to see that B is a linear map.

3. Let $u \in U$ be given and choose $w \in W$ such that (4.4) holds. Then

$$\|Bu\|_Z \leqq \|\Delta - A\|_{\mathcal{L}(W,Z)} \|w\|_W \leqq K_0 \|\Delta - A\|_{\mathcal{L}(W,Z)} \|u\|_U$$

and therefore $B: U \to Z$ is bounded.

LEMMA 4.3. *Let the operators A and B be defined by* (4.3) *and* (4.5), *respectively. Furthermore, let $x \in X$ and $u \in U$ satisfy $Ax + Bu \in X$. Then*

$$x \in W, \quad \Gamma x = u, \quad \Delta x = Ax + Bu.$$

Furthermore there exists a constant $K > 0$ such that

$$\|x\|_W \leqq K[\|x\|_X + \|u\|_U + \|Ax + Bu\|_X]$$

for all $x \in X$, $u \in U$ with $Ax + Bu \in X$.

Proof. Let $x \in X$ and $u \in U$ satisfy $Ax + Bu \in X$ and choose $w \in W$ such that (4.4) holds. Then

$$A(x - w) = Ax + Bu - (A + B\Gamma)w = Ax + Bu - \Delta w \in X.$$

By Remark 4.2(iii), this implies that $x \in W$ and

$$\begin{aligned}
\|x\|_W &\leqq \|w\|_W + \|x - w\|_W \\
&\leqq \|w\|_W + K_1[\|x - w\|_X + \|A(x - w)\|_X] \\
&\leqq [1 + K_1\|\mathrm{id}\|_{\mathcal{L}(W,X)} + K_1\|\Delta\|_{\mathcal{L}(W,X)}]\|w\|_W \\
&\quad + K_1\|x\|_X + K_1\|Ax + Bu\|_X \\
&\leqq K[\|u\|_U + \|x\|_X + \|Ax + Bu\|_X].
\end{aligned}$$

Finally, we obtain again from Remark 4.2(iii) that $\Gamma x = \Gamma w = u$ and from (4.5) that $\Delta x = Ax + B\Gamma x = Ax + Bu$. □

The above operators A and B allow us to reformulate the boundary control system (4.1) as a Cauchy problem of the type (3.1). More precisely, we introduce the following concept of a weak solution for (4.1).

DEFINITION 4.4 (*weak solution*). Let the operators $A \in \mathscr{L}(X, Z)$ and $B \in \mathscr{L}(U, Z)$ be defined as above. Moreover, let $x_0 \in X$ and $u(\cdot) \in L^2[0, T, U]$ be given. Then $x(\cdot) \in \mathscr{C}[0, T; X] \cap W^{1,2}[0, T; Z]$ is said to be a weak solution of (4.1) if

$$\begin{aligned} &\frac{d}{dt} x(t) = Ax(t) + Bu(t), \qquad 0 \leqq t \leqq T, \\ &x(0) = x_0 \end{aligned} \tag{4.6}$$

is satisfied in Z (almost everywhere).

It follows from the definition of the operator B (Remark 4.2(iv)) that every strong solution $x(\cdot) \in \mathscr{C}[0, T; W] \cap \mathscr{C}^1[0, T; X]$ of (4.1) is a weak solution in the sense of Definition 4.4. Moreover we have the following result.

PROPOSITION 4.5. *Suppose that the operator A defined by* (4.3) *is the infinitesimal generator of a strongly continuous semigroup $S(t)$ on X and that $\Gamma \in \mathscr{L}(W, U)$ is onto. Furthermore let $B \in \mathscr{L}(U, Z)$ be defined by* (4.5). *Then the following statements are equivalent.*

(i) *System* (4.1) *is well posed.*

(ii) *The operators A and B satisfy hypothesis* (H1) *of* § 2 *with $p = 2$.*

(iii) *For every $x_0 \in X$ and every $u(\cdot) \in L^2[0, T; U]$ there exists a unique weak solution $x(\cdot) \in \mathscr{C}[0, T; X] \cap W^{1,2}[0, T; Z]$ of* (4.1) *depending continuously on x_0 and $u(\cdot)$. Moreover, the weak (and in particular the strong) solutions of* (4.1) *are given by*

$$x(t) = S(t)x_0 + \int_0^t S(t-s)Bu(s)\, ds \in X, \qquad 0 \leqq t \leqq T. \tag{4.7}$$

Proof. It is a well-known semigroup theoretic result that the solutions of (4.6), and therefore the weak solutions of (4.1), are given by (4.7). Furthermore, it follows from Remark 2.2(ii) that (ii) is equivalent to (iii).

In order to prove that (ii) implies (i), suppose that (H1) is satisfied and let $x(t)$ be given by (4.7) with $x_0 \in W$ and $u(\cdot) \in \mathscr{C}^1[0, T; U]$ satisfying $\Gamma x_0 = u(0)$. Then it is a well-known result from semigroup theory that $x(\cdot) \in \mathscr{C}[0, T; X] \cap \mathscr{C}^1[0, T; Z]$ satisfies

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t) \\ &= S(t)[Ax_0 + Bu(0)] + \int_0^t S(t-s)B\dot{u}(s)\, ds \\ &= S(t)\Delta x_0 + \int_0^t S(t-s)B\dot{u}(s)\, ds, \qquad 0 \leqq t \leqq T. \end{aligned}$$

By (H1) and Remark 2.1(ii), this implies that $\dot{x}(\cdot) \in \mathscr{C}[0, T; X]$ and

$$\sup_{0 \leqq t \leqq T} \|\dot{x}(t)\|_X < \sup_{0 \leqq t \leqq T} \|S(t)\|_{\mathscr{L}(X)} \|\Delta\|_{\mathscr{L}(W,X)} \|x_0\|_W + b_T \|\dot{u}(\cdot)\|_{L^2[0,T;U]}.$$

Applying Lemma 4.3 to the term $Ax(t) + Bu(t) = \dot{x}(t) \in X$, we obtain that $x(\cdot) \in \mathscr{C}[0, T; W] \cap \mathscr{C}^1[0, T; X]$ satisfies (4.1). Since every strong solution of (4.1) is given

by (4.7), $x(t)$ is in fact the unique solution. Furthermore, we obtain from Lemma 4.3 that

$$\|x(t)\|_W \leqq K[\|x(t)\|_X + \|u(t)\|_U + \|\dot{x}(t)\|_X].$$

Since

$$\sup_{0\leqq t\leqq T} \|x(t)\|_X \leqq \sup_{0\leqq t\leqq T} \|S(t)\|_{\mathscr{L}(X)}\|\mathrm{id}\|_{\mathscr{L}(W,X)}\|x_0\|_W + b_T\|u(\cdot)\|_{L^2[0,T;U]}$$

and

$$\sup_{0\leqq t\leqq T} \|u(t)\|_U \leqq \|\Gamma\|_{\mathscr{L}(W,U)}\|x_0\|_W + \sqrt{T}\|\dot{u}(\cdot)\|_{\mathscr{L}^2[0,T;U]}$$

for $u(\cdot)\in \mathscr{C}^1[0, T; U]$ with $u(0)=\Gamma x_0$, this shows that system (4.1) is well posed.

Conversely, suppose that system (4.1) is well posed in the sense of Definition 4.1, let $v(\cdot)\in \mathscr{C}^1[0, T; U]$ and define

$$x(t)=\int_0^t S(t-s)B\int_0^s v(\tau)\,d\tau\,ds, \qquad u(t)=\int_0^t v(\tau)\,d\tau, \qquad 0\leqq t\leqq T.$$

Then $x(\cdot)\in C^1[0, T; Z]$ and

$$Ax(t)+Bu(t)=\dot{x}(t)=\int_0^t S(t-s)Bv(s)\,ds, \qquad 0\leqq t\leqq T.$$

Hence $x(\cdot)\in \mathscr{C}^1[0, T; X]$ and we obtain from Lemma 4.3 that $x(\cdot)\in \mathscr{C}[0, T; W]$, $\Gamma x(t)=u(t)$ and $\Delta x(t)=Ax(t)+Bu(t)$. Hence $x(t)$ is a strong solution of (4.1) in the sense of Definition 4.1(i) and satisfies the inequality

$$\left\|\int_0^T S(T-s)Bv(s)\,ds\right\|_X = \|\dot{x}(T)\|_X \leqq K\|v(\cdot)\|_{L^2[0,T;U]}.$$

This shows that the operators A and B satisfy the hypothesis (H1). □

Having established hypothesis (H1) we are now in a position to apply the perturbation result of § 2 (Theorem 2.2) to the boundary control system (4.1).

COROLLARY 4.6. *If system* (4.1) *is well posed, then the following statements hold.*

(i) *For every* $F\in \mathscr{L}(X, U)$ *the operator*

$$A_F x=\Delta x, \qquad \mathscr{D}(A_F)=\{x\in W \mid \Gamma x = Fx\} \tag{4.8}$$

is the infinitesimal generator of a strongly continuous semigroup $S_F(t)$ *on* X.

(ii) *For every* $x_0\in \mathscr{D}(A_F)$ *the function* $x(t)=S_F(t)x_0$, $t\geqq 0$, *is continuous in* W, *continuously differentiable in* X, *and satisfies the closed loop boundary control equations*

$$\begin{aligned} &\frac{d}{dt}x(t)=\Delta x(t), \qquad x(0)=x_0,\\ &\Gamma x(t)=Fx(t), \qquad t\geqq 0, \end{aligned} \tag{4.9}$$

where the derivative has to be understood in the space X.

(iii) *If* U *is finite dimensional, then* $S_F(t)$ *extends to a strongly continuous semigroup on* Z *whose infinitesimal generator is given by the extended operator* $A+BF: X\to Z$.

Proof. By Proposition 4.5, the operators A and B defined by (4.3) and (4.5), respectively, satisfy hypothesis (H1) of § 2. Hence it follows from Theorem 2.3(i) that the operator

$$A_F x=Ax+BFx, \qquad \mathscr{D}(A_F)=\{x\in X \mid Ax+BFx\in X\}$$

generates a strongly continuous semigroup of X (note that the proof of this result in

[14, Thm. 1.3.7] does not require U to be finite dimensional). Lemma 4.3 shows that this operator A_F coincides with the one defined by (4.8). This proves statement (i).

In order to prove statement (ii), let $x_0 \in \mathcal{D}(A_F)$ be given and define $u(t) = FS_F(t)x_0$, $t \geqq 0$. Then $u(t)$ is continuously differentiable for $t \geqq 0$ and satisfies $u(0) = Fx_0 = \Gamma x_0$. Hence (4.1) admits a unique strong solution $x(t)$, $t \geqq 0$, which by definition of the operators A and B also satisfies (4.6) and is therefore given by (4.7). This implies

$$x(t) = S(t)x_0 + \int_0^t S(t-s)BFS_F(s)x_0\, ds = S_F(t)x_0,$$

by definition of the semigroup $S_F(t)$.

Statement (iii) is an immediate consequence of Theorem 2.3(iii). □

So far we have shown that the general theory of § 2 also covers abstract boundary control systems. In particular, we have reformulated the boundary control system (4.1) in the semigroup theoretic framework with an unbounded input operator. A very similar approach has been developed by Ho and Russell in [8] under only slightly more restrictive assumptions. However, [8] does not contain any feedback results and also the above Proposition 4.5 seems to be new. Furthermore we point out that earlier results in this direction for various classes of partial differential equations can be found, e.g., in the classical work by Lions-Magenes [22], in the more recent papers by Washburn [19], Lasiecka-Triggiani [20], [21] and in the book by Curtain-Pritchard [5] (this list is by no means complete). Another general approach has been presented by Fattorini [7]. In [7] the input operator is bounded, however, there are derivatives in the input function which do not appear in our approach.

In [2] and [3] Curtain has used Fattorini's results for the construction of finite dimensional compensators which leads to integral terms in the loop. These integral terms will disappear if we apply the approach of this section to obtain existence results for finite dimensional compensators. More precisely, we have to assume that the operators A, B and C, introduced in this section, satisfy hypothesis (H2) of § 2, or respectively, hypothesis (H3) and the assumptions of Proposition 2.6. Under these conditions it follows readily from Theorem 2.5 that there exists a finite dimensional compensator of the form (2.9) such that the closed loop system (4.1), (4.2), (2.9) is exponentially stable.

Starting from (4.1), (4.2), the following problems have to be solved for the construction of the compensator.

1. Find the operators A and B.
2. Determine the spectrum of the operator A and the reduced subsystem (2.15).
3. Find the stabilizing operators $F: X \to U$ and $G: Y \to X$.
4. Determine the eigenvalues and eigenvectors of A_F to approximate G.

To illustrate this procedure, we consider the heat equation with Neumann boundary conditions and boundary control which has also been treated in [3] with different methods.

Example 4.7. Consider the parabolic PDE

$$z_t = \pi^{-2} z_{\xi\xi}, \qquad 0 < \xi < 1, \quad t > 0, \tag{4.10.1}$$

$$z_\xi(0, t) = u(t), \qquad z_\xi(1, t) = 0, \qquad t > 0, \tag{4.10.2}$$

$$z(\xi, 0) = z_0(\xi), \qquad 0 < \xi < 1, \tag{4.10.3}$$

$$y(t) = \int_0^1 c(\xi) z(\xi, t)\, d\xi, \qquad t > 0. \tag{4.10.4}$$

This system can be written in the abstract form (4.1) with

$$X = L^2[0,1], \quad W = \{\phi \in H^2[0,1] | \dot{\phi}(1) = 0\}, \quad U = \mathbb{R},$$
$$\Delta\phi = \pi^{-2}\ddot{\phi}, \qquad \Gamma\phi = \dot{\phi}(0),$$
$$Z^* = \mathcal{D}(A^*) = \mathcal{D}(A) = \{\psi \in H^2[0,1] | \dot{\psi}(0) = 0 = \dot{\psi}(1)\}.$$

The operator A satisfies (H3) and has a complete set of eigenvectors $\phi_0(\xi) \equiv 1$, $\phi_n(\xi) = \sqrt{2}\cos n\pi\xi$, corresponding to the eigenvalues $\lambda_0 = 0$, $\lambda_n = -n^2$, $n \in \mathbb{N}$. In order to determine the operator $B: \mathbb{R} \to Z$, let us choose any $\phi \in W$ such that $\Gamma\phi = 1$, e.g. $\phi(\xi) = -(\xi-1)^2/2$. Then, for every $\psi \in Z^*$, the following equation holds

$$\begin{aligned} B^*\psi &= \langle B^*\psi, \Gamma\phi\rangle = \langle \psi, B\Gamma\phi\rangle = \langle \psi, \Delta\phi - A\phi\rangle_{Z^*,Z} \\ &= \langle \psi, \Delta\phi\rangle_H - \langle A^*\psi, \phi\rangle_H \\ &= \frac{1}{\pi^2}\int_0^1 \psi(\xi)\ddot{\phi}(\xi)\,d\xi - \frac{1}{\pi^2}\int_0^1 \ddot{\psi}(\xi)\phi(\xi)\,d\xi \\ &= \frac{1}{\pi^2}\int_0^1 [\psi(\xi)\ddot{\phi}(\xi) + \dot{\psi}(\xi)\dot{\phi}(\xi)]\,d\xi \\ &= \frac{1}{\pi^2}[\psi(1)\dot{\phi}(1) - \psi(0)\dot{\phi}(0)] \\ &= -\frac{1}{\pi^2}\psi(0). \end{aligned}$$

It has been shown in Pritchard-Salamon [12] that these operators A and B satisfy hypothesis (H1) and therefore system (4.10) is well posed in $X = L^2[0,1]$ in the sense of Definition 4.1 (see Proposition 4.5). The spectral projection of $L^2[0,1]$ onto the eigenspace $X_\Lambda = \{\alpha\phi_0 | \alpha \in \mathbb{R}\}$ of A corresponding to the unstable part $\Lambda = \{0\}$ of the spectrum is given by

$$P_\Lambda\phi(\xi) = \int_0^1 \phi(\tau)\,d\tau, \qquad 0 < \xi < 1.$$

With the choice of $\{\phi_0\}$ as a basis of X_Λ, this operator splits into $P_\Lambda = \iota_\Lambda\pi_\Lambda$, where $\pi_\Lambda: L^2[0,1] \to R$ and $\iota_\Lambda: \mathbb{R} \to L^2[0,1]$ are given by

$$\pi_\Lambda\phi = \int_0^1 \phi(\tau)\,d\tau, \qquad \iota_\Lambda x_\Lambda(\xi) = x_\Lambda, \qquad 0 \leqq \xi \leqq 1.$$

Then the reduced finite dimensional system (2.15) is described by the "matrices"

$$A_\Lambda = 0, \quad B_\Lambda = -\pi^{-2}, \quad C_\Lambda = \int_0^1 c(\xi)\,d\xi.$$

This sytem is controllable and obervable if and only if

$$C_\Lambda = \int_0^1 c(\xi)\,d\xi \neq 0. \tag{4.11}$$

Stabilizing matrices are given e.g. by

$$F_\Lambda = \frac{\pi^2}{4}, \qquad G_\Lambda = -C_\Lambda^{-1}$$

so that $A_\Lambda + B_\Lambda F_\Lambda = -\frac{1}{4}$, $A_\Lambda + G_\Lambda C_\Lambda = -1$. Then the operator A_F with $F = F_\Lambda \pi_\Lambda$: $L^2[0,1] \to \mathbb{R}$ given by

$$\mathcal{D}(A_F) = \left\{ \phi \in H^2[0,1] \,|\, \dot{\phi}(1) = 0,\ \dot{\phi}(0) = \frac{\pi^2}{4} \int_0^1 \phi(\xi)\, d\xi \right\},$$

$$A_F \phi = \frac{1}{\pi^2} \ddot{\phi}.$$

The eigenvectors and eigenvalues of A_F coincide with those of A except for $\lambda_0 = 0$ which is now replaced by $\lambda_F = -\frac{1}{4}$. The corresponding normalized eigenfunction is

$$\phi_F(\xi) = \sqrt{2} \sin \frac{\pi}{2} \xi.$$

We will choose $W = \text{span}\{\phi_F\}$ and the maps

$$(\iota_F w)(\xi) = \phi_F(\xi) w, \qquad 0 < \xi < 1, \quad w \in \mathbb{R},$$

$$(\pi_F \phi)(\xi) = \int_0^1 \phi_F(\xi) \phi(\xi)\, d\xi, \qquad \phi \in L^2[0,1],$$

So that $\iota_F \pi_F : L^2[0,1] \to W$ is the orthogonal projection onto W and $\pi_F \iota_F = 1$.

Let us now consider the case that $c(\xi) = \xi$ for $0 \leqq \xi \leqq 1$. Then $C_\Lambda = \frac{1}{2}$ and we choose $G_\Lambda = -\pi g / 2\sqrt{2}$, $g > 0$. With this choice the operator $G : \mathbb{R} \to L^2[0,1]$ is given by

$$[Gy](\xi) = [\iota_\Lambda G_\Lambda y](\xi) = -\frac{\pi g}{2\sqrt{2}} y, \qquad 0 \leqq \xi \leqq 1.$$

We replace this operator by

$$[\hat{G}y](\xi) = [\iota_F \pi_F G y](\xi) = -gy\sqrt{2} \sin \frac{\pi}{2} \xi, \qquad 0 \leqq \xi \leqq 1,$$

whose range is in W. Since the perturbed operator $A + \hat{G}C$ generates an analytic semigroup it satisfies the "spectrum determined growth" assumption. Furthermore, its spectrum is given by

$$\sigma(A + \hat{G}C) = \{-\omega^2 \,|\, g[1 - \cos \omega\pi] = \omega^3 [8K + \pi^2/2\sqrt{2} - \sqrt{2}\pi^2\omega^2] \sin \omega\pi,\ \omega \neq 0\}$$

if $g > 0$. In the case $g < 0$ there is an additional positive eigenvalue $\lambda_0 = \omega^2 > 0$ where

$$-g \frac{e^{\omega\pi} + e^{-\omega\pi} - 2}{e^{\omega\pi} - e^{-\omega\pi}} = \omega^3 [8K + \pi^2/2\sqrt{2} + \sqrt{2}\pi^2\omega^2].$$

We conclude that $A + \hat{G}C$ generates an exponentially stable semigroup if and only if $g > 0$. Hence the operators F and $\hat{G}$ satisfy the hypothesis (H2) with the one-dimensional subspace $W = \text{span}\{\phi_F\}$. In this case the compensator (2.9) is described by the "matrices"

$$M = \pi_F (A_F + \hat{G}C) \iota_F = -\frac{1}{4} - \frac{4\sqrt{2}}{\pi^2} g,$$

$$H = \pi_F \hat{G} = -g,$$

$$K = F_\Lambda \pi_\Lambda \iota_F = \frac{\pi}{\sqrt{2}}.$$

Hence the first order system

$$\dot{w} = -\frac{1}{4} w - g\left[\frac{4\sqrt{2}}{\pi^2} w - y\right],$$
$$u = \frac{\pi}{\sqrt{2}} w, \tag{4.12}$$

defines a stablizing compensator for the parabolic PDE (4.10) with $c(\xi) \equiv \xi$ if and only if $g > 0$.

Remark 4.8. The results of this section show that the abstract framework of § 2 is general enough to cover both FDEs and PDEs. We mention that the approach of this section can also be applied to damped hyperbolic systems. Hence this paper represents a complete generalization of the compensator design of Schumacher [16] to infinite dimensional systems with unbounded control action. However, the degree of unboundedness which we can allow for the input/output operators is not as general as one would desire. For example, for the parabolic PDE (4.10) we cannot allow simultaneously Neumann boundary control and point observation. Also we cannot allow Dirichlet boundary control when the output operator is an arbitrary functional on $L^2[0, 1]$. A general theory which covers these cases would require the consideration of *unbounded output operators* as well. The extension of our theory to this case seems to involve some further difficulties and would be an interesting problem for future investigations.

Remark 4.9. Using the abstract approach outlined in § 2, it is possible to directly extend the results of Schumacher [15], [17] on tracking and regulation in infinite dimensions to unbounded control action. A different approach is to use the extended system formulation discussed in [2], which results in integral control action and this can be found in Curtain [4].

Note. Stronger results on finite dimensional compensators for some classes of functional differential equations have recently been developed by Kamen-Khargonekar-Tannenbaum [23], Nett [24], Logemann [25] using frequency domain methods.

REFERENCES

[1] K. P. M. Bhat and H. N. Koivo, *Modal characterization of controllability and observability for time delay systems.* IEEE Trans. Automat. Control, AC-21 (1976), pp. 292-293.

[2] R. F. Curtain, *On semigroup formulations of unbounded observations and control action for distributed systems*, International MTNS-Symposium, 1983, Beer-Sheva, Israel.

[3] ———, *Finite dimensional compensators for parabolic distributed systems with unbounded control and observation*, this Journal, 22 (1984), pp. 255-276.

[4] ———, *Tracking and regulation for distributed parameter systems with unbounded observation and control*, Mathematica Aplicada e Computacional, 2 (1983), pp. 199-218.

[5] R. F. Curtain and A. J. Pritchard, *Infinite Dimensional Linear Systems Theory*, Lecture Notes in Computer and Information Sciences 8, Springer-Verlag, Berlin 1978.

[6] M. C. Delfour, *The linear quadratic optimal control problem with delays in state and control variables: a state space approach*, Centre de Recherche de Mathématiques Appliquées, Université de Montréal, CRMA-1012, 1981.

[7] H. O. Fattorini, *Boundary control systems*, this Journal, 6 (1968), pp. 349-385.

[8] L. F. Ho and D. L. Russell, *Admissable input elements for systems in Hilbert space and a Carleson measure criterion*, this Journal, 21 (1983), pp. 614-640.

[9] A. Manitius, *Completeness and F-completeness of eigenfunctions associate with retarded functional differential equations*, J. Differential Equations, 35 (1980), pp. 1-29.

[10] L. PANDOLFI, *Feedback stabilization of functional differential equations*, Boll. Un. Mat. Ital., 12 (1975), pp. 626-635.

[11] R. S. PHILLIPS, *A note on the abstract Cauchy problem*, Proc. Nat. Acad. Science U.S.A., 40 (1954), pp. 244-248.

[12] A. J. PRITCHARD AND D. SALAMON, *The linear quadratic optimal control problem for infinite dimensional systems with unbounded input and output operators*, MRC, University of Wisconsin-Madison, TSR-2624, 1984.

[13] D. SALAMON, *Observers and duality between dynamic observation and state feedback for time delay systems*, IEEE Trans. Automat. Control, AC-25 (1980), pp. 1187-1192.

[14] ———, *Control and Observation of Neutral Systems*, RNM 91, Pitman, London 1984.

[15] J. M. SCHUMACHER, *Dynamic feedback in finite- and infinite-dimensional space*. Mathematical Centre Tracts, No. 143, Mathematisch Centrum, Amsterdam, 1981.

[16] ———, *A direct approach to compensator design for distributed parameter systems*, this Journal, 21 (1983), pp. 823-836.

[17] ———, *Finite dimensional regulators for a class of infinite dimensional systems*, Systems Control Lett., 3 (1983), pp. 7-12.

[18] R. B. VINTER AND R. H. KWONG, *The infinite time quadratic control problem for linear systems with state and control delays: an evolution equation approach*, this Journal, 19 (1981), pp. 139-153.

[19] D. WASHBURN, *A bound on the boundary input map for parabolic equations with applications to time optimal control*, this Journal, 17 (1979), pp. 652-671.

[20] I. LASIECKA AND R. TRIGGIANI, *A cosine operator approach for modelling $L^2[0, \tau; L^2(\Gamma)]$-boundary input hyperbolic equations*, Appl. Math. Optim., 7 (1981), pp. 35-93.

[21] ———, *Regularity of hyperbolic equations under $L^2[0, T; L^2(\Gamma)]$-Dirichlet boundary terms*, this Journal, to appear.

[22] J. L. LIONS AND E. MAGENES, *Nonhomogeneous Boundary Value Problems and Applications, Vols.* I & II, Springer-Verlag, New York, 1972.

[23] E. W. KAMEN, P. P. KHARGONEKAR AND A. TANNENBAUM, *Stabilization of time delay systems using finite dimensional compensators*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 75-78.

[24] C. N. NETT, *The fraction representation approach to robust linear feedback design*, Ph.D. thesis, Dept. Electrical, Computer and Systems Engineering, Renssalaer Polytechnic Institute, Troy, NY 1984.

[25] H. LOGEMANN, *Finite dimensional stabilization of infinite dimensional systems: A frequency domain approach*, FS Dynamische Systeme, Univ. Bremen, Report Nr. 124, 1984.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 4, July 1986

MINIMAL ORDER ESTIMATION OF MULTIVARIABLE DISCRETE-TIME STOCHASTIC LINEAR SYSTEMS*

YORAM BARAM† AND URI SHAKED‡

Abstract. Minimal order estimation of the state of a discrete-time, multi-input, multi-output stochastic linear system is considered. The relationship between the given system structure and the order of the minimal estimator is investigated and a necessary and sufficient condition for order minimality is derived in terms of the internal system structure.

Key words. stochastic systems, estimation, filtering

AMS(MOS) subject classifications. 93E03, 93E10, 93E11

1. Introduction. This paper is concerned with the minimal order of the state estimator of a given state-space system. Order minimality of the estimator implies nonexistence of a lower order state estimator for the system. It is well known that an order reduction can be obtained when the observation noise is colored or when it is white with a singular covariance matrix [1]. In this paper we show that an order reduction may be obtained due to the internal structure of the system, even when the observation noise is white with a nonsingular covariance. In fact, we assume a nonsingular observation noise covariance and obtain a necessary and sufficient condition for the minimality of the state estimator in terms of the internal system structure and parameters.

Consider the system

$$x_{n+1} = Ax_n + Bw_n, \qquad x_n \in R^p, \tag{1.1a}$$

$$y_n = Cx_n + u_n, \qquad y_n \in R^q, \tag{1.1b}$$

with

$$E\{x_0\} = E\{w_n\} = E\{u_n\} = 0, \qquad E\{w_n x_0^T\} = E\{u_n x_0^T\} = 0,$$

and

$$E\{w_n w_k^T\} = Q\delta_{n,k}, \qquad E\{u_n u_k\} = P\delta_{n,k},$$

where

$$Q > 0, \qquad P > 0 \quad \text{and} \quad \delta_{n,k} = \begin{cases} 1 & n = k \\ 0 & n \neq k \end{cases}.$$

We further assume

$$E\{w_n u_k^T\} = 0.$$

Let us denote by $Y_n^- = \{y_{n-k}, k \geqq 0\}$ the space of past observations at time n and by $x_{n|k}$ the linear mean-square projection of x_n on Y_k^-. We assume that the system is asymptotically stable and that $E\{x_{n|n-1}x_{n|n-1}^T\}$ has a constant value Π and $E\{(x_n - x_{n|n-1})(x_n - x_{n|n-1})^T\}$ has a constant value Σ. It is well known that Π and Σ satisfy the

* Received by the editors June 5, 1984, and in revised form May 13, 1985.
† Department of Electrical Engineering, Technion, Israel Institute of Technology, Haifa, Israel.
‡ Department of Electronic Systems, Tel Aviv University, Tel Aviv, Israel.

equations

$$\Pi - A\Pi A^T = AKRK^TA^T, \tag{1.2}$$

$$\Sigma - A\Sigma A^T + AKC\Sigma A^T = BQB^T, \tag{1.3}$$

where

$$K = \Sigma C^T R^{-1}, \qquad R = C\Sigma C^T + P.$$

The state predictions are obtained from the stationary Kalman filter

$$x_{n+1|n} = Ax_{n|n-1} + AK\nu_n \tag{1.4}$$

where $\nu_n = y_n - Cx_{n|n-1}$.

Let us denote by T the matrix whose columns are the eigenvectors of Π corresponding to nonzero eigenvalues and define

$$z_n = T^T x_{n|n-1}. \tag{1.5}$$

Then z_n has orthonormal components and dim z_n = rank Π. Furthermore, we have the inverse transformation

$$x_{n|n-1} = Tz_n. \tag{1.6}$$

Substituting (1.5) and (1.6) into (1.4) we get

$$z_{n+1} = T^TATz_n + T^TAK\nu_n \tag{1.7}$$

with $\nu_n = y_n - CTz_n$. Estimation updates can be obtained as

$$x_{n|n} = Tz_n + K\nu_n. \tag{1.8}$$

Since, clearly, there is no vector of dimension lower than that of z_n, from which $x_{n|n-1}$ can be obtained by a linear transformation, it seems justified to call (1.7) a minimal state estimator of the system (1.1).

2. Filter order minimality. Recall that to each Jordan block J_i, $i = 1, \cdots, t$, in the spectral decomposition of A there corresponds one eigenvalue α_i and one-eigenrow v_i^T (we note that α_i is not necessarily different from α_j for $i, j = 1, \cdots, t$). We have the following result.

THEOREM 1. *The filter* (1.4) *is of minimal order if and only if* (A, B) *is reachable,* A *is not singular and for any eigenrow* v_i^T *of* A

$$C(\alpha_i^{-1}I - A)^{-1}BQB^Tv_i \neq 0. \tag{2.1}$$

Proof. It follows from (1.3) that

$$(zI - A)\Sigma(z^{-1}I - A^T) + A\Sigma(z^{-1}I - A^T) + (zI - A)\Sigma A^T + \bar{K}C\Sigma A^T = BQB^T,$$

where z is the Z transform variable and where we have denoted

$$\bar{K} = AK.$$

We multiply both sides of this equation by $(zI - A)^{-1}$, on the left, and by $(z^{-1}I - A^T)^{-1}C^T$, on the right, and obtain

$$\begin{aligned}\Sigma C^T + (zI - A)^{-1}A\Sigma C^T + \Sigma A^T(z^{-1}I - A^T)^{-1}C^T \\ + (zI - A)^{-1}\bar{K}C\Sigma A^T(z^{-1}I - A^T)^{-1}C^T \\ = (zI - A)^{-1}BQB^T(z^{-1}I - A^T)^{-1}C^T.\end{aligned}$$

Since $A\Sigma C^T = \bar{K}R$, we readily find then that

$$\Sigma C^T + (zI - A)^{-1}\bar{K}R[I + \bar{K}^T(z^{-1}I - A^T)^{-1}C^T] = (zI - A)^{-1}BQB^T(z^{-1}I - A^T)^{-1}C^T. \tag{2.2}$$

We consider next the partial fraction expansion of the two sides of the last equation. Since the eigenstructure of A may consist of Jordan blocks of dimensions greater than one, we consider the contribution of each Jordan block to the partial fraction expansion separately. It is well known that the terms in the partial fraction expansion of $(zI - A)^{-1}$ that correspond to the Jordan block of dimension r_i of the eigenvalue α_i are

$$\vartheta_{i,k,j}(z) = u_i^{(k)}v_i^{(j)T}(z - \alpha_i)^{-(r_i - j - k)}, \qquad j, k = 0, 1, \cdots, r_i - 1, \tag{2.3}$$

where $u_i^{(0)}$ and $v_i^{(0)T}$ are the eigenvector and the eigenrow that correspond to this Jordan block and $u_i^{(k)}$ and $v_i^{(k)T}$ are the corresponding kth order pseudo-eigenvector and pseudo-eigenrows, respectively [2].

It follows from the asymptotic stability of the system (1.1) that the terms in the partial fraction expansion of $(zI - A)^{-1}BQB^T(z^{-1}I - A^T)^{-1}C^T$ that correspond to the eigenvalues of A are of the form $\vartheta_{i,k,j}(z)BQ \lim_{z\to\alpha_i} B^T(z^{-1}I - A^T)^{-1}C^T$. Obviously there may be terms in the expansion that correspond to $(z - \alpha_i)^{-r_i}$. These terms result from different Jordan blocks that correspond to α_i and α_j where $\alpha_i = \alpha_j$. They include $\vartheta_{j,0,0}(z)BQ \lim_{z\to\alpha_j} B^T(z^{-1}I - A^T)^{-1}C^T$, for all $\alpha_j = \alpha_i$ and $r_i = r_j$ and also other dyadic terms that result from Jordan blocks of $\alpha_j = \alpha_i$ which are of dimension greater than r_i (i.e., $\vartheta_{j,k,r_j - r_i - k}(z)BQ \lim_{z\to\alpha_i} B^T(z^{-1}I - A^T)^{-1}C^T$, $k \geqq r_j - r_i \geqq 0$). Since the eigenvectors and pseudo-eigenvectors of A are all independent, it is possible, given the matrix coefficient of $(z - \alpha_i)^{-r_i}$, in the partial fraction expansion of $(zI - A)^{-1}BQB^T(z^{-1}I - A^T)^{-1}C^T$, to find the term that corresponds to $(z - \alpha_i)^{-r_i}$ and the eigenrow $v_i^{(0)T}$ of the specific Jordan block of α_i, namely, $\vartheta_{i,0,0}(z)BQ \lim_{z\to\alpha_i} B^T(z^{-1}I - A^T)^{-1}C^T$. The corresponding term in the partial fraction expansion of $(zI - A)^{-1}\bar{K}R[I + \bar{K}^T(z^{-1}I - A^T)^{-1}C^T]$ in (2.2) is clearly $\vartheta_{i,0,0}(z)\bar{K}R \lim_{z\to\alpha_i} F^T(z^{-1})$, where $F(z) = I + C(zI - A)^{-1}\bar{K}$ is the return difference matrix of the one step ahead stationary Kalman filter. It follows therefore from (2.2) and (2.3) that

$$v_i^{(0)T}\bar{K}R \lim_{z\to\alpha_i} F^T(z^{-1}) = v_i^{(0)T}BQ \lim_{z\to\alpha_i} B^T(z^{-1}I - A^T)^{-1}C^T. \tag{2.4}$$

By the Popov-Belevitch-Hautus test for reachability [3, p. 135], $(A, \bar{K}R^{1/2})$ is reachable if and only if none of the eigenrows of A lie in the left annihilating subspace of $\bar{K}$. Since the zeros of $F^T(z^{-1})$ are the reciprocals of the poles of the Kalman filter, they lie outside the unit circle and $\lim_{z\to\alpha_i} F^T(z^{-1})$ is therefore nonsingular. The eigenrow $v_i^{(0)T}$ will lie then in the annihilating subspace of $\bar{K}$ if and only if

$$v_i^{(0)T}BQ \lim_{z\to\alpha_i} B^T(z^{-1}I - A^T)^{-1}C^T = 0. \tag{2.5}$$

It follows from (2.5) that the pair $(A, \bar{K}R^{1/2})$ is not reachable if and only if one of the following three cases holds.

(i) The matrix A is singular: The limit for $\alpha_i = 0$ will then yield the required zero in (2.5), independently of the specific structure of B and the eigenrows of A.

(ii) The pair (A, B) is nonreachable: In this case $v_i^{(0)T}B = 0$, which readily satisfies (2.5), independently of the specific structure of C.

(iii) The pair (A, B) is reachable and $\alpha_i \neq 0$ for each i, but (2.5) is satisfied for some i.

Since reachability of $(A, \bar{K}R^{1/2})$ implies that Π has a full rank [4, p. 64] and since $\dim z_n = \operatorname{rank}\Pi$, it follows that the state estimator (1.4) is minimal if and only if A is

nonsingular, (A, B) is reachable and none of the eigenrows of A satisfy (2.5). The proof is thus completed.

Let us now examine the conditions of Theorem 1. The matrix A is known to be nonsingular if the system is obtained by discretizing a continuous time linear system. Suppose that the pair (A, B) is reachable. Then by the Popov-Belevitch-Hautus test $B^T v_i \neq 0$ for all i. Suppose further that $r \geqq q$, i.e., that there are no less outputs than inputs. Then a sufficient condition for minimality of the filter is that the matrix $C(\alpha_i^{-1} I - A)^{-1} B$ is nonsingular. It follows that a sufficient condition for filter minimality in this case is that α_i^{-1} is not a zero of the system, i.e., that none of the system's zeros is a reciprocal of any of its poles. Reciprocal pole-zero pairs are known as all-pass elements as they represent a flat (unity) spectral density. In the single-input, single-output case, we immediately obtain that when the pair (A, B) is reachable, a necessary and sufficient condition for filter minimality is that the system contains no all-pass pairs. Condition (2.1) may then b regarded as a generalization of the no-all-pass condition to the multivariable case.

Finally, we note that the concept of filter order reduction, in the sense of this paper, is different from that of degeneracy of the Ricatti equation, which has received considerable attention in the literature (e.g. [5], [6]). The minimal filter order is equal to the rank of the matrix Π and not to that of the matrix Σ. Consequently, the filter order may not be minimal even when Σ is of full rank, as condition (2.1) need not be satisfied even if the system is reachable and observable. In the single-input, single-output case, when the system is reachable and observable and contains all-pass elements, Σ will be of full rank but Π will be rank deficient, yielding a state estimator of order lower than that of the given system.

3. Conclusion. This paper has investigated the relationships between a discrete time, stochastic linear system in state-space form and its minimal state estimator. A necessary and suffidient condition for the minimality of the Kalman filter as a state estimator has been obtained in terms of the internal system structure.

REFERENCES

[1] A. E. BRYSON AND D. E. JOHANSEN, *Linear filtering for time-varying systems using measurements containing colored noise*, IEEE Trans. Automat. Control, AC-10 (1965), pp. 4-10.

[2] U. SHAKED AND M. DIXON, *Generalized minimal realization of transfer function matrices*, Internat. J. Control, 25 (1977), pp. 785-803.

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

[4] B. D. O. ANDERSON AND J. B. MOORE, *Optimal Filtering*, Prentice-Hall, Englewood Cliffs, NJ, 1979.

[5] M. GEVERS AND T. KAILATH, *Constant, predictable and degenerate directions of the discrete-time Riccati equation*, Automatica, 9 (1973), pp. 699-711.

[6] L. M. Silverman, *Discrete Riccati equations: alternative algorithms, asymptotic properties and system theoretic interpretation*, in Control and Dynamic Systems, C. T. Leondes, ed., Vol. 12, 1976, pp. 313-386.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 4, July 1986

018

ITERATIVE TECHNIQUES FOR THE NASH SOLUTION IN QUADRATIC GAMES WITH UNKNOWN PARAMETERS*

G. P. PAPAVASSILOPOULOS†

Abstract. We study adaptive schemes for repeated quadratic Nash games in a deterministic and a stochastic framework. The convergence of the schemes is demonstrated under certain conditions.

Key words. Nash equilibrium, adaptive Games, stochastic approximation

AMS(MOS) subject classifications. 60G42, 62L20, 90D05, 90D15

1. Introduction. The object of this paper is the study of a static quadratic Nash game where the players do not have knowledge of the parameters involved in the description of the cost of their opponents and of their opponent's information. The game is played repeatedly and at each stage the players know the past actions of their opponents. The only dynamics involved are in the accumulation of the information on their opponent's previous actions; apart from this dynamic aspect, the problem considered is a repeated static game. We examine both the deterministic and stochastic case, consider some adaptive schemes for updating the players decisions, and we show convergence to the optimal decisions (in the mean square sense and with probability one for the stochastic case), under some conditions. The scheme for the stochastic case is actually a stochastic approximation algorithm of the Robbins-Monro type.

The underlying motivation for the present paper is to study situations of conflict where the players do not know some of the parameters involved in the description of the others' cost functionals, or in the state equation. Such situations have been and are being studied for the single player—i.e., control problem—case and come under the name of Adaptive Control; the corresponding problems for situations of conflict, i.e., Adaptive Games, has received very little attention up to now. The problem studied here can be considered as a very simple type of adaptive game where the players adapt their decisions as to converge in the limit to the solution of a static Nash game. It should be noted that the strategies exhibited in this paper do not constitute a Nash equilibrium pair for the construed dynamic—dynamic due to the dynamic information—game; but similarly, the adaptive control strategy in the self-tuning regulator problem [5], converges in the limit to the optimal solution without being necessarily optimal at each stage. Adaptive games are important for several reasons. For example, when two players are involved in a situation of conflict, it is reasonable to assume that each player knows his own objective, but not that of his opponent; in addition, he might not know several of the parameters of the dynamic system which couples him with the other. In decentralized control, we think of decentralization as a scheme according to which each controller knows his own objective and information but not those of the others. If each controller knew the objectives of the others—as is implicitly assumed in many existing decentralized schemes—then the notion of decentralization is weakened. Although considerable progress has been achieved for the centralized controller, single objective adaptive control [4]-[6], the area of adaptive games is in

* Received by the editors June 26, 1984, and in revised form April 1, 1985. This research was supported in part by the U.S. Air Force Office of Scientific Research under grant AFOSR-82-0174 and by the University of Southern California Faculty Research and Innovation Fund.

† University of Southern California, Department of Electrical Engineering-Systems, Los Angeles, California 90089-0781.

its infancy. The only work that the author is familiar with in this area is [7] and [8]. In [7], adaptive schemes based on self-tuning for stochastic Nash and Stackelberg games are considered, where the players have the same information. (In the present paper the information of the players is different.) In [8] two adaptive schemes are studied for repeated Stackelberg games in a determinisitic framework.

The structure of the paper is as follows. In § 2 we consider the determinisitic case and study three simple adaptive schemes. In § 3 we consider an adaptive scheme for the stochastic case. The stochastic scheme is a Robbins-Monro type of stochastic approximation algorithm. Although several results exist for such algorithms, many of which can be used to provide convergence for the scheme considered here, the conditions of convergence that they would obtain for our scheme are more stringent than those that we prove here. In each section we provide several comments relating the results with previous work, expand on their meaning and provide appropriate motivation. Finally, we have a conclusions section.

2. Deterministic case. Let $J_1, J_2: R^{m_1} \times R^{m_2} \to R$ be two functions defined by:

$$J_i(u_1, u_2) = \tfrac{1}{2} u_i' u_i + u_i' R_i u_j + u_i' c_i, \qquad i \neq j, \quad i, j = 1, 2 \tag{1}$$

where $u_i \in R^{m_i}$, R_1, R_2 are real constant matrices and c_1, c_2 are real constant vectors of appropriate dimensions. A pair (u_1^*, u_2^*) is a Nash equilibrium if it satisfies ([1], [2]):

$$J_1(u_1^*, u_2^*) \leqq J_1(u_1, u_2^*) \quad \forall u_1 \in R^{m_1}, \tag{2}$$

$$J_2(u_1^*, u_2^*) \leqq J_2(u_1^*, u_2) \quad \forall u_2 \in R^{m_2}, \tag{3}$$

or equivalently if

$$R \begin{bmatrix} u_1^* \\ u_2^* \end{bmatrix} + c = 0, \quad R = \begin{bmatrix} 1 & R_1 \\ R_2 & 1 \end{bmatrix}, \quad c = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}. \tag{4}$$

J_i and u_i are the cost and the decision of player i.

Let us assume that player i knows R_i and c_i, but not R_j and c_j $(j \neq i)$; then he cannot solve (4) for u_i^*. Consider also that this game is played repeatedly at times $t = 1, 2, 3, \cdots$, that at time t, player i knows $I_t^i = \{u_{1,1}, \cdots, u_{1,t-1}, u_{2,1}, \cdots, u_{2,t-1}\}$ and plays u_{it} which is chosen as a function of I_t^i, i.e.,

$$u_{it} = F_i(I_t^i, t), \qquad i = 1, 2, \quad t = 2, 3, \cdots. \tag{5}$$

The question is: For what F_1, F_2 the recursion (5) will converge to a solution of (4). Let us now examine three possible choices of F_1, F_2.

CASE 1.

$$F_i(I_t^i, t) = -R_i u_{j,t-1} - c_i, \qquad i = 1, 2, \quad i \neq j. \tag{6}$$

The meaning of (6) is that player 1 minimizes $J_1(u_1, u_{2,t-1})$, i.e., he reacts only to the last announced decision of player 2. Recursion (5) assumes the form:

$$\begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix} = \begin{bmatrix} u_{1,t-1} \\ u_{2,t-1} \end{bmatrix} - \left(R \begin{bmatrix} u_{1,t-1} \\ u_{2,t-1} \end{bmatrix} + c \right), \qquad t \geqq 2. \tag{7}$$

Recursion (7) will converge to a solution of (4) for any initial condition $(u_{1,1}, u_{2,1})$ if and only if all the eigenvalues of the matrix R lie within the open disc of radius 1 centered at the point 1 in the complex plane, i.e.,

$$|\lambda(R) - 1| < 1 \tag{8}$$

((8) is equivalent to $|\lambda(R_1, R_2)| < 1$). Condition (8) also guarantees that (4) has a unique solution.

CASE 2.

$$F_i(I_t^i, t) = -R_i[u_{j,t-1} + \theta u_{j,t-2} + \cdots + \theta^{t-2} u_{j,1}] \frac{1-\theta}{1-\theta^{t-1}} - c_i, \tag{9}$$

$$1 > \theta \geqq 0, \quad i = 1, 2, \quad i \neq j.$$

The meaning of (9) is that player 1 minimizes J_1 with respect to u_1, with u_2 fixed to a value that is a weighted average of $u_{2,t-1}, \cdots, u_{2,1}$ where more weight is put on the recent values of u_2. We assume that both players use the same θ. Recursion (9) can be written equivalently:

$$\begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix} = \begin{bmatrix} u_{1,t-1} \\ u_{2,t-1} \end{bmatrix} - \frac{1-\theta}{1-\theta^{t-1}} \left(R \begin{bmatrix} u_{1,t-1} \\ u_{2,t-1} \end{bmatrix} + c \right), \qquad t \geqq 2. \tag{10}$$

Recursion (10) will converge to a solution of (4) for any initial condition $(u_{1,1}, u_{2,1})$ if and only if all the eigenvalues of the matrix R lie within the open disc of radius $(1-\theta)^{-1}$ centered at the point $(1-\theta)^{-1}$ in the complex plane, i.e.,

$$\left| \lambda(R) - \frac{1}{1-\theta} \right| < \frac{1}{1-\theta}. \tag{11}$$

Condition (11) also guarantees that (4) has a unique solution. (Notice that as $t \to +\infty$, $\theta^{t-1} \to 0$ and thus $(1-\theta)R$ in (10) assumes the role of R in (7).) Obviously, for $\theta = 0$, (11) reduces to (8) and (10) to (7).

CASE 3.

$$F_i(I_t^i, t) = -R_i[u_{j,t-1} + u_{j,t-2} + \cdots + u_{j,1}] \frac{1}{t-1} - c_i, \qquad i = 1, 2, \quad i \neq j. \tag{12}$$

The meaning of (12) is that player 1 minimizes J_1 with respect to u_1, with u_2 fixed to the arithmetic mean of $u_{2,t-1}, \cdots, u_{2,1}$. Recursion (12) can be written equivalently:

$$\begin{bmatrix} u_{1,t} \\ u_{2,t} \end{bmatrix} = \begin{bmatrix} u_{1,t-1} \\ u_{2,t-1} \end{bmatrix} - \frac{1}{t-1} \left(R \begin{bmatrix} u_{1,t-1} \\ u_{2,t-1} \end{bmatrix} + c \right), \qquad t \geqq 2. \tag{13}$$

Recursion (13) will converge to a solution of (4), for any initial condition $(u_{1,1}, u_{2,1})$ if and only if all the eigenvalues of R has positive real parts, i.e.,

$$\operatorname{Re} \lambda(R) > 0 \tag{14}$$

(for proof see Appendix A, Lemma A3). Condition (14) also guarantees that (4) has a unique solution. Notice that as $\theta \to 1$, (11) reduces to (14).

Remark 1. Obviously (8)$\Rightarrow$(11)$\Rightarrow$(14). If (8) holds, (7) converges faster than (10) and if (11) holds, (10) converges faster than (13).

Remark 2. In all three cases we assumed that both players use the same scheme. Nonetheless, it might happen that they use different ones. It is easy to verify that if player 1 uses scheme 1 and player 2 uses scheme 2, the region of convergence is larger than if both were using scheme 1 and worse than if both were using scheme 2. Similar results holds for the other combinations.

Remark 3. If we consider (10) with $\theta > 1$, i.e., more weight is assigned to the old measurements, the scheme will not converge. This can be easily verified by considering the scalar version of (10) with $c = 0$:

$$u_t = u_{t-1} \left(1 - r \frac{1-\mu}{\mu} \frac{\mu^{t-1}}{1-\mu^{t-1}} \right), \qquad \mu = \frac{1}{\theta}$$

which for $t \to +\infty$ behaves like

$$y_t = y_{t-1}\left(1 - r\frac{1-\mu}{\mu}\mu^{t-1}\right)$$

(since $0 < \mu < 1$) and is easily seen to fail to converge.

Remark 4. Conditions (8), (11) and (14) can be expressed equivalently in terms of the eigenvalues of R_1R_2.

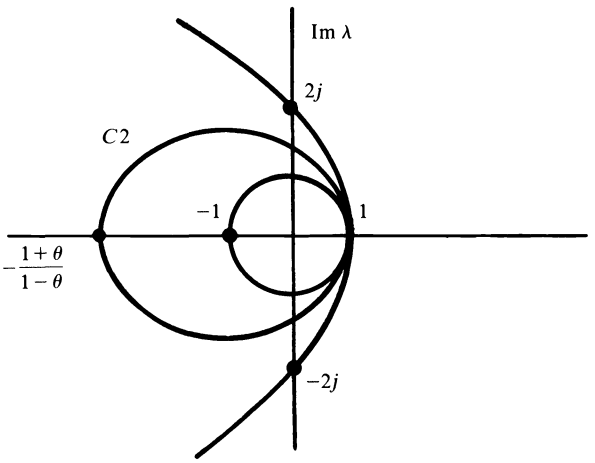


FIG. 1

Condition (8) corresponds to $|\lambda(R_1R_2)| < 1$, i.e., inside the unit disc, (11) corresponds to

$$(1-\theta)|\lambda| \pm 2\theta \cos\frac{\varphi}{2}|\lambda|^{1/2} - (1+\theta) < 0,$$

$$\lambda(R_1R_2) = |\lambda|e^{j\varphi},$$

i.e., inside the curve $C2$ of Fig. 1. Condition (14) corresponds to eigenvalues of R_1R_2 being inside the parabola defined by

$$\operatorname{Re}\lambda + \tfrac{1}{4}(\operatorname{Im}\lambda)^2 < 1, \qquad \lambda = \lambda(R_1R_2).$$

Remark 5. If (8) (or equivalently $|\lambda(R_1R_2)| < 1$) holds, the solution of (4) is called in game theory a stable equilibrium, and the game is called stable [1]. The reason is that if player i deviates from u_i^*, then player $j (j \neq i)$ responds according to scheme (6) and to that player i responds according to scheme (6) and so on and eventually they both converge back to (u_1^*, u_2^*). Obviously the notion of stable equilibrium depends on the reaction scheme that the players employ. If schemes (9) or (12) are used as reaction schemes, we have an enlarged class of stable games.

Remark 6. Since the scheme of Case 3 (12) has the best convergence region out of the three schemes, in the next section we will deal with the stochastic analogue of (12).

Remark 7. All three schemes considered can actually be viewed as schemes for solving $Ru + c = 0$ (see (4)), by using an iteration of the form:

$$u_{n+1} = u_n - D_n[Ru_n + c] \tag{15}$$

where D_n has to have the structure

$$D_n = \begin{bmatrix} D_n^1 & 0 \\ 0 & D_n^2 \end{bmatrix}.$$

(Iterative solutions of linear equations is a vast subject, see for e.g. [16].) Scheme (13) employed: $D_n^i = (1/n)I$. We can create new schemes which converge under weaker conditions than (14) by allowing $D_n^i = (1/n)D^i$ where D^1, D^2 are properly chosen constant matrices. For example, if R_1, R_2 are scalars, (14) is equivalent to $1 > r_1 r_2$; but if we use $D_n^i = (1/n)d_i$ in (15), the convergence condition becomes

$$\operatorname{Re}\lambda\left(\begin{bmatrix} d_1 & 0 \\ 0 & d_2 \end{bmatrix}\begin{bmatrix} 1 & r_1 \\ r_2 & 1 \end{bmatrix}\right) > 0$$

which is equivalent to:

$$d_1 + d_2 > 0, \qquad d_1 d_2 (1 - r_1 r_2) > 0,$$

and can always be satisfied for some d_1, d_2 as long as $1 \neq r_1 r_2$. Notice that $1 \neq r_1 r_2$ is the necessary and sufficient condition for solvability of (4) for any c.

Remark 8. Another way of going about the problem of this section is to consider that at each stage, each player uses a certain scheme to estimate the R and C of his opponent and then calculates his action by solving (4) wherein he employs the estimates of the R and c of his opponent. In such a scheme, each player should know at each stage not only the previous actions of his opponent—as in our scheme—but also the rationale according to which his opponent calculates his actions. This is necessary in order just to estimate his opponent's parameters at each stage. Nonetheless, such an additional knowledge can be permitted and the convergence of the resulting scheme studied. Finally, it should be noted that the problem considered here and the schemes proposed, besides having their own merit, provide a certain motivation for the scheme considered for the stochastic case of the next section.

3. The stochastic case. Let x be a Gaussian random vector in R^n with zero mean and unit covariance matrix. Let

$$y_i = C_i x, \qquad i = 1, 2 \tag{16}$$

represent the measurements of the two players, where C_1, C_2 are fixed real matrices of dimensions $n_1 \times n$, $n_2 \times n$ respectively. Let Γ_i be the set of all measurable γ_i: $R^{n_i} \to R^{m_i}$ functions with $E[\gamma_i(y_i)'\gamma_i(y_i)] < +\infty$. Set $u_i = \gamma_i(y_i)$ and let

$$J_i(\gamma_1, \gamma_2) = E[\tfrac{1}{2}u_i' u_i + u_i' R_i u_j + u_i' S_i x], \qquad i \neq j, \quad i, j = 1, 2 \tag{17}$$

represent the costs of the two players. R_1, R_2, S_1, S_2 are fixed real matrices of appropriate dimensions. A pair $(\gamma_1^*, \gamma_2^*) \in \Gamma_1 \times \Gamma_2$ is called a Nash equilibrium if it satisfies

$$\begin{aligned} &J_1(\gamma_1^*, \gamma_2^*) \leqq J_1(\gamma_1, \gamma_2^*) \quad \forall \gamma_1 \in \Gamma_1, \\ &J_2(\gamma_1^*, \gamma_2^*) \leqq J_2(\gamma_1^*, \gamma_2) \quad \forall \gamma_2 \in \Gamma_2. \end{aligned} \tag{18}$$

For background concerning the formulation of the stochastic Nash game see [18]. (18) is equivalent to (see [2], [3]):

$$\gamma_1^*(y_1) + R_1 E[\gamma_2^*(y_2)|y_1] + S_1 E[x|y_1] = 0, \tag{19a}$$

$$\gamma_2^*(y_2) + R_2 E[\gamma_1^*(y_1)|y_2] + S_2 E[x|y_2] = 0. \tag{19b}$$

It is known (see [3]) that if no eigenvalue of $R_1 R_2$ equals the inverse of any arbitrary but finite product of powers of the squares of the canonical correlation coefficients of

y_1, y_2 (i.e., of $\sigma_1, \sigma_2, \cdots$), then (19) has a unique solution which as to be linear in the information. The set of values where the eigenvalues of R_1R_2 should not lie is a countable isolated set of points in $[1, +\infty)$ and thus it is generically true that (19) admits a unique solution which has to be linear in the information. We can assume without loss of generality (see [3, Lemma 1]) that

$$n_1 \leqq n_2 C_1 C_1' = I_{n_1 \times n_1}, \quad C_2 C_2' = I_{n_2 \times n_2}, \quad C_1 C_2' = \begin{bmatrix} \sigma_1 & & 0 & \vdots & \\ & \sigma_2 & & \vdots & \\ & & \ddots & \vdots & 0 \\ 0 & & & \sigma_{n_1} \vdots & \end{bmatrix}_{n_1 \times n_2} \tag{20}$$

$$1 \geqq \sigma_1 \geqq \sigma_2 \geqq \sigma_{n_1} \geqq 0$$

and then $\gamma_i^*(y_i) = L_i y_i$ where L_1, L_2 are the solutions to the system:

$$\begin{aligned} L_1 + R_1 L_2 C_2 C_1' + S_1 C_1' &= 0, \\ L_2 + R_2 L_1 C_1 C_2' + S_2 C_2' &= 0. \end{aligned} \tag{21}$$

Let us assume that player i knows R_i, S_i, C_i, but not R_j, S_j, C_j, $i \neq j$; then he cannot solve (21) for L_i. Consider also that this game is played repeatedly at times $t = 1, 2, 3, \cdots$, that at time t player i knows

$$I_t^i = \{u_{1,1}, \cdots, u_{1,t-1}, u_{2,1}, \cdots, u_{2,t-1}, y_{i,1}, \cdots, y_{i,t}\} \tag{22}$$

where y_{it} is the measurement of player i at time t. We assume that

$$y_{it} = C_i x_t \tag{23}$$

where the x_t's are independent Gaussian vectors with zero mean and unit covariance. At time t, player 1 employs the following scheme for finding u_{1t}:

$$u_{1t} + R_1 \left(\frac{1}{t-1} \sum_{k=1}^{t-1} u_{2,k} y_{1,k}' \right) y_{1t} + S_1 C_1' y_{1t} = 0. \tag{24}$$

A justification of this scheme is the following: at time t player 1 has to solve (19a) for u_{1t} and thus he has to calculate $E[u_{2,t}|y_{1t}]$, $E[x|y_{1t}]$. If u_{2t} is linear in y_{2t}, then u_{2t}, y_{1t} are jointly Gaussian and thus

$$E[u_{2,t}|y_{1t}] = E[u_{2t} y_{1t}'](E[y_{1t} y_{1t}'])^{-1} y_{1,t}. \tag{25}$$

Player 1 approximates $E[u_{2t} y_{1t}']$ by $1/(t-1)\sum_{k=1}^{t-1}(u_{2,k} y_{1k}^T)$; a motivation for this approximation is the following: If player 1 knew all the parameters of (16), (17), he would then solve equation (19) at state t, employing (23); due to the independence of the x_t's, $1/(t-1)\sum_{k=1}^{t-1}(u_{2k} y_{1k}^T)$ would provide a reasonable approximation of $E[u_{2t}|y_{1t}]$, since u_{2k} would be independent of u_{2l}, y_{1l}, $l \neq k$. By overlooking the lack of independence of u_{2k} on u_{2l}, y_{1l}, $l \neq k$, he still employs the above approximation, hoping that things will work out. The convergence results of Theorems 1′ and 2′ provide a posterior justification for the reasonableness of this approximation.

By our assumption (20) $E[y_{1t} y_{1t}'] = I$ and $E[x_t|y_{1t}] = S_1 C_1' y_{1t}$. (24) yields that u_{1t} is linear in y_{1t}, i.e., $u_{1t} = L_{1t} y_{1t}$ where L_{1t} satisfies

$$L_{1t} + R_1 \left[\frac{1}{t-1} \sum_{k=1}^{t-1} u_{2k} y_{1k}' \right] + S_1 C_1' = 0. \tag{26}$$

A similar equation is satisfied by L_{2t}, if we consider that u_{2t} is calculated by an equation corresponding to (24) and $u_{2t} = L_{2t} y_{2t}$. The equations for L_{1t}, L_{2t} can be written

recursively as:

$$L_{1t} = L_{1,t-1} - \frac{1}{t-1}[L_{1,t-1} + R_1 L_{2,t-1} y_{2,t-1} y'_{1,t-1} + S_1 C'_1], \tag{27a}$$

$$L_{2t} = L_{2,t-1} - \frac{1}{t-1}[L_{2,t-1} + R_2 L_{1,t-1} y_{1,t-1} y'_{2,t-1} + S_2 C'_2]. \tag{27b}$$

Recursion (27) is the recursion that we intend to study and show that under some conditions converges to the solution of (21) in the q.m. sense and w.p.1. The initial condition L_{11}, L_{12} of (27) is taken to be an arbitrary pair of real constant matrices and we are interested in convergence for any initial condition. The recursion (27) defines a Markovian stochastic process (L_{1t}, L_{2t}) and is obviously a stochastic approximation algorithm of the Robbins-Monro type [9] for solving (21). Recursion (27) is a stochastic analogue of the scheme of Case 3 of the deterministic case.

Let us now study the convergence of (27). Let us call l_{it}, m_{it}, c_i, d_i the ith columns of L_{1t}, L_{2t}, $S_1C'_1$, $S_2C'_2$ respectively, i.e.,

$$\begin{aligned} L_{1t} &= [l_{1t}, \cdots, l_{n_1 t}], \qquad L_{2t} = [m_{1t}, \cdots, m_{n_2 t}], \\ S_1 C'_1 &= [c_1, \cdots, c_{n_1}], \qquad S_2 C'_2 = [d_2, \cdots, d_{n_2}]. \end{aligned} \tag{28}$$

Let

$$\bar{l}_{it} = E[l_{it}], \qquad \bar{m}_{it} = E[m_{it}]. \tag{29}$$

Using (20) and the fact that L_{1t} depends on $y_{11}, \cdots, y_{1t-1}, y_{21}, \cdots, y_{2,t-1}$, we obtain from (27):

$$\bar{l}_{it} = \bar{l}_{it-1} - \frac{1}{t-1}[\bar{l}_{i,t-1} + \sigma_i R_1 \bar{m}_{i,t-1} + c_i], \qquad i = 1, \cdots, n_1 \tag{30a}$$

$$\bar{m}_{it} = \bar{m}_{it-1} - \frac{1}{t-1}[\bar{m}_{i,t-1} + \sigma_i R_2 \bar{l}_{i,t-1} + d_i], \qquad i = 1, \cdots, n_1 \tag{30b}$$

and

$$\bar{m}_{it} = \bar{m}_{i,t-1} - \frac{1}{t-1}[\bar{m}_{i,t-1} + d_i], \qquad i = n_1 + 1, \cdots, n_2. \tag{30c}$$

Recursion (30c) converges for any initial condition (see Lemma A3). Recursions (30a) and (30b) can be written as

$$\begin{bmatrix} \bar{l}_{it} \\ \bar{m}_{it} \end{bmatrix} = \begin{bmatrix} \bar{l}_{i,t-1} \\ \bar{m}_{i,t-1} \end{bmatrix} - \frac{1}{t-1}\left(\begin{bmatrix} I & \sigma_i R_1 \\ \sigma_i R_2 & I \end{bmatrix} \begin{bmatrix} \bar{l}_{i,t-1} \\ \bar{m}_{i,t-1} \end{bmatrix} + \begin{bmatrix} c_i \\ d_i \end{bmatrix} \right) \tag{31}$$

and using Lemma A3 yields that (31) converges for any initial condition if and only if

$$\operatorname{Re} \lambda \left(\begin{bmatrix} I & \sigma_i R_1 \\ \sigma_i R_2 & I \end{bmatrix} \right) > 0. \tag{32}$$

It is easy to see that if (32) holds for σ_1 then it holds for any σ_i, $0 \leqq \sigma_i \leqq \sigma_1$. We thus have proven the following theorem.

THEOREM 1'. *The means of L_{1t}, L_{2t} as defined by the recursion* (27) *converge to a solution of* (21) *for any initial condition, if and only if*

$$\operatorname{Re} \lambda \left(\begin{bmatrix} I & \sigma_1 R_1 \\ \sigma_1 R_2 & I \end{bmatrix} \right) > 0. \tag{33}$$

It is easy to see that if (33) holds then (21) has a unique solution. If we want (27) to converge to a solution of (21) not only for any initial condition, but also for any pair of measurements, i.e., any C_1, C_2, we have to consider $\sigma_1 = 1$ in (33) which is exactly the condition for convergence of Case 3 of the deterministic case.

Next we will show that L_{1t}, L_{2t} converge to a solution of (21) in the mean square sense, under condition (33). For simplicity and w.l.o.g. we will assume $S_1 C_1' = 0$, $S_2 C_2' = 0$. We can write (27) component-wise in tems of l_{it}, m_{it} and then form the products $l_{it} l_{jt}'$, $i, j = 1, \cdots, n_1$, $m_{it} m_{jt}'$, $i, j = 1, \cdots, n_2$ and $l_{it} m_{mt}'$, $i = 1, \cdots, n_1$, $j = 1, \cdots, n_2$. These products satisfy recursions that can be easily calculated, and taking expectations of which result in a recursion which gives the evolution of $E(l_{it} l_{jt}')$, $E(m_{it} m_{jt}')$, $E(l_{it} m_{jt}')$ in terms of $E(l_{i,t-1} l_{j,t-1}')$, $E(m_{i,t-1}, m_{j,t-1}')$ $E(l_{i,t-1} m_{j,t-1}')$. Before writing down this recursion we introduce some notation:

$$\Lambda_{ij}^t = E[l_{it} l_{jt}'], \qquad i, j = 1, \cdots, n_1, \tag{34a}$$

$$M_{ij}^t = E[m_{it} m_{jt}'], \qquad i, j = 1, \cdots, n_2, \tag{34b}$$

$$K_{ij}^t = E[l_{it} m_{jt}'], \quad i = 1, \cdots, n_1, \quad j = 1, \cdots, n_2, \tag{34c}$$

$$N_t = \left[\begin{array}{ccc|ccc} \Lambda_{11}^t & \cdots & \Lambda_{1n_1}^t & K_{1,1}^t & \cdots & K_{1,n_2}^t \\ \vdots & & \vdots & \vdots & & \vdots \\ \Lambda_{n_1 1}^t & \cdots & \Lambda_{n_1 n_1}^t & K_{n_1,1} & \cdots & K_{n_1,n_2}^t \\ \hline (K_{11}^t)' & \cdots & (K_{n_1,1}^t)' & M_{11}^t & \cdots & M_{1,n_2}^t \\ \vdots & & \vdots & \vdots & & \vdots \\ (K_{1,n_2}^t)' & \cdots & (K_{n_1 n_2}^t)' & M_{n_2,1} & \cdots & M_{n_2,n_2}^t \end{array}\right], \tag{35}$$

$$Q = \left[\begin{array}{ccc|ccccc} I & & 0 & \sigma_1 R_1 & & 0 & 0 & 0 \\ & \ddots & & & \sigma_2 R_1 & & \vdots & \\ & & & & & \ddots & & \\ 0 & & I & 0 & & \sigma_{n_1} R_1 & 0 \cdots & 0 \\ \hline \sigma_1 R_2 & & 0 & I & & & & \\ & \ddots & & & & & & \\ & 0 & \sigma_{n_1} R_2 & & & & & 0 \\ 0 & \cdots & 0 & & 0 & & \ddots & \\ \vdots & & \vdots & & & & & \\ 0 & \cdots & 0 & & & & & I \end{array}\right]. \tag{36}$$

Then N_t satisfies

$$N_t = N_{t-1} - \frac{1}{t-1}[N_{t-1} Q' + Q N_{t-1}] + \frac{1}{(t-1)^2} \mathscr{L}(N_{t-1}). \tag{37}$$

where $\mathscr{L}(\cdot)$ denotes a linear time invariant function of its argument. (For details of this derivation, see Appendix B.)

Using Lemma A4 we conclude that N_t goes to zero for any initial condition if and only if the matrix Q has eigenvalues with positive real parts which is easily seen to be equivalent to (33). We thus have proven.

THEOREM 2'. *L_{1t}, L_{2t} as defined by recursion* (27) *converge to a solution of* (21) *for any initial condition, in the mean square sense, if and only if* (33) *holds.*

Next, we will show that (L_{1t}, L_{2t}) converges under (33) for any initial condition to the solution of (21) with probability 1 (i.e., a.s. convergence). We again assume for simplicity and w.l.o.g. that $S_1C_1'=0$, $S_2C_2'=0$. We will use the theorem in paragraph 3 of [11] (or [13, Lemma 3.5]) which we restate here and which is an easy consequence of the martingale convergence theorem of Doob.

LEMMA 1. *Let $\{V_t\}$ be a sequence of random variables such that $E(V_1)$ exists. Let A be a real number and suppose $V_t \geqq A$. Furthermore, assume that $\sum_{t=1}^{\infty} E(E[V_{t+1}-V_t|V_1, \cdots, V_t]^+)$ converges. Then the sequence $\{V_t\}$ converges with probability* 1.

(Recall that if x is a random variable: $x^+=\frac{1}{2}(|x|+x)$.) Let $x_t=(l_{1t}', \cdots, l_{n_1,t}', m_{1,t}', \cdots, m_{n_2,t}')'$. We will prove that x_t converges to 0 w.p.l. or equivalently that $V_t=\|x_t\|^2$ does. Let $A=0$. From (27) we can easily obtain (see Appendix C)

$$|E[V_{t+1}-V_t|V_1, \cdots, V_t]| \leqq \frac{\alpha}{t} V_t$$

for some positive number α and thus

$$E[V_{t+1}-V_t|V_1, \cdots, V_t]^+ \leqq \frac{\alpha}{t} V_t.$$

In order to fulfill the assumption of Lemma 1, it suffices to show that

$$\sum_{t=1}^{\infty} \frac{\alpha}{t} E(V_t) < +\infty. \tag{38}$$

$E[V_t]=\operatorname{tr} N_t$ holds, and thus it suffices to show that

$$\sum_{t=1}^{\infty} \frac{\operatorname{tr} N_t}{t} < +\infty. \tag{39}$$

From (37) we obtain

$$N_{t+1}=N_1-Q\left[\sum_{t=1}^{t} \frac{N_k}{k}\right]-\left[\sum_{k=1}^{t} \frac{N_k}{k}\right] Q'+\mathscr{L}\left(\sum_{k=1}^{t} \frac{N_k}{k^2}\right). \tag{40}$$

If we assume that Q has eigenvalues with real parts (40) can be solved for $\sum_{k=1}^{t}(N_k/k)$ to yield

$$\sum_{k=1}^{t} \frac{N_k}{k}=\mathscr{L}'\left(N_{t+1}, N_1, \sum_{k=1}^{t} \frac{N_k}{k^2}\right).$$

Since N_k converges, it is bounded and so is $\sum_{k=1}^{t}(N_k/k^2)$. Thus $\sum_{k=1}^{t}(N_k/k)$ is uniformly bounded and thus (39) and (38) are bounded. We thus conclude that $\|x_t\|^2=V_t$ converges with probability 1. $\|x_t\|^2$ converges to 0 in the mean square sense by Theorem 2′ and thus in probability and thus it has a subsequence converging to zero with probability one [17, Thms. 2, 5, 3, p. 93]. Since we just showed that $\|x_t\|^2$ converges with probability one, this limit has to be zero. Let us now summarize the results of this section in a theorem.

THEOREM. *L_{1t}, L_{2t} as defined by recursion* (27) *converge to a solution of* (21) *for any initial condition, in the mean square sense and with probability one if and only if*

$$\operatorname{Re} \lambda\left(\begin{bmatrix} I & \sigma_1 R_1 \\ \sigma_1 R_2 & I \end{bmatrix}\right) > 0$$

(*under this condition* (21) *admits a unique solution*).

Remark 1. N_t, (37), goes to zero but it does not have to converge monotonically.

Remark 2. One can construct the stochastic analogues of the deterministic schemes of Cases 1 and 2, if a different—appropriate—approximation is used for $E[u_{2t}|y_{1t}]$ in (25). A little reflection, though, will persuade the reader that these schemes will converge under conditions more stringent than (33).

Remark 3. For a repeated Stackelberg game one can consider schemes similar to those considered here, if one assumes that the Leader does not know the parameters involved in the Follower's cost. An idea of this sort was recently studied in a deterministic framework in [8].

Remark 4. It should be clear from (30) and (37) that the rate of convergence of the means and the covariances of l_{it}, m_{it} depend on the eigenvalues of the matrices in (32) for $\sigma_i = 1, \sigma_1, \cdots, \sigma_{n_1}$, or equivalently of Q. Actually, a recursion of the form (A1) with $\bar{\lambda} = \operatorname{Re}(\lambda) > 0$ goes to zero like $(n^{\bar{\lambda}})^{-1}$, (see [12]). Thus if λ_m denotes the real part of the eigenvalues of Q, $m = 1, \cdots, n_1 + n_2$ and $\bar{\lambda} = \min \operatorname{Re}(\lambda_m)$ the mean converges no slower than $(t^{\bar{\lambda}})^{-1}$, the covariances no slower than $(t^{2\bar{\lambda}})^{-1}$, the third moments no slower than $(t^{3\bar{\lambda}})^{-1}$ and so on. Thus if one were to consider whether $t^{\theta}[L_{1t}, L_{2t}]$ converges weakly to a Gaussian random variable as $t \to \infty$, θ should be chosen equal to $\bar{\lambda}$ so that the second moments converge to a nonzero constant, but then automatically all the moments will also do so. Thus in general one cannot have asymptotic normality of $n^{\theta}[L_{1t}, L_{2t}]$ for some $\theta > 0$. As a matter of fact, Theorem (1) of [12] cannot be applied since its assumption (A4) fails for the stochastic approximation algorithm (27), considered here, as should be expected from the above remarks. Finally, it should be pointed out that the fact that the rate of convergence of the algorithm is given by $t^{-\bar{\lambda}}$ and $t^{-2\bar{\lambda}}$ for the first and second moments, is a useful fact when implementing it, in deciding when to stop, what is the probability of error when stopping in a finite number of iterations, etc.

Remark 5. Stochastic approximation has been an object of intensive study (see [9]-[15]). Several of the results available can be used to prove convergence of the iteration (27) but they demand conditions stronger than (33), or they are not applicable to it. For example, in [9] it is required that in the scheme $x_{n+1} = x_n - (1/n)y_n$, y_n is uniformly bounded. Assumptions III and IV of [10] do not hold for (27). In proving asymptotic normality [12], he uses Assumption (A4) which does not hold for (27). Assumptions A5, A5′ of [11] do not hold for our scheme. Lemma 3.1 and Theorem 4.3 of [13] can be applied to (27) but result in more stringent conditions than (33). The convergence analysis of [15] demands boundness of the second term in (27) which is not applicable to our case. Assumption iii in [14, Problem 1, p. 92] does not hold for (27).

4. Conclusions. There are several directions in which this research can be continued. One of them is the corresponding problem for the Stackelberg game (see Remark 3 in § 3). The dynamic case where the players are also coupled through the evolution of a discrete time equation is obviously important and useful. We hope that the analysis presented here will be helpful in such further research.

Appendix A.

LEMMA A1. *Consider the scalar recursion*

$$x_{n+1} = \left(1 - \frac{\lambda}{n}\right) x_n, \qquad n = 1, 2, 3, \cdots \tag{A1}$$

where λ *and* x_1 *are complex numbers. Then* $x_n \to 0$ *for any* x_1 *if and only if* $\operatorname{Re}(\lambda) > 0$.

(*If we set* $t_n = 1 + \cdots + 1/n$, *we see that* (A1) *is a discrete approximation of* $\dot{x} = -\lambda x$ *and thus* $\operatorname{Re}(\lambda) > 0$ *is expected in order to have asymptotic stability of* (A1).)

LEMMA A2. *Consider the scalar recursion*

$$x_{n+1} = \left(1 - \frac{\lambda}{n} + O\left(\frac{1}{n^2}\right)\right) x_n, \qquad n = 1, 2, 3, \cdots \tag{A2}$$

where x *and* x_1 *are complex numbers. Then* $x_n \to 0$ *for any* x_1 *if and only if* $\operatorname{Re}(\lambda) > 0$.

Proof. It is an immediate consequence of Lemma A1 since λ/n dominates $O(1/n^2)$. □

LEMMA A3. *Consider the recursion*

$$x_{n+1} = \left(I - \frac{1}{n} A + O\left(\frac{1}{n^2}\right)\right) x_n, \qquad n = 1, 2, 3, \cdots \tag{A3}$$

where A *is a real square matrix and* x_1 *is a vector. Then* $x_n \to 0$ *for any* x_1 *if and only if* $\operatorname{Re} \lambda(A) > 0$.

Proof. We bring A to its Jordan form and apply Lemma A2. It is helpful to notice that if P is a real symmetric matrix

$$x'_{n+1} P x_{n+1} = x'_n P x_n - \frac{1}{n} x'_n [PA + A'P] x_n + x'_n O\left(\frac{1}{n^2}\right) x_n$$

and thus if A has $\operatorname{Re} \lambda(A) > 0$, we can find a positive definite P so that $A'P + PA > 0$. Therefore if n is sufficiently large

$$\frac{1}{n} x'_n [PA + A'P] x_n > x'_n O\left(\frac{1}{n^2}\right) x_n$$

and thus $x'_{n+1} P x_n < x'_n P x_n$ and consequently x_n is bounded. This justifies the fact that the $1/n$ term dominates in (A3). □

LEMMA A4. *Consider the recursion*

$$N_{t+1} = N_t - \frac{1}{t}[N_t Q' + Q N_t] + \frac{1}{t_2} \mathscr{L}(N_t), \qquad t = 1, 2, \cdots \tag{A4}$$

where N_t, Q *are square matrices.* $N_t \to 0$ *for any initial condition if and only if* $\operatorname{Re} \lambda(Q) > 0$.

Proof. Let x_t be the vector composed of the columns of N_t. We can write the recursion equivalently as

$$x_{t+1} = x_t - \frac{1}{t} A x_t + \frac{1}{t^2} \mathscr{L}(x_t).$$

It can be checked that $\operatorname{Re} \lambda(A) > 0$ if and only if $\operatorname{Re} \lambda(Q) > 0$ and thus Lemma A3 can be applied. □

It should be pointed out that if x_n evolves as in (A1), and λ is real, x_n behaves like $n^{-\lambda}$ (see 12, (2.3)). If λ is complex, then (A2) implies that $|x_n|^2$ behaves like n^{-2a} and thus $|x_n|$ behaves like n^{-a}, i.e., $n^{-\operatorname{Re}\lambda}$. Consequently x_{n+1} in (A3) behaves like $n^{-\bar{\lambda}}$, where $\bar{\lambda} = \min \operatorname{Re} \lambda\ (A)$ and N_t in (A4) behaves like $t^{-2\bar{\lambda}}$ where $\bar{\lambda} = \min \operatorname{Re} \lambda(Q)$.

Appendix B. Let l_{it}, m_{it}, c_i, d_i be as in (28). For convenience, let

$$y_{i,t-1} = \begin{bmatrix} y_1 \\ \vdots \\ y_{n_1} \end{bmatrix}, \qquad y_{2,t-1} = \begin{bmatrix} z_1 \\ \vdots \\ z_{n_2} \end{bmatrix}. \tag{B1}$$

Equation (27) can be written as

$$l_{it} = l_{i,t-1} - \frac{1}{t-1}\left[l_{i,t-1} + y_i R_1 \sum_{j=1}^{n_2} z_j m_{j,t-1} + c_i \right], \qquad i = 1, \cdots, n_1, \tag{B2}$$

$$m_{it} = m_{i,t-1} - \frac{1}{t-1}\left[m_{i,t-1} + z_i R_2 \sum_{j=1}^{n_1} y_j l_{j,t-1} + d_i \right], \qquad i = 1, \cdots, n_2. \tag{B3}$$

For convenience, let us drop the subscript $t-1$ from $l_{i,t-1}$, $m_{i,t-1}$. From (B2), (B3), we obtain:

$$\begin{aligned} l_{it}l'_{jt} = {} & l_i l'_j - \frac{1}{t-1}\left[2l_i l'_j + y_j \sum_{l=1}^{n_2} z_l l_i m'_l R'_1 + y_i R_1 \sum_{k=1}^{n_2} z_k m_k l'_j + l_i c'_j + c_i l'_j \right] \\ & + \frac{1}{(t-1)^2}\left[l_i l'_j + y_j \sum_{l=1}^{n_2} z_l l_i m'_l R'_1 + y_i R_1 \sum_{k=1}^{n_2} z_k m_k l'_j + y_i y_j R_1 \sum_{k,l=1}^{n_2} z_k z_l m_k m'_l R'_1 \right. \\ & \left. + y_i R_i \sum_{k=1}^{n_2} z_k m_k c'_j + y_j \sum_{l=1}^{n_2} z_l c_i m'_l R'_1 + l_i c'_j + c_i l'_j + c_i c'_j \right], \\ & i, j = 1, \cdots, n_1, \end{aligned} \tag{B4}$$

$$\begin{aligned} m_{it}m'_{jt} = {} & m_i m'_j - \frac{1}{t-1}\left[2m_i m'_j + z_j \sum_{l=1}^{n_1} y_l m_i l'_l R'_2 + z_i R_2 \sum_{k=1}^{n_1} y_k l_k m'_j + m_i d'_j + d_i m'_j \right] \\ & + \frac{1}{(t-1)^2}\left[m_i m'_j + z_j \sum_{l=1}^{n_1} y_l m_i l'_l R'_2 + z_i R_2 \sum_{k=1}^{n_1} y_k l_k m'_j + z_i z_j R_2 \sum_{k,l=1}^{n_1} y_k y_l l_k l'_l R'_2 \right. \\ & \left. + z_i R_2 \sum_{k=1}^{n_1} y_k l_k d'_j + z_j \sum_{l=1}^{n_1} y_l d_i l'_l R'_2 + m_i d'_j + d_i m'_j + d_i d'_j \right], \\ & i, j = 1, \cdots, n_2, \end{aligned} \tag{B5}$$

$$\begin{aligned} l_{it}m'_{jt} = {} & l_i m'_j - \frac{1}{t-1}\left[2l_i m_j + z_j \sum_{l=1}^{n_1} y_l l_i l'_l R'_2 + y_i R_1 \sum_{k=1}^{n_2} z_k m_k m'_j + l_i d'_j + c_i m'_j \right] \\ & + \frac{1}{(t-1)^2}\left[l_i m'_j + z_j \sum_{l=1}^{n_1} y_l l_i l'_l R'_2 + y_i R_1 \sum_{k=1}^{n_2} z_k m_k m'_j + y_i z_j R_1 \sum_{k=1}^{n_2} \sum_{l=1}^{n_1} z_k y_l m_k l'_l R'_2 \right. \\ & \left. + y_i R_1 \sum_{k=1}^{n_2} z_k m_k d'_j + z_j \sum_{l=1}^{n_2} y_l c_i l'_l R'_2 + l_i d'_j + c_i m'_j + c_i d'_j \right], \\ & i = 1, \cdots, n_1, \quad j = 1, \cdots, n_2. \end{aligned} \tag{B6}$$

Let Λ_{ij}^t, M_{ij}^t, K_{ij}^t be defined as in (34), let c_i, $d_i = 0$ for simplicity and w.l.o.g.. We take expectation in (B4)-(B6) and drop for convenience the superscript $t-1$ from Λ_{ij}^{t-1}, M_{ij}^{t-1}, K_{ij}^{t-1} in the right-hand side. (When taking expectations, we use the fact that l_i^{t-1}, m_i^{t-1} are independent of $y_{1,t-1}$, $y_{2,t-1}$.) We obtain:

$$\begin{aligned} \Lambda_{ij}^t = {} & \Lambda_{ij} - \frac{1}{t-1}\left[2\Lambda_{ij} + \sigma_j K_{ij} R'_1 + \sigma_i R_1 (K_{ij})' \right] \\ & + \frac{1}{(t-1)^2}\left[\Lambda_{ij} + \sigma_j K_{ij} R'_1 + \sigma_i R_1 (K_{ij})' \right. \\ & \left. + \begin{cases} \sigma_i \sigma_j R_1 (M_{ij} + M_{ji}) R'_1, & \text{if } i \neq j \\ R_1 \left(\sum_{\substack{k=1 \\ k \neq i}}^{n_2} M_{kk} + E(y_i^2 z_i^2) M_{ii} \right) R'_1, & \text{if } i \neq j \end{cases} \right], \\ & i, j = 1, \cdots, n_1, \end{aligned} \tag{B7}$$

$$M_{ij}^{t} = M_{ij} - \frac{1}{t-1}[2M_{ij} + \sigma_j K_{ij} R_2' + \sigma_i R_1 (K_{ij})'] + \frac{1}{(t-1)^2}\mathscr{L}_2(\Lambda_{ij}\text{'s})$$

(B8)

$$i, j = 1, \cdots, n_2 \text{ and } \sigma_i = 0 \quad \text{if } i > n_1,$$

$$\sigma_j = 0 \quad \text{if } j > n_1,$$

$$K_{ij}^{t} = K_{ij} - \frac{1}{t-1}[2K_{ij} + \sigma_j \Lambda_{ij} R_2' + \sigma_i R_1 M_{ij}] + \frac{1}{(t-1)^2}\mathscr{L}_3(\Lambda_{ij}, M_{ij}, K_{ij}\text{'s}),$$

(B9)

$$i = 1, \cdots, n_1, \quad \sigma_i = 0 \text{ if } i > n_1,$$

$$j = 1, \cdots, n_2, \quad \sigma_j = 0 \text{ if } j > n_2.$$

Defining N_t and Q as in (35), (36) we see that (B7)-(B9) can be written in compact form as in (37).

Appendix C. Let $x_t = (l_{it}', \cdots, l_{n_1,t}', m_{1,t}', \cdots, m_{n_2,t}')$. Using (27) or the equivalent (B2), (B3) we have

$$x_{t+1} = x_t - \frac{1}{t}[R(y_{1t}, y_{2t})x_t] \tag{C1}$$

where the definition of $R(y_{1t}, y_{2t}) = \bar{R}_t$ is obvious from (B2), (B3). From (C1) we obtain

$$\|x_{t+1}\|^2 = \|x_t\|^2 - \frac{2}{t} x_t' \bar{R}_t x_t + \frac{1}{t^2} x_t' \bar{R}_t' \bar{R}_t x_t. \tag{C2}$$

It holds

$$\begin{aligned} &E[\|x_{t+1}\|^2 - \|x_t\|^2 | \|x_1\|^2, \cdots, \|x_t\|^2] \\ &\quad = E[E[\|x_{t+1}\|^2 - \|x_t\|^2 | \|x_1\|^2, \cdots, \|x_t\|^2] | x_1, \cdots, x_t], \end{aligned} \tag{C3}$$

$$\begin{aligned} E[x_t' \bar{R}_t x_t | \|x_1\|^2, \cdots, \|x_t\|^2] &= E[E[x_t' \bar{R}_t x_t | \|x_1\|^2, \cdots, \|x_t\|^2] | x_1, \cdots, x_t] \\ &= E[E[x_t' \bar{R}_t x_t | x_1, \cdots, x_t] | x_1, \cdots, x_t] | \|x_1\|^2, \cdots, \|x_t\|^2] \\ &= E[x_t' E[\bar{R}_t | x_1, \cdots, x_t] x_t | \|x_1\|^2, \cdots, \|x_t\|^2] \\ &= E[x_t' R_1 x_t | \|x_1\|^2, \cdots, \|x_t\|^2], \end{aligned} \tag{C4a}$$

since R_t depends only on y_{1t}, y_{2t} which are independent of $x_1, \cdots, x_t$ and where R_1 is a constant matrix defined by

$$E[R(y_{1t}, y_{2t})] = R_1. \tag{C4b}$$

Similarly

$$E[x_t' \bar{R}_t \bar{R}_t x_t | \|x_1\|^2, \cdots, \|x_t\|^2] = E[x_t' R_2 x_t | \|x_1\|^2, \cdots, \|x_t\|^2] \tag{C4c}$$

where R_2 is a constant matrix defined by

$$E[R'(y_{1,t}, y_{2,t}) R(y_{1t}, y_{2t})] = R_2. \tag{C5}$$

From (C3)-(C5) we obtain

$$\begin{aligned} &E[\|x_{t+1}\|^2 - \|x_t\|^2 | \|x_1\|^2, \cdots, \|x_t\|^2] \\ &\quad = E\left[x_t' \left(-\frac{2}{t} R_1 + \frac{1}{t^2} R_2\right) x_t | \|x_1\|^2, \cdots, \|x_t\|^2\right]. \end{aligned} \tag{C6}$$

It holds

$$-\frac{\alpha}{t} I \leqq -\frac{2}{t} R_1 + \frac{1}{t^2} R_2 \leqq \frac{\alpha}{t} I \tag{C7}$$

for some positive constant α and thus

$$\begin{aligned} &\left| \left[E \left[x_t' \left(-\frac{2}{t} R_1 + \frac{1}{t^2} R_2 \right) x_t \middle| \|x_1\|^2, \cdots, \|x_t\|^2 \right] \right] \right| \\ &\leqq \frac{\alpha}{t} E[\|x_t\|^2 | \|x_1\|^2, \cdots, \|x_t\|^2] = \frac{\alpha}{t} \|x_t\|^2. \end{aligned} \tag{C8}$$

Let $V_t = \|x_t\|^2$; then from (C6) and (C7) we obtain

$$|E[V_{t+1} - V_t | V_1, \cdots, V_t]| \leqq \frac{\alpha}{t}. \tag{C9}$$

REFERENCES

[1] T. BASAR AND G. J. OLSDER, *Dynamic Noncooperative Game Theory*, Academic Press, New York, 1982.

[2] T. BASAR, *Equilibrium solutions in two-person quadratic decision problems with statis information structures*, IEEE Trans. Automat. Control, AC-20 (1975), pp. 320-328.

[3] G. P. PAPAVASSILOPOULOS, *Solution of some stochastic quadratic Nash and leader-follower games*, this Journal, 19 (1981), pp. 651-666.

[4] I. D. LANDAU, *A survey of modal reference adaptive techniques-theory and applications*, Automatica, 10 (1974), pp. 353-379.

[5] K. J. ASTROM AND B. WITTENMARK, *On self tuning regulators*, Automatica, 9 (1973), pp. 185-199.

[6] P. R. KUMAR, *Optimal adaptive control of linear quadratic gaussian systems*, this Journal, 21 (1983), pp. 163-178.

[7] Y. M. CHAN, *Self-tuning methods for multiple controller systems*, Ph.D. Thesis, Dept. Electrical Engineering, Univ. Illinois at Urbana-Champaign, 1981.

[8] T. L. TING, J. B. CRUZ, JR. AND R. A. MILITO, *Adaptive incentive controls for Stackelberg games with unknown cost functionals*, American Control Conference, June 1984, San Diego, CA.

[9] H. ROBBINS AND S. MONRO, *A stochastic approximation method*, Ann. Math. Statist., 22 (1951), pp. 400-407.

[10] K. L. CHUNG, *On a stochastic approximation method*, Ann. Math. Statist., 25 (1954), pp. 463-483.

[11] J. R. BLUM, *Multidimensional stochastic approximation methods*, Ann. Math. Statist., 25 (1954), pp. 737-744.

[12] J. SARKS, *Asymptotic distribution of stochastic approximation procedures*, Ann. Math. Statist., 29 (1958), pp. 373-405.

[13] L. SCHMETTERER, *Multidimensional stochastic approximation*, in Multivariate Analysis—II, P. Krishnaiah, ed., Academic Press, New York, 1969.

[14] M. T. WASAN, *Stochastic Approximation*, Cambridge Univ. Press, Cambridge, 1969.

[15] H. J. KUSHNER AND D. S. CLARK, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Springer-Verlag, Berlin 1978.

[16] J. M. ORTEGA AND W. C. RHEINBOLDT, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[17] R. B. ASH, *Real Analysis and Probability*, Academic Press, New York, 1972.

[18] J. B. HARSANGI, *Games with incomplete information played by Bayesian players*, I-III, Management Science, 14 (1969-68), pp. 159-182; 320-334; 486-502.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 5, September 1986

THE LINEAR-QUADRATIC OPTIMAL CONTROL PROBLEM WITH DELAYS IN STATE AND CONTROL VARIABLES: A STATE SPACE APPROACH*

M. C. DELFOUR†

Abstract. This paper generalizes the results of Vinter and Kwong to a control operator with a more general delay structure than the one in A. Ichikawa and to delay systems with both finite and infinite memories. The linear-quadratic optimal control problem over both finite and infinite time horizons is treated. Several forms of associated abstract Riccati differential and algebraic equations are presented. The properties of the kernel of the solution are studied and a coupled set of matrix ordinary and partial differential equations is obtained.

Key words. optimal control, delays, state space, state control

AMS(MOS) subject classifications. 34, 29, 93

1. Introduction. The linear quadratic optimal control theory of systems with delays in state and control variables has been studied by several authors from different view points (cf. A. Ichikawa [1], [2], [3], Koivo and Lee [1], R. H. Kwong [1], [2], [3], Kwong and Willsky [1], [2], L. Pandolfi [1], [2], Vinter and Kwong [1]). Recently state-space techniques have been emphasized and two concepts of state have been proposed: the first one by A. Ichikawa [1], [2], [3] and the second one by Vinter and Kwong [1]. Both approaches lead to some Riccati equations. The work of A. Ichikawa [3] aims at a general theory for a family of evolution equations with a control operator containing a *finite number of pure delays*; the work of Vinter and Kwong [1] deals with differential delay equations in $\mathbb{R}^n$ and a less general control operator which does not contain pure delays.

The object of this paper is to generalize the results of Vinter and Kwong [1] to a *control operator with a more general delay structure* than the one in A. Ichikawa [3] and to *delay systems with both finite and infinite memory.* Finite time and infinite time optimal control problems will be covered. We do not attempt to extend our result to abstract evolution equations in a Hilbert space. We emphasize the basic state space theory of a differential system with delays in state and control variables (§ 4). Three sets of differential equations are derived for the decoupling operator: the standard operator Riccati equation, a new weak form of the equation, and a coupled set of matrix ordinary and partial differential equations for the kernel of the decoupling operator. The first one is a generalization of the one of Vinter and Kwong [1] and several numerical methods are available to approximate its solution. The second one is new and can advantageously be used to construct direct Galerkin full discretization (in space and time variables) methods nicely incorporating the natural characteristics of delay systems. A detailed discussion of those methods is unfortunately beyond the scope of this paper. The last set of coupled matrix differential equations is a complement to the first two. In specific applications and for certain classes of delay systems, it can become a very useful tool in the analysis of the structure of the feedback law. A recent and striking example of this philosophy can be found in the recent work of A. Manitius

* Received by the editors April 22, 1981, and in final revised form June 3, 1985. This research was supported in part by the Natural Sciences and Engineering Research Council of Canada under grant A-8730 and by a FCAC grant from the Ministère de l'Education du Québec.

† Centre de Recherche de Mathématiques Appliquées, Université de Montréal, Montréal, Québec, Canada H3C 3J7.

on the design of feedback controllers for a wind tunnel. He used this set of equations to obtain the structure of the feedback law. He then numerically adjusted the parameters to meet other design criteria. So it is not possible to priviledge a single tool in the analysis and solution of linear quadratic problems. The issue is not to play one method against another one. Those tools and others are available to the designer who uses his experience and judgment in specific applications.

The presence of pure delays in the control operator makes it an unbounded operator (in the state space formulation). This creates new difficulties in the modelling and the solution of this linear quadratic problem; it necessitates special techniques which bear a certain similarity with those encountered in the analysis of parabolic systems with boundary control through a Dirichlet condition (cf. Delfour and Sorine [1]). It must be stressed that the techniques of A. Ichikawa [3] only apply to control operators with a special delay structure (a finite number of pure delays and an integral term on the length of the memory); they do not apply to the general delay structure we shall consider in this paper.

In § 2, we give the fundamental lemma which asserts that the conditions of Borisovic and Turbabin [1] are always verified for linear time-invariant delay-differential equations with finite or infinite memory. This is a special case of M. Delfour [3]. We also extend the concepts of hereditary operators and state found in Vinter and Kwong [1].

In § 3, the statement and solution of the linear quadratic optimal control problem on $[0, T]$ are provided. It is shown that the decoupling operator $\Pi(t)$ is a linear continuous transformation of the space M^2 ($M^2 = \mathbb{R}^n \times L^2(-h, 0; \mathbb{R}^n)$, $0 < h \leqq +\infty$, cf. Notation). However this is not sufficient to obtain a Riccati differential equation for $\Pi(t)$ and it is necessary to embed our optimal control problem into a larger family with initial data and right-hand side in a "bigger" space.

In § 4, the method of transposition is used to make sense of the solution of the state equation for more general data. Fundamental isomorphisms are characterized and an *integration by parts formula* is obtained; perturbation theorems are also presented. To our knowledge the results contained in Theorem 4.5 are new.

In § 5, the original optimal control problem (OCP) is embedded in an enlarged class of OCP's. The solution of this new OCP leads to an optimality system, the use of decoupling techniques and the construction of a Riccati differential equation. It is interesting to notice that the decoupling operator is a "smoothing operator." We essentially follow the method of J. L. Lions [1].

In § 6, we construct the matrix function associated with the decoupling operator $P(t)$, study its properties and derive a set of coupled ordinary and partial differential equations for it. We recover as a special case the equations given by Koivo and Lee [1] and R. H. Kwong [3] for a single delay in the state and control variables. Our results also generalize the linear quadratic optimal control theory of delay systems without delays in the control variable (cf. M. C. Delfour [4] and [1], and Delfour, Lee and Manitius [1]).

In § 7, we solve the infinite time optimal control problem, derive a Riccati equation for the decoupling operator P and a set of coupled partial differential matrix equations for the kernel of P. To our knowledge the techniques used in that section are new.

Some of the results presented in this paper have been announced without proofs in M. C. Delfour [2], [5].

Notation. $\mathbb{R}$ will denote the field of real numbers and, for an arbitrary integer $n \geqq 1$, $\mathbb{R}^n$ will be the n-dimensional Euclidean space. The norm of x in $\mathbb{R}^n$ and the inner product of x and y in $\mathbb{R}^n$ will be written $|x|$ and $x \cdot y$, respectively.

Given $-\infty \leqq a \leqq b \leqq +\infty$, $I(a, b) = \mathbb{R} \cap [a, b]$. For a real Banach space X, $L^2(a, b; X)$ will denote the space of all equivalence classes of square integrable Lebesgue measurable functions on $I(a, b)$ into X. The derivative of a function x on $I(a, b)$ into X will be denoted by $\dot{x}$, dx/dt, Dx or $D_t x$ (in the distributional sense). The Sobolev space of all functions y in $L^2(a, b; X)$ with distributional derivatives $D_t^j y$, $j = 1, \cdots, m$, in $L^2(a, b; X)$ will be written $H^m(a, b; X)$.

$C(a, b; X)$ will be the Banach space of all bounded continuous functions x from $I(a, b)$ into X; when $a = -\infty$, $b = +\infty$, or both, $C_0(a, b; X)$ will be the Banach space of bounded continuous functions vanishing at infinity $C_0(a, b; X) = \{x \in C(a, b; X) | \forall \varepsilon > 0, \exists \text{ compact } K \subset I(a, b), \text{ such that } |x(t)| < \varepsilon, \forall t \in K^c\}$, where K^c is the complement of K with respect to $I(a, b)$,

$$K^c = \{t \in I(a, b) | t \notin K\};$$

when a and b are finite we define $C_0(a, b; X)$ as $C(a, b; X)$. $C_c(a, b; X)$ will be the subspace of functions of $C(a, b; X)$ with compact support in $]a, b[$. It is not to be confused with the space of bounded continuous functions with support in $I(a, b)$. In general, the two spaces do not coincide except on $I(-\infty, \infty)$. In addition to the above function spaces, we shall also use the notation

$$\mathscr{F}_{\text{loc}}(a, \infty; X) = \{y: I(a, \infty) \to X | \forall T > a, y|_{I(a,T)} \in \mathscr{F}(a, T; X)\}$$

for any function space $\mathscr{F}$ (for instance, $\mathscr{F}$ can be C, L^2, H^1, etc.).

Given a real measure μ on a σ-algebra of subsets of a set S, $|\mu|$ will denote the total variation of μ (cf. W. Rudin [1, pp. 117–118]). The total variation of an $n \times m$ matrix β of real measures $\{\beta_{ij} | 1 \leqq i \leqq n, 1 \leqq j \leqq m\}$ is defined as

$$|\beta| = \left\{ \sum_{i=1}^{n} \sum_{j=1}^{m} |\beta_{ij}|^2 \right\}^{1/2},$$

where $|\beta_{ij}|$ is the total variation of β_{ij}.

Given two real Banach spaces X and Y, $\mathscr{L}(X, Y)$ will denote the space of all bounded (or continuous) linear maps from X to Y. The topological dual of X will be written X' and the duality pairing $(x^*, x) \to \langle x^*, x \rangle_X : X' \times X \to \mathbb{R}$. The transpose of an operator T in $\mathscr{L}(X, Y)$ will be denoted $T^* \in \mathscr{L}(Y', X')$.

When X is a Hilbert space, the inner product in $L^2(a, b; X)$ will be denoted $(\ ,\ )_2$. The inner product of two elements $\phi = (\phi^0, \phi^1)$ and $\psi = (\psi^0, \psi^1)$ in the space $M^2 = X \times L^2(-h, 0; X)$, $0 < h \leqq +\infty$, will be defined and denoted as follows:

$$((\phi, \psi)) = \phi^0 \cdot \psi^0 + (\phi^1, \psi^2)_2.$$

$\mathscr{D}(]a, b[; \mathbb{R}^n)$ will denote the vector space of all infinitely continuously differentiable functions from $]a, b[$ into $\mathbb{R}^n$.

In the paper we shall introduce several continuous injections (typ. $L: X \to Y$). The same notation will be used to denote the injection of a space of functions from $I(a, b)$ to X into another space of functions from $I(a, b)$ to Y (typ. $L: L^2(a, b; X) \to L^2(a, b; Y)$, $(Lx)(t) = Lx(t)$ for almost all t in $I(a, b)$). Also the integral of a function $t \to f(t): I(a, b) \to X$ will often be written $\int_a^b f\, dt$ instead of $\int_a^b f(t)\, dt$. This notation is more economical.

Given a Hilbert space V, we shall say that an element T in $\mathscr{L}(V', V)$ is symmetrical if

$$\forall v, w \in V', \qquad \langle v, Tw \rangle_V = \langle w, Tv \rangle_V;$$

we shall say that T is positive if

$$\forall v \in V', \qquad \langle v, Tv\rangle_V \geqq 0.$$

Given an $n \times m$ matrix M ($n \geqq 1$, $m \geqq 1$, integers) the transposed of the $m \times n$ matrix M will be denoted M^T.

2. System description and basic theory. Let h, $0 < h \leqq +\infty$, be the length of the memory of our system. When h is finite, we say that it has *finite memory*; when $h = +\infty$, we say that it has *infinite memory*. Let $B: C_0(-h, 0; \mathbb{R}^m) \to \mathbb{R}^n$ and $L: C_0(-h, 0; \mathbb{R}^n) \to \mathbb{R}^n$ be continuous linear maps. Consider the system of equations

$$\begin{aligned} &\frac{dx}{dt}(t) = Lx_t + Bu_t + f^0(t), \qquad t \geqq 0, \\ &x(0) = \phi^0, \quad x_0 = \phi^1, \quad u_0 = w, \end{aligned} \tag{2.1}$$

where $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$, $(\phi^0, \phi^1) \in M^2$, $w \in L^2(-h, 0; \mathbb{R}^m)$, $f^0 \in L^2_{\text{loc}}(0, \infty; \mathbb{R}^n)$, $u \in L^2_{\text{loc}}(0, \infty; \mathbb{R}^m)$ and u_t and x_t are defined as follows:

$$u_t(\theta) = \begin{cases} u(t+\theta), & t+\theta \geqq 0, \\ w(t+\theta), & \text{otherwise}, \end{cases} \qquad x_t(\theta) = \begin{cases} x(t+\theta), & t+\theta \geqq 0, \\ \phi^1(t+\theta), & \text{otherwise}. \end{cases} \tag{2.2}$$

2.1. Existence and uniqueness theory. For a fixed real number $a > 0$ and a continuous function y in $C_0(-h, a; \mathbb{R}^n)$ (resp. v in $C_0(-h, a; \mathbb{R}^m)$), the functions $t \to (\mathscr{L}y)(t) = Ly_t$ (resp. $t \to (\mathscr{B}v)(t) = Bv_t$) belong to $C(0, a; \mathbb{R}^n)$. Though the function y (resp. v) does not have a pointwise meaning for a function y in $L^2(-h, a; \mathbb{R}^n)$ (resp. v in $L^2(-h, a; \mathbb{R}^m)$), we shall show that it globally makes sense as an element of $L^2(0, a; \mathbb{R}^n)$. This amounts to showing that the three conditions of Borisovic and Turbabin [1] are always verified for delay systems with both finite and infinite memories.

LEMMA 2.1. *Fix h, $0 < h \leqq +\infty$, and the continuous linear map $B: C_0(-h, 0; \mathbb{R}^m) \to \mathbb{R}^n$ (resp. $L: C_0(-h, 0; \mathbb{R}^n) \to \mathbb{R}^n$).*

(i) *There exists a unique $n \times m$ (resp. $n \times n$) matrix β (resp. η) of real regular Borel measures such that*

$$\begin{aligned} &Bw = \int_{-h}^{0} d_\theta \beta w(\theta) \quad \forall w \in C_0(-h, 0; \mathbb{R}^m) \\ &\left(\text{resp. } L\phi = \int_{-h}^{0} d_\theta \eta \phi(\theta) \quad \forall \phi \in C_0(-h, 0; \mathbb{R}^n).\right. \end{aligned} \tag{2.3}$$

(ii) *Let $a > 0$ be a fixed real number. For each v in $C_c(-h, a; \mathbb{R}^m)$ (resp. y in $C_c(-h, a; \mathbb{R}^n)$) the function $\mathscr{B}v$ (resp. $\mathscr{L}y$),*

$$t \to (\mathscr{B}v)(t) = Bv_t : [0, a] \to \mathbb{R}^n \qquad (\text{resp. } t \to (\mathscr{L}y)(t) = Ly_t : [0, a] \to \mathbb{R}^n) \tag{2.4}$$

is continuous and the map

$$\begin{aligned} &v \to \mathscr{B}v : C_c(-h, a; \mathbb{R}^m) \to L^2(0, a; \mathbb{R}^n) \\ &(\text{resp. } y \to \mathscr{L}y : C_c(-h, a; \mathbb{R}^n) \to L^2(0, a; \mathbb{R}^n)) \end{aligned} \tag{2.5}$$

is linear and continuous.

(iii) *The map $\mathscr{B}$ (resp. $\mathscr{L}$) extends to a continuous linear map defined on $L^2(-h, a; \mathbb{R}^m)$ (resp. $L^2(-h, a; \mathbb{R}^n)$)*

$$\mathscr{B} : L^2(-h, a; \mathbb{R}^m) \to L^2(0, a; \mathbb{R}^n) \qquad (\text{resp. } \mathscr{L} : L^2(-h, a; \mathbb{R}^n) \to L^2(0, a; \mathbb{R}^n)). \tag{2.6}$$

Proof. Cf. Appendix to § 2. □

Remark 2.1. For convenience we shall often use the notation Bv_t (resp. Ly_t) instead of $(\mathcal{B}v)(t)$ (resp. $(\mathcal{L}y)(t)$) for v in $L^2(-h, a; \mathbb{R}^m)$ (resp. y in $L^2(-h, a; \mathbb{R}^n)$).

Remark 2.2. For h finite, Lemma 2.1 is essentially Theorem 2.1 in Delfour and Manitius [1]. However, to our knowledge, this result is new for $h = +\infty$.

Lemma 2.1 says that we can now make sense of the right-hand side of (2.1) for all:

$$T>0, \quad x \text{ in } L^2(0, T; \mathbb{R}^n), \quad u \text{ in } L^2(0, T; \mathbb{R}^m), \tag{2.7}$$
$$\phi^1 \text{ in } L^2(-h, 0; \mathbb{R}^n), \quad w \text{ in } L^2(-h, 0; \mathbb{R}^m).$$

Since the maps $\mathcal{B}$ and $\mathcal{L}$ are defined between spaces of L^2-functions, it is possible to rewrite (2.1) in a way which separates the solution x and the control u defined on $[0, \infty[$ from the initial functions ϕ^1 and w on $I(-h, 0)$,

$$\dot{x} = \mathcal{L}(e_+^0 x + e_-^0 \phi^1) + \mathcal{B}(e_+^0 u + e_-^0 w) + f^0, \qquad x(0) = \phi^0, \tag{2.8}$$

where for an arbitrary function z defined on $I(a, b)$, $-\infty \leqq a < b \leqq +\infty$, and an arbitrary $s \in I(a, b)$

$$(e_+^s z)(t) = \begin{cases} z(t), & t \in I(s, b), \\ 0, & \text{otherwise}, \end{cases} \quad \left(\text{resp. } (e_-^s z)(t) = \begin{cases} 0, & \text{otherwise} \\ z(t), & t \in I(a, s) \end{cases} \right). \tag{2.9}$$

Notice that the terms in ϕ^1 and w are zero outside the interval $I(0, h)$. This suggests the introduction of the following *structural operators*

$$H: L^2(-h, 0; \mathbb{R}^n) \to L^2(-h, 0; \mathbb{R}^n), \qquad K: L^2(-h, 0; \mathbb{R}^m) \to L^2(-h, 0; \mathbb{R}^n) \tag{2.10}$$

which are defined for each α in $I(-h, 0)$ as

$$(H\phi^1)(\alpha) = \mathcal{L}(e_-^0 \phi^1)(-\alpha), \qquad (Kw)(\alpha) = \mathcal{B}(e_-^0 w)(-\alpha). \tag{2.11}$$

It is readily seen that for ϕ^1 in $C_0(-h, 0; \mathbb{R}^n)$ and w in $C_0(-h, 0; \mathbb{R}^m)$

$$(H\phi^1)(\alpha) = \int_{-h}^{\alpha} d\eta(\theta)\phi^1(\theta - \alpha), \qquad (Kw)(\alpha) = \int_{-h}^{\alpha} d\beta(\theta) w(\theta - \alpha). \tag{2.12}$$

In Delfour and Manitius [1, Thm. 2.1, p. 470], the first identity (2.12) was used to define H on $C(-h, 0; \mathbb{R}^n)$, $h < +\infty$; then it was shown that H has a continuous extension to $L^2(-h, 0; \mathbb{R}^n)$ which coincides with the operator defined by the first identity (2.11).

In view of (2.8) and (2.11), (2.1) is of the form

$$\dot{x} = \mathcal{L}(e_+^0 x) + \mathcal{B}(e_+^0 u) + f^0 + \mathcal{E}(\xi^1), \qquad x(0) = \xi^0, \tag{2.13}$$

where

$$\xi = (\xi^0, \xi^1) = (\phi^0, H\phi^1 + Kw) \in M^2, \tag{2.14}$$

and $\mathcal{E}$ is the continuous linear map from $L^2(-h, 0; \mathbb{R}^n)$ into $L^2_{\text{loc}}(0, \infty; \mathbb{R}^n)$ defined as

$$\mathcal{E}(\xi^1)(t) = (e_+^{-h}\xi^1)(-t), \qquad t \in I(0, \infty). \tag{2.15}$$

But (2.13) still makes sense for arbitrary ξ^1 in $L^2(-h, 0; \mathbb{R}^n)$. We summarize our conclusions in the next theorem.

THEOREM 2.2. (i) *Equation* (2.1) (*or, more accurately,* (2.8)) *has a unique solution* x *in* $H^1_{\text{loc}}(0, \infty; \mathbb{R}^n)$ *and for each* $T > 0$ *there exists a constant* $c > 0$ *such that*

$$\|x\|_{H^1(0,T;\mathbb{R}^n)} \leqq c[\|u\|_{L^2(0,T;\mathbb{R}^m)} + \|f^0\|_{L^2(0,T;\mathbb{R}^n)} + \|(\phi^0, H\phi^1 + Kw)\|_{M^2}]. \tag{2.16}$$

(ii) *For arbitrary* $\xi=(\xi^0,\xi^1)$ *in* M^2 (ξ *not related to* ϕ), (2.13) *has a unique solution* x *in* $H^1_{\text{loc}}(0,\infty;\mathbb{R}^n)$ *and for each* $T>0$, *there exists a constant* $c>0$ *such that*

$$\|x\|_{H^1(0,T;\mathbb{R}^n)}\leqq c[\|u\|_{L^2(0,T;\mathbb{R}^m)}+\|f^0\|_{L^2(0,T;\mathbb{R}^n)}+\|\xi\|_{M^2}]. \tag{2.17}$$

Proof. By standard arguments. See also M. Delfour [3]. □

2.2. Associated semigroups. Let w, u and f^0 be zero in (2.1). The solution x of (2.1) generates a strongly continuous semigroup $\{S(t)\}$ of class C_0 on M^2

$$S(t)(\phi^0,\phi^1)=(x(t),x_t) \tag{2.18}$$

with infinitesimal generator

$$\begin{aligned} &A(\phi^0,\phi^1)=(L\phi^1, D_\theta\phi^1) \quad \forall\phi=(\phi^0,\phi^1)\in\mathcal{D}(A),\\ &\mathcal{D}(A)=\{(\phi^0,\phi^1)\colon \phi^1\in H^1(-h,0;\mathbb{R}^n),\ \phi^0=\phi^1(0)\}. \end{aligned} \tag{2.19}$$

Notice that the above characterization of the operator A implies the following lemma which can also be proved directly.

LEMMA 2.3. $H^1(-\infty,0;\mathbb{R}^n)\subset C_0(-\infty,0;\mathbb{R}^n)$ *and the injection is continuous.*

Proof. Cf. Appendix to § 2. □

Associate with system (2.1) the following "transposed" system

$$\frac{dz}{dt}(t)=L^Tz_t,\quad t\geqq 0,\quad z(0)=\psi^0,\quad z_0=\psi^1, \tag{2.20}$$

where the operator $L^T\colon C_0(-h,0;\mathbb{R}^n)\to\mathbb{R}^n$,

$$L^T\psi=\int_{-h}^{0}d\eta(\theta)^T\psi(\theta),\qquad \psi\in C_0(-h,0;\mathbb{R}^n), \tag{2.21}$$

is in some sense the "transpose" of L. As above the solution of system (2.20) generates a strongly continuous semigroup $\{S^T(t)\}$ of class C_0 on M^2

$$S^T(t)(\psi^0,\psi^1)=(z(t),z_t),\qquad t\geqq 0, \tag{2.22}$$

with infinitesimal generator

$$A^T(\psi^0,\psi^1)=(L^T\psi^1, D_\theta\psi^1),\quad \forall(\psi^0,\psi^1)\in\mathcal{D}(A^T),\quad \mathcal{D}(A^T)=\mathcal{D}(A). \tag{2.23}$$

From the theory of semigroups of class C_0 we know that for each ψ in $\mathcal{D}(A^T)$ the map $t\to S^T(t)\psi$ defined on $[0,\infty[$ belongs to $C^1(0,T;M^2)\cap C(0,T;\mathcal{D}(A^T))$ for all $T>0$, when $\mathcal{D}(A^T)$ is endowed with the graph topology of A^T.

We shall denote by V *the Hilbert space* $\mathcal{D}(A^T)$ *endowed with the norm*

$$\|\psi\|_V=[\|\psi\|^2_{M^2}+\|A^T\psi\|^2_{M^2}]^{1/2}. \tag{2.24}$$

The restriction of $S^T(t)$ to V also generates a strongly continuous semigroup $\{S^T_V(t)\}$ of class C_0 on V with infinitesimal generator

$$\begin{aligned} &A^T_V(\psi^0,\psi^1)=(L^T\psi^1, D_\theta\psi^1)\quad \forall(\psi^0,\psi^1)\in\mathcal{D}(A^T_V),\\ &\mathcal{D}(A^T_V)=\{(\psi^0,\psi^1)\colon \psi^1\in H^2(-h,0;\mathbb{R}^n),\ \psi^0=\psi^1(0),\ L^T\psi^1=D\psi^1(0)\}. \end{aligned} \tag{2.25}$$

The space V is isomorphic to the Sobolev space $H^1(-h,0;\mathbb{R}^n)$ since the map

$$\psi\to J\psi=(\psi(0),\psi)\colon H^1(-h,0;\mathbb{R}^n)\to V \tag{2.26}$$

is clearly a continuous bijection. In fact one could define the analogue $\{\bar S^T_V(t)\}$ of the

semigroup $\{S_V^T(t)\}$ directly on $H^1(-h, 0; \mathbb{R}^n)$ as follows:

$$[\bar{S}_V^T(t)]\psi](\theta) = y(t+\theta), \qquad \theta \in I(-h, 0), \tag{2.27}$$

where y is the solution of

$$\frac{dy}{dt} = L^T y_t, \qquad y_0 = \psi. \tag{2.28}$$

Its infinitesimal generator would be

$$(\bar{A}_V^T\psi)(\theta) = \begin{cases} L^T\psi, & \theta = 0, \\ D\psi(\theta), & \theta \neq 0, \end{cases} \qquad \mathcal{D}(\bar{A}_V^T) = \{\psi \in H^2(-h, 0; \mathbb{R}^n): L^T\psi = D\psi(0)\}. \tag{2.29}$$

2.3. Extension of the state space theory of Vinter and Kwong. We have seen that a knowledge of the pair $(\phi^0, H\phi^1 + Kw)$ at time 0 is sufficient to solve (2.1). So at time $t > 0$ it is sufficient to know

$$\hat{x}(t) = (x(t), Hx_t + Ku_t) \quad \text{in } M^2 \tag{2.30}$$

in order to solve (2.1) for all times greater than or equal to t. For the more general case of (2.13) with an arbitrary ξ^1, the definition of the L^2-component $\hat{x}^1(t)$ or $\hat{x}(t)$ must be modified as follows

$$[\hat{x}^1(t)](\theta) = (H(e_+^0 x)_t + K(e_+^0 u)_t)(\theta) + \mathcal{E}(\xi^1)(t-\theta). \tag{2.31}$$

More generally for arbitrary initial time $s \geqq 0$ and initial function ξ^1, (2.13) becomes

$$\dot{x}(t) = (\mathcal{L}(e_+^s x) + \mathcal{B}(e_+^s u) + f^0)(t) + \mathcal{E}(\xi^1)(t-s), \quad t \geqq s, \quad x(s) = \xi^0. \tag{2.32}$$

Denote by $x(t; s, \xi, u, f^0)$ the solution of (2.32) at time $t \geqq s$ and by $\hat{x}(t; s, \xi, u, f^0) = (\hat{x}^0(t), \hat{x}^1(t))$ in M^2 its state at time $t \geqq s$ defined as follows:

$$\begin{aligned} &\hat{x}^0(t) = x(t; s, \xi, u, f^0), \\ &\hat{x}^1(t)(\theta) = (H(e_+^s x)_t + K(e_+^s u)_t)(\theta) + \mathcal{E}(\xi^1)(t-s-\theta), \qquad \theta \in I(-h, 0). \end{aligned} \tag{2.33}$$

In their paper, Vinter and Kwong [1] have obtained an equation for the evolution of the state $\hat{x}(t) = \hat{x}(t; s, \xi, u, f^0)$ in M^2 in the case where the operator B is of the form

$$Bw = B_0 w(0) + \int_{-h}^{0} B_1(\theta) w(\theta)\, d\theta, \tag{2.34}$$

where B_0 and $B_1(\theta)$ are $n \times m$ matrices and the elements of the matrix $B_1(\theta)$ belong to $L^2(-h, 0; \mathbb{R})$. In this special situation, for any $T > 0$ and u in $L^2(-h, T; \mathbb{R}^m)$ the map $\mathcal{B}$ associated with B reduces to

$$(\mathcal{B}u)(t) = B_0 u(t) + \int_{-h}^{0} B_1(\theta) u(t+\theta)\, d\theta. \tag{2.35}$$

The object of this section is to extend the above mentioned results to all continuous linear $B: C_0(-h, 0; \mathbb{R}^m) \to \mathbb{R}^n$. This will be given in Theorem 2.7. It necessitates the following definition and technical lemmas.

DEFINITION 2.4. To fix h, $0 < h \leqq +\infty$, we fix the arbitrary continuous linear map $B: C_0(-h, 0; \mathbb{R}^m) \to \mathbb{R}^n$ (resp. $L: C_0(-h, 0; \mathbb{R}^n) \to \mathbb{R}^n$) and its representation in terms of the $n \times m$ (resp. $n \times n$) matrix β (resp. η) of real regular Borel measures.

(i) We shall denote by B^T (resp. L^T) the following continuous linear map

$$\phi \to B^T\phi = \int_{-h}^{0} d_\theta \beta^T \phi(\theta) : C_0(-h, 0; \mathbb{R}^n) \to \mathbb{R}^m, \tag{2.36}$$

$$\left(\text{resp. } \phi \to L^T\phi = \int_{-h}^{0} d_\theta \eta^T \phi(\theta) : C_0(-h, 0; \mathbb{R}^n) \to \mathbb{R}^n\right)$$

where β^T (resp. η^T) is the transposed $n \times m$ matrix β (resp. $n \times n$-matrix η) of real regular Borel measures.

(ii) For each ϕ in $C_0(-h, 0; \mathbb{R}^n)$ define the function

$$(K^T\phi)(\alpha) = \int_{-h}^{\alpha} d_\theta \beta^T \phi(\theta - \alpha) \qquad (\text{resp. } (H^T\phi)(\alpha) = \int_{-h}^{\alpha} d_\theta \eta^T \phi(\theta - \alpha)) \tag{2.37}$$

in $L^2(-h, 0; \mathbb{R}^m)$ (resp. $L^2(-h, 0; \mathbb{R}^n)$). As in Lemma 2.1 and in the discussion from (2.10) to (2.12), the linear map K^T (resp. H^T) extends to a continuous linear map $K^T : L^2(-h, 0; \mathbb{R}^n) \to L^2(-h, 0; \mathbb{R}^m)$ (resp. $H^T : L^2(-h, 0; \mathbb{R}^n) \to L^2(-h, 0; \mathbb{R}^n)$). □

LEMMA 2.5. *If H^* and K^* are the respective transposes of the maps $H : L^2(-h, 0; \mathbb{R}^n) \to L^2(-h, 0; \mathbb{R}^n)$ and $K : L^2(-h, 0; \mathbb{R}^m) \to L^2(-h, 0; \mathbb{R}^n)$, then $H^* = H^T$ and $K^* = K^T$.*

Proof. Cf. Delfour and Manitius [1, pp. 473-474]. □

LEMMA 2.6. *Let $a > 0$ be a fixed real number. For each z in $C_c(-h, a; \mathbb{R}^n)$ the function $\mathcal{B}^T z$ (resp. $\mathcal{L}^T z$) is defined as*

$$t \to (\mathcal{B}^T z)(t) = B^T z_t : [0, a] \to \mathbb{R}^m \tag{2.38}$$

$$(\text{resp. } t \to (\mathcal{L}^T z)(t) = L^T z_t : [0, a] \to \mathbb{R}^n).$$

(i) *The function $\mathcal{B}^T z$ (resp. $\mathcal{L}^T z)(t)$ is continuous and the continuous linear map*

$$z \to \mathcal{B}^T z : C_c(-h, a; R^n) \to L^2(0, a; \mathbb{R}^m) \tag{2.39}$$

$$(\text{resp. } z \to \mathcal{L}^T z : C_c(-h, a; \mathbb{R}^n) \to L^2(0, a; \mathbb{R}^n))$$

extends to a continuous linear map

$$\mathcal{B}^T : L^2(-h, a; R^n) \to L^2(-h, a; \mathbb{R}^m) \tag{2.40}$$

$$(\text{resp. } \mathcal{L}^T : L^2(-h, a; R^n) \to L^2(-h, a; \mathbb{R}^n)).$$

(ii) *For all $t > s$, z in $L^2(s, t; R^n)$ and x in $L^2(s, t; \mathbb{R}^n)$ (resp. u in $L^2(s, t; \mathbb{R}^m)$)*

$$\int_s^t z(t-r) \cdot (\mathcal{L} e_+^s x)(r)\, dr = \int_s^t (\mathcal{L}^T e_+^0 z)(t-r) \cdot x(r)\, dr, \tag{2.41}$$

$$(\text{resp. } \int_s^t z(t-r) \cdot (\mathcal{B} e_+^s u)(r)\, dr = \int_s^t (\mathcal{B}^T e_+^0 z)(t-r) \cdot u(r)\, dr).$$

Proof. Cf. Appendix to § 2. □

In the following theorem the map $j : V \to M^2$ is defined as the injection of V into M^2

$$j(\psi^0, \psi^1) = (\psi^0, \psi^1), \tag{2.42}$$

where V is endowed with the topology generated by the norm (2.24). *We shall also identify the elements of the topological dual $(M^2)'$ of M^2 with those of M^2.* This means that we obtain the following chain of continuous dense injections

$$V \xrightarrow{j} M^2 \equiv (M^2)' \xrightarrow{j^*} V'.$$

THEOREM 2.7. (i) *For each pair* $0 \leqq s \leqq t$,

$$j^*\hat{x}(t; s, \xi, u, f^0) = j^*S^T(t-s)^*\xi + \int_s^t S_V^T(t-r)^*[(B^TJ^{-1})^*u(r) + j^*(f^0(r), 0)]\, dr, \tag{2.43}$$

where $J: H^1(-h, 0; \mathbb{R}^n) \to V$ *is defined by* (2.26), *the same notation* B^T *is used for the restriction to the subspace* $H^1(-h, 0; \mathbb{R}^n)$ *of the map* $B^T: C_0(-h, 0; \mathbb{R}^n) \to \mathbb{R}^m$ ($H^1(-h, 0; \mathbb{R}^n)$ *is a subspace of* $C_0(-h, 0; \mathbb{R}^n)$, $0 < h \leqq +\infty$).

(ii) *For each pair* $0 \leqq s < T$, *the function* $\hat{x}(t) = \hat{x}(t; s, \xi, u, f^0)$, $s \leqq t \leqq T$, *is the unique solution in the space*

$$\mathcal{W}(s, T; M^2, V') = \left\{ y \in C(s, T; M^2) \,\middle|\, \frac{dj^*y}{dt} \in L^2(s, T; V') \right\} \tag{2.44}$$

of the equation

$$\frac{d}{dt}(j^*y(t)) = (A^T)^*y(t) + (B^TJ^{-1})^*u(t) + j^*(f^0(t), 0) \text{ in } V', \qquad y(s) = \xi, \tag{2.45}$$

where j^*, $(B^TJ^{-1})^*$ *and* $(A^T)^*$ *are the transposes of the maps* j, B^TJ^{-1} *and* $A^T: V \to M^2$.

(iii) *In weak form the first equation* (2.45) *is equivalent to*

$$\frac{d}{dt}((j\psi, y(t))) = ((A^T\psi, y(t))) + B^TJ^{-1}\psi \cdot u(t) + \psi^0 \cdot f^0(t), \qquad t \geqq s, \tag{2.46}$$

for all $\psi = (\psi^0, \psi^1)$ *in* V (*recall that for each* $\psi = (\psi^0, \psi^1) \in V$, $\psi^1 \in H^1(-h, 0; \mathbb{R}^n)$ *and* $\psi^0 = \psi^1(0)$).

(iv) *When* B *is of the form*

$$Bw = B_0w(0) + \int_{-h}^0 B_1(\theta)w(\theta)\, d\theta, \tag{2.47}$$

then

$$B^TJ^{-1}\phi = B_M^Tj\phi \quad \forall \phi \in V, \tag{2.48}$$

where $B_M^T: M^2 \to U$ *is the continuous linear map*

$$B_M^T(\phi^0, \phi^1) = B_0^T\phi^0 + \int_{-h}^0 B_1(\theta)^T\phi^1(\theta)\, d\theta. \tag{2.49}$$

For such a B, (2.43), (2.45) *and* (2.46) *reduce to*

$$\hat{x}(t; s, \xi, u, f^0) = S^T(t-s)^*\xi + \int_s^t S^T(t-r)^*[(B_M^T)^*u(r) + (f^0(r), 0)]\, dr, \tag{2.50}$$

$$\frac{d}{dt}(j^*y(t)) = (A^T)^*y(t) + j^*[(B_M^T)^*u(t) + (f^0(t), 0)] \text{ in } V', \qquad y(s) = \xi \text{ in } M^2, \tag{2.51}$$

$$\frac{d}{dt}((j\psi, y(t))) = ((A^T\psi, y(t))) + ((j\psi, (B_M^T)^*u(t) + (f^0(t), 0))), \tag{2.52}$$

for all ψ *in* V.

Proof. Cf. Appendix to § 2. □

Remark 2.3. Part (iv) of the theorem is precisely the special case covered by Vinter and Kwong [1]. It corresponds to the subfamily of all continuous linear operators

$B: C_0(-h, 0; \mathbb{R}^n) \to \mathbb{R}^m$ for which the operator $(B^T J^{-1})^*: U \to V'$ is continuous with values in M^2. This condition is equivalent to say that the continuous linear map $\psi \to B^T J^{-1} \psi: V \to U$ is continuous for the M^2-topology. So it can be extended to a continuous linear map $B_M^T: M^2 \to U$ such that $B^T J^{-1} = B_M^T j$.

3. The linear quadratic optimal control problem on [0, T]. Fix $T > 0$. Associate with the solution of (2.1) the quadratic cost function

$$J^T(u, \phi^0, \phi^1, w) = \int_0^T [Qx(t) \cdot x(t) + Nu(t) \cdot u(t)]\, dt, \tag{3.1}$$

where "$\cdot$" denotes the inner product in $\mathbb{R}^n$ or $\mathbb{R}^m$. Q and N are $n \times n$ and $m \times m$ symmetrical matrices. Q is positive semi-definite and N is positive definite:

$$Q^* = Q \geqq 0, \qquad N^* = N > 0. \tag{3.2}$$

The optimal control problem on the finite time horizon $[0, T]$ consists in finding a u^* in $L^2(0, T; \mathbb{R}^m)$ such that

$$J^T(u^*, \phi^0, \phi^1, w) = \operatorname{Inf}\{J^T(u, \phi^0, \phi^1, w): u \in L^2(0, T; \mathbb{R}^m)\}. \tag{3.3}$$

We have seen that system (2.1) can be expressed in state form by introducing the state $\hat{x}(t)$ which is a solution of (2.50) with $s = 0$ and ξ given by identity (2.14). The cost function (3.1) can be redefined in terms of $\hat{x}(t)$,

$$J^T(u, \phi^0, \phi^1, w) = \int_0^T [((\tilde{Q}\hat{x}(t), \hat{x}(t))) + Nu(t) \cdot u(t)]\, dt, \tag{3.4}$$

by introducing the linear transformation $\tilde{Q}$ of M^2

$$\tilde{Q}(\psi^0, \psi^1) = (Q\psi^0, 0). \tag{3.5}$$

The minimizing control u^* is completely characterized by the following optimality system

$$\frac{dj^*\hat{x}}{dt} = (A^T)^*\hat{x} + (B^T J^{-1})^* u^* + j^*(f^0, 0) \text{ in } V', \qquad \hat{x}(0) = (\phi^0, H\phi^1 + Kw), \tag{3.6}$$

$$u^*(t) = -N^{-1} B^T p(t), \tag{3.7}$$

$$\frac{djp}{dt} + A^T p + \tilde{Q}\hat{x} = 0, \qquad p(T) = 0, \tag{3.8}$$

where the solution p of (3.8) belongs to $C(0, T; V)$ since $\tilde{Q}\hat{x}(t) = (Qx(t), 0)$ and $p(T) = 0$. So (3.7) makes sense.

The next step consists in decoupling the Hamiltonian system

$$\frac{dj^*\hat{x}}{dt} = (A^T)\hat{x} - Rp + j^*(f^0, 0), \qquad \hat{x}(0) = (\phi^0, H\phi^1 + Kw), \tag{3.9}$$

$$R = (B^T J^{-1})^* N^{-1} B^T J^{-1},$$

$$\frac{djp}{dt} + A^T p + \tilde{Q}\hat{x} = 0, \qquad p(T) = 0. \tag{3.10}$$

To do that, we fix s, $0 \leqq s < T$, and consider the embedded problem

$$\operatorname{Inf}\{J_s^T(u, \xi): u \in L^2(s, T; \mathbb{R}^m)\}, \tag{3.11}$$

where

$$J_s^T(u,\xi)=\int_s^T [((\tilde{Q}y(t),y(t)))+Nu(t)\cdot u(t)]\,dt, \tag{3.12}$$

and y is the solution of

$$\frac{dj^*y}{dt}=(A^T)^*y+(B^TJ^{-1})^*u+j^*(f^0,0), \qquad y(s)=\xi. \tag{3.13}$$

The minimizing control is again completely characterized by the Hamiltonian system

$$\frac{dj^*y}{dt}=(A^T)^*y-Rq+j^*(f^0,0), \qquad y(s)=\xi, \tag{3.14}$$

$$\frac{djq}{dt}+A^Tq+\tilde{Q}y=0, \qquad q(T)=0. \tag{3.15}$$

Following the decoupling theory of J. L. Lions [1]

$$p(t)=\Pi(t)\hat{x}(t)+r(t) \text{ in } [0,T], \tag{3.16}$$

where $\Pi(t)$ is a continuous linear operator from M^2 into V and $r(t)$ a vector in V. $\Pi(t)$ and $r(t)$ are constructed from (3.14)-(3.15) as follows:

$$\frac{dj^*\beta}{dt}=(A^T)^*\beta-R\gamma, \qquad \beta(s)=\xi, \tag{3.17}$$

$$\frac{dj\gamma}{dt}+A^T\gamma+\tilde{Q}\beta=0, \qquad \gamma(T)=0, \tag{3.18}$$

$$\Pi(s)\xi=\gamma(s), \tag{3.19}$$

$$\frac{dj^*\phi}{dt}=(A^T)^*\phi-R\psi+j^*(f_0,0), \qquad \phi(s)=0, \tag{3.20}$$

$$\frac{dj\psi}{dt}+A^T\psi+\tilde{Q}\phi=0, \qquad \psi(T)=0, \tag{3.21}$$

$$r(s)=\psi(s). \tag{3.22}$$

The second step is the study of the properties of $\Pi(s)$ with respect to s. This is easy if $\Pi(s)$ is considered as an operator from M^2 into M^2. By standard arguments (cf. J. L. Lions [1]) it can be shown that for each ξ in M^2, the map $s\to j\Pi(s)\xi$ is at least weakly continuous, bounded and strongly measurable from $[0,T]$ into M^2. It is more difficult to obtain the same properties for the same map from $[0,T]$ with value in V.

To get around this technical problem, we are led to embed system (3.13) into a larger class of systems indexed by the initial time s in $[0,T[$, initial conditions ξ in V' and arbitrary functions f in $L^2(0,T;V')$ instead of $j^*(f^0,0)$. However, we must make sense of (3.13) and the cost function (3.12) for such data. The advantage of this approach is that with V' as space of initial data, the techniques of J. L. Lions [1] are available to show that the new decoupling operator, $P(s)$, is a continuous linear map from V' to V and that for all ξ in V' the map $s\to P(s)\xi:[0,T]\to V$ is weakly continuous bounded and strongly measurable.

The fundamental theory is given in § 4, used in § 5 to derive the Riccati differential equation for P, and, in § 6 to derive the differential equations for the kernel associated

with the decoupling operator. *In the sequel* $\Pi(t)$ *will denote the decoupling operator considered as a continuous linear operator from* M^2 *into* M^2 *and not from* M^2 *into* V. In particular the decoupling identity (3.16) will be rewritten in the form

$$jp(t) = \Pi(t)\hat{x}(t) + jr(t) \quad \text{in } [0, T]. \tag{3.23}$$

4. Fundamental isomorphisms. In this section we establish two fundamental isomorphisms, obtain an "integration by parts formula" and prove a perturbation theorem which will be used to obtain the Riccati differential equation. Since this section is rather technical, we summarize the notation which will be used. Define the space

$$W = R^n \times H^1(-h, 0; R^n) \tag{4.1}$$

endowed with the product norm

$$(\phi^0, \phi^1)_W^2 = |\phi^0|^2 + \|\phi^1\|_{H^1}^2. \tag{4.2}$$

Identify the elements of the dual $(M^2)'$ of M^2 with those of M^2 and define the chain of continuous dense injections:

$$H^1(-h, 0; \mathbb{R}^n) \underset{J^{-1}}{\overset{J}{\rightleftarrows}} V \xrightarrow{l} W \xrightarrow{i} M^2 \equiv (M^2)' \xrightarrow{i^*} W' \xrightarrow{l^*} V' \xrightarrow{J^*} (H^1)',$$

$$i(\phi^0, \phi^1) = (\phi^0, \phi^1), \quad l(\phi^0, \phi^1) = (\phi^0, \phi^1), \quad J(\phi) = (\phi(0), \phi), \quad J^{-1}(\phi^0, \phi^1) = \phi^1. \tag{4.3}$$

Given two spaces X, Y with canonical dense injection $i_X: X \to Y$, define for $T > 0$

$$\mathcal{W}(0, T; X, Y) = \{v \in C(0, T; X): D_t i_X v \in L^2(0, T; Y)\}. \tag{4.4}$$

4.1. Adjoint isomorphism. It is well known from semigroup theory that for $T > 0$, $\psi \in V$ and $g \in L^2(0, T; V)$ the function

$$v(t) = S_V^T(T-t)\psi + \int_t^T S_V^T(r-t)g(r)\, dr \tag{4.5}$$

is the unique solution in the space $\mathcal{W}(0, T; V, M^2)$ to the differential equation

$$D_t jv + A^T v + jg = 0, \qquad v(T) = \psi. \tag{4.6}$$

In addition for $T > 0$, $\psi \in V$ and $g^0 \in L^2(0, T; R^n)$ the function

$$v(t) = (p(t), p^t), \qquad t > 0, \tag{4.7}$$

constructed from the solution of the differential equation

$$\dot{p}(t) + L^T p^t + g^0(t) = 0, \qquad (p(T), p^T) = \psi \in V, \tag{4.8}$$

is the unique solution in $\mathcal{W}(0, T; V, M^2)$ to the differential equation

$$D_t jv + A^T v + (g^0(t), 0) = 0, \qquad v(T) = \psi. \tag{4.9}$$

We combine the two sets of results in one.

LEMMA 4.1. *Fix* $T > 0$. *For all* $\psi \in V$ *and* $g \in L^2(0, T; W)$, *there exists a unique solution* v *in* $\mathcal{W}(0, T; V, M^2)$ *to the differential equation*

$$D_t jv + A^T v + ig = 0, \qquad v(T) = \psi. \tag{4.10}$$

From the above lemma, the map

$$v \to (-(D_t jv + A^T v), v(T)): \mathcal{W}(0, T; V, M^2) \to L^2(0, T; M^2) \times V$$

is injective, linear and continuous. It is also surjective onto the subspace $L^2(0, T; W) \times V$. This suggests the definition of the space

$$\mathcal{V}(0, T) = \{v \in C(0, T; V) | -(D_t jv + A^T v) \in L^2(0, T; W)\} \tag{4.11}$$

endowed with the norm

$$\|v\|^2_{\mathcal{V}(0,T)} = \|v\|^2_{C(0,T;V)} + \|D_t jv + A^T v\|^2_{L^2(0,T;W)}. \tag{4.12}$$

The inclusion

$$-(D_t jv + A^T v) \in L^2(0, T; W)$$

is to be understood as follows

$$\exists w \in L^2(0, T; W) \quad \text{such that } -(D_t jv + A^T v) = iw.$$

Notation 4.1. It will be convenient to associate with each element v of $\mathcal{V}(0, T)$ the function

$$-[D_t + A^T]v \in L^2(0, T; W) \tag{4.13}$$

characterized by the identity

$$i[D_t + A^T]v = Djv + A^T v. \tag{4.14}$$

□

It can be shown by standard arguments that the space $\mathcal{V}(0, T)$ is a Banach space.

THEOREM 4.2. *The map*

$$v \to (-[D_t + A^T]v, v(T)) : \mathcal{V}(0, T) \to L^2(0, T; W) \times V \tag{4.15}$$

is an (algebraic and topological) isomorphism.

4.2. Transposed adjoint isomorphism. From standard semigroup theory we can always construct the continuous linear map

$$(f, \xi) \to x : L^2(0, T; V') \times V' \to C(0, T; V') \tag{4.16}$$

where

$$x(t) = S_V^T(t)^* \xi + \int_0^t S_V^T(t - r)^* f(r)\, dr, \qquad 0 \leqq t \leqq T. \tag{4.17}$$

We shall see in Lemma 4.3 that we can further characterize the image of the map (4.16) and construct an isomorphism between $L^2(0, T; V') \times V'$ and that image. This result is in fact completely general and independent of the special semigroup $S_V^T(t)^*$.

However it will not be sufficient four our purposes. We shall need the following additional property:

$$\exists z \in L^2(0, T; W') \quad \text{such that } l^* z(t) = x(t) \quad \text{a.e. in } [0, T], \tag{4.18}$$

where W' is the topological dual of the product space $W = \mathbb{R}^n \times H^1(-h, 0; \mathbb{R}^n)$. This property will be obtained with the help of the "method of transposition."

For the reader who is not familiar with the method of transposition, pertinent details will be given in Lemma 4.4. This method is widely used in the theory of partial differential equations to make sense of the solution of equations with "nonsmooth data." Its starting point is the construction of the so-called "adjoint isomorphism" which is an isomorphism between two Banach spaces of "smooth functions" for the associated "adjoint system." The transpose of that isomorphism is itself an isomorphism between the duals of the initial spaces of "smooth functions" endowed with their

respective norm topologies. It will be referred to as the "transposed adjoint isomorphism."

Coming back to our specific problem, the "transposed adjoint isomorphism" and the "integration by parts" formula will be given in Theorem 4.5. The first Lemma gives a general construction which applies to any strongly continuous semigroup of class C_0.

LEMMA 4.3. *Let $\mathcal{D}(A_V^T)$ be the domain of the operator $A_V^T: \mathcal{D}(A_V^T) \to V$ defined in* (2.25) *of* § 2.2. *With the topology defined by the norm*

$$\|v\|_{\mathcal{D}} = \{\|v\|_V^2 + \|A_V^T v\|_V^2\}^{1/2}, \tag{4.19}$$

$\mathcal{D}(A_V^T)$ is a Hilbert space. Denote by $j_V: \mathcal{D}(A_V^T) \to V$ the continuous linear injection of the subspace $\mathcal{D}(A_V^T)$ of V into V.

(i) *For all $T > 0$, ξ in V' and f in $L^2(0, T; V')$ the function*

$$x(t) = S_V^T(t)^*\xi + \int_0^t S_V^T(t-r)^* f(r)\, dr \tag{4.20}$$

is the unique solution in $\mathcal{W}(0, T; V', \mathcal{D}(A_V^T)')$ to the equation

$$\frac{dj_V^* x}{dt} - (A_V^T)^* x = j_V^* f, \qquad x(0) = \xi, \tag{4.21}$$

where $\mathcal{D}(A_V^T)'$ is the topological dual of the Hilbert space $\mathcal{D}(A_V^T)$ endowed with the norm (4.19) *and $(A_V^T)^*$ is the transpose of the continuous linear operator $A_V^T: \mathcal{D}(A_V^T) \to V$.*

(ii) *The subspace*

$$\mathcal{X}(0, T) = \{x \in \mathcal{W}(0, T; V', \mathcal{D}(A_V^T)') \mid D_t j_V^* x - (A_V^T)^* x \in L^2(0, T; V')\} \tag{4.22}$$

is a Banach space when it is endowed with the norm

$$\|x\|_{\mathcal{X}(0,T)}^2 = \|x\|_{C(0,T;V')}^2 + \|D_t j^* x - (A_V^T)^* x\|_{L^2(0,T;V')}^2. \tag{4.23}$$

(iii) *The map*

$$x \to ((D_t j_V^* - (A_V^T)^*)x, x(0)) : \mathcal{X}(0, T) \to L^2(0, T; V') \times V' \tag{4.24}$$

is an (algebraic and topological) isomorphism when the product space $L^2(0, T; V') \times V'$ is endowed with the norm

$$\|(f, \xi)\|_{L^2(0,T;V') \times V'} = \{\|f\|_{L^2(0,T;V')}^2 + \|\xi\|_{V'}^2\}^{1/2}. \tag{4.25}$$

Proof. By standard arguments. □

Notation 4.2. It will be convenient to associate with each element of the space $\mathcal{X}(0, T)$ the function

$$[D_t - (A_V^T)^*]x \in L^2(0, T; V') \tag{4.26}$$

which is characterized by the identity

$$j_V^*[D_t - (A_V^T)^*]x = D_t j_V^* x - (A_V^T)^* x. \tag{4.27}$$

With the above characterization

$$[D_t - (A_V^T)^*]j^* x = D_t j^* x - (A^T)^* x \quad \forall x \in \mathcal{W}(0, T; M^2, V'). \tag{4.28}$$ □

At this juncture it is useful to recall our twofold objective: to obtain property (4.18) and an "integration by parts formula." All this will be easily obtained by transposition of the adjoint isomorphism (4.15) of Theorem 4.2. The method of transposition uses the following lemma.

LEMMA 4.4. *Given an (algebraic and topological) isomorphism $T: X \to Y$ between two real Banach spaces X and Y, its transpose $T^*: Y' \to X'$ is also an isomorphism between the topological dual spaces Y' and X' of Y and X endowed with their respective norm topologies. Therefore for each x' in X' the "variational equation"*

$$\langle T^* y', x\rangle_X = \langle x', x\rangle_X \quad \forall x \in X \tag{4.29}$$

has a unique solution y' in Y' which is equal to $(T^*)^{-1}x'$. Moreover the solution y' is continuous with respect to the datum x'.

Proof. Cf. Dunford and Schwartz [1, Vol. I, p. 479, Lemma 7]. The existence of a unique solution to the variational equation (4.37) is a trivial consequence of the continuous invertibility of T^*. □

THEOREM 4.5. (i) *Given (f, ξ) in $L^2(0, T; V') \times V'$, the variational equation*

$$-\int_0^T \langle z(t), ([D_t + A^T]v)(t)\rangle_W \, dt + \langle z_T, v(T)\rangle_V = \int_0^T \langle f(t), v(t)\rangle_V \, dt + \langle \xi, v(0)\rangle_V \quad \forall v \in \mathcal{V}(0, T; V, W), \tag{4.30}$$

has a unique solution (z, z_T) in $L^2(0, T; W') \times V'$ and there exists a constant $c > 0$ such that

$$\|(z, z_T)\|_{L^2(0,T;W')\times V'} \leqq c \|(f, \xi)\|_{L^2(0,T;V')\times V'} \tag{4.31}$$

for all (f, ξ) in $L^2(0, T; V') \times V'$.

(ii) *The solution of* (4.30) *is given in terms of ξ and f by the formulae*

$$l^* z(t) = x(t) \quad \text{a.e. in } [0, T], \tag{4.32}$$

$$z_T = x(T), \tag{4.33}$$

where x is the function

$$x(t) = S_V^T(t)^* \xi + \int_0^t S_V^T(t - r)^* f(r) \, dr. \tag{4.34}$$

(iii) *The map*

$$z \to ([D_t - (A_V^T)^*]l^* z, (l^* z)(0)) : \mathcal{Z}(0, T) \to L^2(0, T; V') \times V' \tag{4.35}$$

is an isomorphism for the Banach space

$$\mathcal{Z}(0, T) = \{z \in L^2(0, T; W') \mid l^* z \in \mathcal{H}(0, T)\} \tag{4.36}$$

endowed with the norm

$$\|z\|^2_{\mathcal{Z}(0,T)} = \|z\|^2_{L^2(0,T;W')} + \|l^* z\|^2_{\mathcal{H}(0,T)}. \tag{4.37}$$

(iv) *For all v in $\mathcal{V}(0, T)$ and z in $\mathcal{Z}(0, T)$*

$$-\int_0^T \langle z(t), ([D_t + A^T]v)(t)\rangle_W \, dt + \langle (l^* z)(T), v(T)\rangle_V = \int_0^T \langle ([D_t - (A_V^T)^*]l^* z)(t), v(t)\rangle_V \, dt + \langle (l^* z)(0), v(0)\rangle_V. \tag{4.38}$$

Proof. Cf. Appendix to § 4. □

Remark 4.1. Part (iii) of Theorem 4.5 says that the spaces $\mathcal{H}(0, T)$ and $\mathcal{Z}(0, T)$ are isomorphic:

$$l^* \mathcal{Z}(0, T) = \mathcal{H}(0, T). \tag{4.39}$$

So for each element x of $\mathscr{X}(0, T)$

$$\exists z = (z^0, z^1) \in \mathscr{Z}(0, T) \subset L^2(0, T; W') \quad \text{such that } x = l^* z, \tag{4.40}$$

where W', the dual space of $W = \mathbb{R}^n \times L^2(0, T; \mathbb{R}^n)$, is identified with the product space $\mathbb{R}^n \times H^1(0, T; \mathbb{R}^n)'$. *This is a generalization of the familiar and convenient product space structure* $M^2 = \mathbb{R}^n \times L^2(-h, 0; \mathbb{R}^n)$. Therefore any element x in $\mathscr{X}(0, T)$ can be represented in terms of a pair (z^0, z^1) in $\mathbb{R}^n \times H^1(0, T; \mathbb{R}^n)'$ as follows:

$$\langle x(t), Jv\rangle_V = z^0(t) \cdot v(0) + \langle z^1(t), v\rangle_{H^1} \tag{4.41}$$

$$\forall v \in H^1(-h, 0; \mathbb{R}^n) \quad \text{almost everywhere in } [0, T].$$

4.3. Perturbation theorem.

THEOREM 4.6. (i) *Let* K, $t \to K(t) : [0, T] \to \mathscr{L}(V, V)$ *be a strongly measurable and bounded operator-valued map. For each* v *in* $C(0, T; V)$ *denote by* Kv *the function*

$$(Kv)(t) = K(t)v(t), \qquad 0 \leqq t \leqq T.$$

The continuous linear map

$$v \to \left(-\left[\frac{d}{dt} + A^T\right] v - lKv, v(T)\right) : \mathscr{V}(0, T) \to L^2(0, T; W) \times V \tag{4.42}$$

is an (algebraic and topological) isomorphism.

(ii) *Let* F, $t \to F(t) : [0, T) \to \mathscr{L}(V', V')$ *be a strongly measurable and bounded operator-valued map. For each* y *in* $C(0, T; V')$ *denote by* Fy *the function*

$$(Fy)(t) = F(t)y(t), \qquad 0 \leqq t \leqq T. \tag{4.43}$$

The continuous linear map

$$y \to \left(\left[\frac{d}{dt} - (A_V^T)^*\right] l^* y - Fl^* y, (l^* y)(0)\right) : \mathscr{Z}(0, T) \to L^2(0, T; V') \times V' \tag{4.44}$$

is an (algebraic and topological) isomorphism.

Proof. Cf. Appendix to § 4. □

5. The new embedding of the control problem and the operator Riccati differential equation. In § 3 we have considered the following problem:

$$\operatorname{Inf}\{J^T(u, \phi^0, \phi^1, w) : u \in L^2(0, T; \mathbb{R}^m)\}, \tag{5.1}$$

where

$$J^T(u, \phi^0, \phi^1, w) = \int_0^T [Qx(t) \cdot x(t) + Nu(t) \cdot u(t)]\, dt \tag{5.2}$$

and

$$\begin{aligned} &\frac{dx}{dt} = Lx_t + Bu_t + f^0(t) \text{ in } [0, T], \qquad x \in H^1(0, T; \mathbb{R}^n), \\ &(x(0), x_0, u_0) = (\phi^0, \phi^1, w) \in \mathbb{R}^n \times L^2(-h, 0; \mathbb{R}^n) \times L^2(-h, 0; \mathbb{R}^m). \end{aligned} \tag{5.3}$$

By using the extended state space theory of Vinter and Kwong (cf. § 2.3), problem (5.1) to (5.3) was embedded in the following larger family of problems indexed by $s \in [0, T[$ and ξ in M^2:

$$\operatorname{Inf}\{J_s^T(u, \xi) : u \in L^2(s, T; \mathbb{R}^m)\}, \tag{5.4}$$

where

$$J_s^T(u,\xi)=\int_s^T [((\tilde{Q}y(t),y(t)))+Nu(t)\cdot u(t)]\,dt \tag{5.5}$$

and

$$\frac{d}{dt}j^*y=(A^T)^*y+(B^TJ^{-1})^*u+j^*(f^0,0) \text{ in } [s,T], \qquad y(s)=\xi\in M^2, \tag{5.6}$$

($\tilde{Q}$ is defined by (3.5)).

5.1. New embedding of the control problem. In § 4 we have seen that system (5.6) can be further embedded in the larger family of problems indexed by $s\in[0,T[$ with initial conditions ξ in V' and right-hand sides f in $L^2(0,T;V')$; there exists a unique y in $\mathscr{X}(0,T)$ such that

$$[D_t-(A_V^T)^*]l^*y=(B^TJ^{-1})^*u+f \text{ in } [s,T], \qquad (l^*y)(s)=\xi. \tag{5.7}$$

The critical property that $y\in\mathscr{X}(0,T)\subset L^2(0,T;W')$ makes it possible to also associate with ξ in V' and u in $L^2(s,T;\mathbb{R}^m)$ the cost function

$$\hat{J}_s^T(u,\xi)=\int_s^T [\langle y,\hat{Q}y+2q\rangle_W+Nu\cdot u]\,dt, \tag{5.8}$$

for q in $L^2(0,T;W)$ and $\hat{Q}$ a continuous linear map from $W'=\mathbb{R}^n\times(H^1)'$ into $W=\mathbb{R}^n\times H^1$ defined as

$$\hat{Q}(\psi^0,\psi^1)=(Q\psi^0,0), \qquad H^1=H^1(-h,0;\mathbb{R}^n). \tag{5.9}$$

When $q=0$ and ξ and f are of the form $\xi=j^*\bar{\xi}$ and $f=j^*(f^0,0)$, it is not too difficult to see that (5.7)-(5.8) reduce to (5.6)-(5.5). Indeed the solution y to (5.7) is of the form $y=i^*\bar{y}$ for some $\bar{y}$ in $\mathscr{W}(s,T;M^2,V')$ (cf. Theorem 2.7(ii)) and by identity (4.28)

$$[D_t-(A_V^T)^*]l^*y=[D_t-(A_V^T)^*]l^*i^*\bar{y}=[D_t-(A_V^T)^*]j^*\bar{y}=\frac{d}{dt}j^*\bar{y}-A^T\bar{y}.$$

Furthermore,

LEMMA 5.1. $\tilde{Q}=i\hat{Q}i^*$.

Proof. For all pairs (ϕ^0,ϕ^1) and (ψ^0,ψ^1) in M^2

$$(((\phi^0,\phi^1),\tilde{Q}(\psi^0,\psi^1)))=\phi^0\cdot Q\psi^0$$

and

$$\langle i^*(\phi^0,\phi^1),\hat{Q}i^*(\psi^0,\psi^1)\rangle_W=\phi^0\cdot Q\psi^0. \qquad \square$$

So by construction $\tilde{Q}=i\hat{Q}i^*$ and

$$\langle y(t),\hat{Q}y(t)\rangle_W=\langle i^*\bar{y}(t),\hat{Q}i^*\bar{y}(t)\rangle_W=((\bar{y}(t),i\hat{Q}i^*\bar{y}(t)))=((\bar{y}(t),\tilde{Q}\bar{y}(t))).$$

Thus problem (5.6)-(5.5) is indeed embedded in problem (5.7)-(5.8).

The new optimal control problem in $[s,T]$ consists in minimizing $\hat{J}_s^T(u,\xi)$ over all u in $L^2(s,T;\mathbb{R}^m)$

$$\operatorname{Inf}\{\hat{J}_s^T(u,\xi)\,|\,u\in L^2(s,T;\mathbb{R}^m)\}. \tag{5.10}$$

5.2. Solution of the optimal control problem. The techniques of J. L. Lions [1] directly apply to problem (5.7), (5.8) and (5.10) and the minimizing control u^* is

completely characterized by the equation

$$\int_s^T [\langle z, \hat{Q}y + q\rangle_W + Nu^* \cdot v]\, dt = 0 \tag{5.11}$$

for all v in $L^2(s, T; \mathbb{R}^n)$ where y is the solution of

$$[D_t - (A_V^T)^*]l^*z = (B^TJ^{-1})^*v \text{ in } [s, T], \qquad (l^*z)(s) = 0. \tag{5.12}$$

As usual we introduce the adjoint system

$$-[D_t + A^T]p = \hat{Q}y + q \text{ in } [s, T], \qquad p(T) = 0, \tag{5.13}$$

with solution p in $\mathcal{V}(0, T)$ (cf. Theorem 4.2 and Notation 4.1). By using identity (4.30) of Theorem 4.5 on the interval $[s, T]$, we readily obtain

$$\int_s^T (B^TJ^{-1}p + Nu^*) \cdot v\, dt = 0 \quad \forall v \in L^2(s, T; \mathbb{R}^m), \tag{5.14}$$

or since N is invertible

$$u^*(t) = -N^{-1}B^TJ^{-1}p(t) \text{ in } [s, T]. \tag{5.15}$$

The substitution of $u = u^*$ (given by expression (5.15)) into (5.7) leads to the following Hamiltonian system

$$[D_t - (A_V^T)^*](l^*y) = -Rp + f, \quad (l^*y)(s) = \xi, \quad R = (B^TJ^{-1})^*N^{-1}B^TJ^{-1}, \tag{5.16}$$

$$-[D_t + A^T]p = \hat{Q}y + q, \qquad p(T) = 0. \tag{5.17}$$

It is important to emphasize that the above and forthcoming results follow the standard arguments of J. L. Lions [1]. For this reason proofs will be omitted. The next theorem uses invariant embedding with one small difference with respect to J. L. Lions [1]. The space of initial conditions V' is a Hilbert space which has not been identified with its topological dual since we have already identified the elements of the dual $(M^2)'$ of M^2 with the elements of M^2. This difference does not affect the proofs but is very essential in our results.

THEOREM 5.2. (i) *For $s = 0$, there exists a family $\{P(t): 0 \leqq t \leqq T\}$ of continuous linear maps from V' into V and a family $\{r(t): 0 \leqq t \leqq T\}$ of vectors in V such that*

$$p(t) = P(t)(l^*y)(t) + r(t), \qquad 0 \leqq t \leqq T. \tag{5.18}$$

$P(t)$ and $r(t)$ are obtained by the following rules:

a) *Solve*

$$\begin{aligned} &[D_t - (A_V^T)^*](l^*\phi) = -R\psi \text{ in } [s, t], \qquad (l^*\phi)(s) = h, \\ &-[D_t + A^T]\psi = \hat{Q}\phi \text{ in } [s, T], \qquad \psi(T) = 0, \end{aligned} \tag{5.19}$$

$$P(s)h = \psi(s). \tag{5.20}$$

b) *Solve*

$$\begin{aligned} &[D_t - (A_V^T)^*](l^*\beta) = -R\gamma + f \text{ in } [s, T], \qquad (l^*\beta)(s) = 0, \\ &-[D_t + A^T]\gamma = \hat{Q}\beta + q \text{ in } [s, T], \qquad \gamma(T) = 0, \end{aligned} \tag{5.21}$$

$$r(s) = \gamma(s). \tag{5.22}$$

(ii) *The operator $P(s)$ is symmetrical and positive*

$$\begin{aligned} &\forall h, k \text{ in } V', \quad \langle h, P(s)k\rangle_V = \langle k, P(s)h\rangle_V \\ &\forall h \text{ in } V', \quad \langle h, P(s)h\rangle_V \geqq 0. \end{aligned} \tag{5.23}$$

For all h in V', the map

$$t \to P(t)h : [0, T] \to V \tag{5.24}$$

is weakly continuous, bounded and strongly measurable. In particular

$$\exists c > 0 \quad \text{such that } \forall t \in [0, T], \quad \|P(t)\| \leqq c. \tag{5.25}$$

The proof of this theorem uses standard arguments and will be omitted.

5.3. Riccati differential equation for P. The remaining considerations and the final results are based on the perturbation Theorem 4.6 and a lemma. Theorem 4.6 will be used to perturb the adjoint isomorphism (4.15) by the operator $K(t) = P(t)R$; the lemma will say that if $q(t)$ is chosen equal to $-lP(t)f(t)$ in (5.17), then the pair of solutions (y, p) to system (5.16)-(5.17) verifies the identity $p(t) = P(t)y(t)$, $0 \leqq t \leqq T$.

LEMMA 5.3. *Recall the Hamiltonian system* (5.16)-(5.17), *identity* (5.18) *and the characterization of r in Theorem* 5.2(i). *The following conditions are equivalent*:

(i) $\forall s, \quad 0 \leqq s \leqq T, \quad r(s) = 0,$

(ii) $\forall s, \quad 0 \leqq s \leqq T, \quad p(s) = P(s)(l^* y)(s),$

(iii) *for all h in V' and s, $0 \leqq s \leqq T$,*

$$\int_s^T [\langle \phi(t), q(t)\rangle_W + \langle f(t), \psi(t)\rangle_V]\, dt = 0, \tag{5.26}$$

where (ϕ, ψ) is the solution of system (5.19) *on $[s, T]$ with initial condition h at time s. In particular if*

$$q(t) = -lP(t)f(t), \qquad 0 \leqq t \leqq T, \tag{5.27}$$

condition (5.26) *is verified and $r(s) = 0$, $0 \leqq s \leqq T$.*

Proof. Cf. Appendix to § 5. □

THEOREM 5.4. (i) *For each x in $\mathscr{Z}(0, T)$ the function Pl^*x,*

$$(P(l^*x))(t) = P(t)[l^*x](t) = P(t)l^*x(t), \qquad 0 \leqq t \leqq T, \tag{5.28}$$

is the unique solution in $\mathscr{V}(0, T)$ to the equation

$$[D_t + A^T]Pl^*x - lP[D_t - (A_V^T)^*]l^*x - lPR(Pl^*x) + \hat{Q}x = 0, \qquad P(T) = 0. \tag{5.29}$$

The family $P = \{P(t) \in \mathscr{L}(V', V) | 0 \leqq t \leqq T\}$ is the unique family with the above property in the class $(\mathscr{P})$ of all families of weakly continuous symmetrical operators $Z = \{Z(t) \in \mathscr{L}(V', V) | 0 \leqq t \leqq T\}$.

(ii) *The map $Pl^* : \mathscr{Z}(0, T) \to \mathscr{V}(0, T)$ is linear and continuous.*

(iii) *The function r is the unique solution in $\mathscr{V}(0, T)$ to the equation*

$$-[D_t + A^T]r + PRr = Pf + q \text{ in } [0, T], \qquad r(T) = 0. \tag{5.30}$$

Proof. Cf. Appendix to § 5. □

Equation (5.29) is not the usual form of the Riccati equation. We now make this connection.

THEOREM 5.5. *The following statements are equivalent to each other and to Theorem* 5.4(i).

(i) *For each x in $\mathcal{W}(0, T; M^2, V')$, the function Pj^*x is the unique solution in $\mathcal{V}(0, T)$ to the equation*

$$\frac{d}{dt}(jPj^*x)+A^TPj^*x+jP[-D_t(j^*x)+(A^T)^*x-RPj^*x]+\tilde{Q}x=0, \qquad P(T)=0. \tag{5.31}$$

In the class $(\mathcal{P})$ the solution P is unique.

(ii) *For each h in M^2, the function $t\to P(t)j^*h$ is the unique solution in $\mathcal{V}(0, T; V, W)$ to the equation*

$$\frac{d}{dt}(jP(t)j^*h)+A^TP(t)j^*h+jP(t)(A^T)^*h-jP(t)RP(t)j^*h+\tilde{Q}h=0, \tag{5.32}$$

$$P(T)=0.$$

P is the unique solution in the class $(\mathcal{P})$.

Proof. Cf. Appendix to § 5. □

6. Differential equation for the kernel of the equation $P(t)$. In this section we construct the kernel $P(t, \alpha, \theta)$ associated with the operators Π and P and derive a set of coupled partial differential matrix equations for it. We recover, as a special case, the equations derived by H. N. Koivo and E. B. Lee [1] and R. H. Kwong [3] for a single delay.

Recall that

$$V=\{(\phi^0, \phi^1)\in M^2 | \phi^1\in H^1(-h, 0; \mathbb{R}^n) \text{ and } \phi^0=\phi^1(0)\}, \tag{6.1}$$

that J is the isomorphism

$$\phi\to J\phi=(\phi(0), \phi): H^1\to V, \qquad H^1=H^1(-h, 0; \mathbb{R}^n), \tag{6.2}$$

and that j, i and l are the continuous injections

$$(\phi^0, \phi^1)\to j(\phi^0, \phi^1)=(\phi^0, \phi^1): V\to M^2, \tag{6.3}$$

$$(\phi^0, \phi^1)\to l(\phi^0, \phi^1)=(\phi^0, \phi^1): V\to W, \tag{6.4}$$

$$(\phi^0, \phi^1)\to i(\phi^0, \phi^1)=(\phi^0, \phi^1): W\to M^2, \tag{6.5}$$

$$j=il,$$

where V is endowed with the norm (2.24).

6.1. Definition and properties of the matrix kernel $P(t, \alpha, \theta)$. Recall the definition of $P(t)$, the results of Theorem 5.1 and the definition of $\Pi(t)$ at the end of § 3.

PROPOSITION 6.1. *For all t in $[0, T]$*

$$\Pi(t)=jP(t)j^*. \tag{6.6}$$

Proof. From definitions. □

The subspace V of M^2 endowed with the norm (2.24) is essentially isomorphic to H^1. It will be convenient to introduce on H^1, an operator equivalent to $P(t)$ on V.

DEFINITION 6.2. For each t in $[0, T]$

$$P_J(t)=J^{-1}P(t)(J^*)^{-1}. \tag{6.7}$$

□

It is readily seen that for each t, $P_J(t)$ is a continuous linear operator from $(H^1)'$ into H^1 and that the family $\{P_J(t): 0\leqq t\leqq T\}$ has properties analogous to those of the

family $\{P(t): 0 \leqq t \leqq T\}$:

$$\forall h, k \in (H^1)', \qquad \langle h, P_J(t)k\rangle_{H^1} = \langle k, P_J(t)h\rangle_{H^1},$$

$$\forall h \in (H^1)', \qquad \langle h, P_J(t)h\rangle_{H^1} \geqq 0,$$

$$\exists c > 0, \quad \forall t \in [0, T], \quad \|P_J(t)\| \leqq c, \tag{6.8}$$

$$\forall h \in (H^1)', \qquad t \to P_J(t)h : [0, T] \to H^1$$

is weakly continuous, bounded and strongly measurable.

DEFINITION 6.3. For each $(t, \alpha, \theta) \in [0, T] \times I(-h, 0) \times I(-h, 0)$, let $P(t, \alpha, \theta)$ be the $n \times n$ matrix defined as follows:

$$(e, d) \to d \cdot P(t, \alpha, \theta)e = \langle \delta_\theta d, P_J(t)(\delta_\alpha e)\rangle_{H^1} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}, \tag{6.9}$$

where for each ζ in $I(-h, 0)$ δ_ζ is the continuous linear map

$$\delta_\zeta : \mathbb{R}^n \to (H^1)', \quad \langle \delta_\zeta c, \phi\rangle_{H^1} = c \cdot \phi(\zeta) \quad \forall c \in \mathbb{R}^n \quad \forall \phi \in H^1. \qquad \square$$

It is easy to check that $P(t, \alpha, \theta)$ is well defined. The following lemma will be useful in the study of the properties of $P(t, \alpha, \theta)$.

LEMMA 6.4. (i) *For each e in $\mathbb{R}^n$, the function*

$$\alpha \to \delta_\alpha e : [-h, 0] \to (H^1)' \tag{6.10}$$

is uniformly continuous,

$$\forall \theta, \alpha \in I(-h, 0), \qquad \|\delta_\theta e - \delta_\alpha e\|_{(H^1)'} \leqq |\theta - \alpha|^{1/2}|e| \tag{6.11}$$

and there exists a constant $c' > 0$ such that

$$\forall \alpha \in I(-h, 0), \qquad \|\delta_\alpha e\|_{(H^1)'} \leqq c'|e|. \tag{6.12}$$

(ii) *For all ψ in $L^2(-h, 0; \mathbb{R}^n)$*

$$\int_{-h}^{0} \delta_\theta \psi(\theta)\, d\theta = J^* j^*(0, \psi) \quad \text{in } (H^1)'. \tag{6.13}$$

Proof. (i) For each ϕ in H^1

$$\langle \delta_\theta e - \delta_\alpha e, \phi\rangle_{H^1} = e \cdot [\phi(\theta) - \phi(\alpha)] = e \cdot \int_\alpha^\theta D\phi(\xi)\, d\xi,$$

$$|\langle \delta_\theta e - \delta_\alpha e, \phi\rangle| \leqq |e||\theta - \alpha|^{1/2}\|D\phi\|_{L^2} \leqq |e||\theta - \alpha|^{1/2}\|\phi\|_{H^1}.$$

By definition of the norm in $(H^1)'$, we obtain (6.11). To prove (6.12) recall the definition of δ_α and the norm in $(H^1)'$:

$$\|\delta_\alpha e\|_{(H^1)'} = \sup\{|\langle \delta_\alpha e, v\rangle_{H^1}| : \|v\|_{H^1} = 1\} = \sup\{|e \cdot v(\alpha)| : \|v\|_{H^1} = 1\}$$
$$\leqq |e| \sup\{c'\|v\|_{H^1} : \|v\|_{H^1} = 1\} = c'|e|,$$

since the injection of $H^1(-h, 0; \mathbb{R}^n)$ into the space of bounded continuous functions $C(-h, 0; \mathbb{R}^n)$ is continuous.

(ii) For each v in $H^1(-h, 0; \mathbb{R}^n)$ and almost all θ in $I(-h, 0)$

$$\langle \delta_\theta \psi(\theta), v\rangle_{H^1} = \psi(\theta) \cdot v(\theta),$$

$$\int_{-h}^{0} \langle \delta_\theta \psi(\theta), v\rangle_{H^1}\, d\theta = \int_{-h}^{0} \psi(\theta) \cdot v(\theta)\, d\theta = (((0, \psi), jJv)) = \langle J^* j^*(0, \psi), v\rangle_{H^1}.$$

Therefore

$$\left\langle \int_{-h}^{0} \delta_\theta \psi(\theta)\, d\theta, v \right\rangle_{H^1} = \langle J^* j^*(0, \psi), v \rangle_{H^1}. \qquad \square$$

THEOREM 6.5. (i) *For all* (t, α, θ) *in* $[0, T] \times I(-h, 0) \times I(-h, 0)$

$$P(t, \alpha, \theta)^* = P(t, \theta, \alpha). \tag{6.14}$$

(ii) *The matrix function*

$$(t, \alpha, \theta) \to P(t, \alpha, \theta) : [0, T] \times I(-h, 0) \times I(-h, 0) \to \mathscr{L}(\mathbb{R}^n, \mathbb{R}^n) \tag{6.15}$$

is continuous and bounded.

(iii) *There exists a constant* $c > 0$ *such that*

$$\sup \left\{ \int_{-h}^{0} |P(t, \theta, \alpha)|^2\, d\theta \,|(t, \alpha) \in [0, T] \times I(-h, 0) \right\} \leqq c, \tag{6.16}$$

$$\sup \left\{ \int_{-h}^{0} |P(t, \theta, \alpha)|^2\, d\alpha \,|(t, \theta) \in [0, T] \times I(-h, 0) \right\} \leqq c, \tag{6.17}$$

$$\sup \left\{ \int_{-h}^{0} \left| \frac{\partial P}{\partial \alpha}(t, \theta, \alpha) \right|^2 d\alpha \,|(t, \theta) \in [0, T] \times I(-h, 0) \right\} \leqq c, \tag{6.18}$$

$$\sup \left\{ \int_{-h}^{0} \left| \frac{\partial P}{\partial \theta}(t, \theta, \alpha) \right|^2 d\theta \,|(t, \alpha) \in [0, T] \times I(-h, 0) \right\} \leqq c. \tag{6.19}$$

When h *is finite the matrix function*

$$\left.\begin{aligned} &(t, \alpha, \theta) \to P(t, \alpha, \theta), \\ &(t, \alpha, \theta) \to \frac{\partial P}{\partial \alpha}(t, \alpha, \theta), \\ &(t, \alpha, \theta) \to \frac{\partial P}{\partial \theta}(t, \alpha, \theta), \end{aligned}\right\} : [0, T] \times I(-h, 0) \times I(-h, 0) \to \mathscr{L}(\mathbb{R}^n, \mathbb{R}^n), \tag{6.20}$$

all belong to $L^2([0, T] \times I(-h, 0) \times I(-h, 0); \mathscr{L}(\mathbb{R}^n, \mathbb{R}^n))$.

(iv) *For each* ϕ *in* M^2 *the function*

$$t \to P_J(t) J^* j^* \phi : [0, T] \to H^1 = H^1(-h, 0; \mathbb{R}^n) \tag{6.21}$$

is continuous and the functions

$$t \to P_J(t) J^* j^* \phi : [0, T] \to L^2 = L^2(-h, 0; \mathbb{R}^n), \tag{6.22}$$

$$t \to (P_J(t) J^* j^* \phi)(0) : [0, T] \to \mathbb{R}^n, \tag{6.23}$$

belong to $H^1(0, T; L^2)$ *and* $H^1(0, T; \mathbb{R}^n)$, *respectively.*

(v) *For each* $\phi = (\phi^0, \phi^1)$ *in* M^2, t *in* $[0, T]$ *and* α *in* $I(-h, 0)$

$$(P_J(t) J^* j^* \phi)(\alpha) = P(t, 0, \alpha)\phi^0 + \int_{-h}^{0} P(t, \theta, \alpha)\phi^1(\theta)\, d\theta. \tag{6.24}$$

Moreover the following functions are continuous

$$t \to P(t, 0, \alpha)\phi^0 + \int_{-h}^{0} P(t, \theta, \alpha)\phi^1(\theta)\, d\theta : [0, T] \to \mathbb{R}^n, \tag{6.25}$$

$$t \to P(t, 0, \cdot)\phi^0 + \int_{-h}^{0} P(t, \theta, \cdot)\phi^1(\theta)\, d\theta : [0, T] \to L^2, \tag{6.26}$$

$$t \to DP(t,0,\cdot)\phi^0 + D\int_{-h}^{0} P(t,\theta,\cdot)\phi^1(\theta)\,d\theta : [0,T] \to L^2, \tag{6.27}$$

where

$$\left[DP(t,0,\cdot)\phi^0 + D\int_{-h}^{0} P(t,\theta,\cdot)\phi^1(\theta)\,d\theta\right](\alpha) = \frac{\partial P}{\partial \alpha}(t,0,\alpha)\phi^0 + \frac{\partial}{\partial \alpha}\int_{-h}^{0} P(t,\theta,\alpha)\phi^1(\theta)\,d\theta; \tag{6.28}$$

the functions

$$t \to \frac{\partial P}{\partial t}(t,0,0)\phi^0 + \frac{\partial}{\partial t}\int_{-h}^{0} P(t,\theta,0)\phi^1(\theta)\,d\theta : [0,T] \to \mathbb{R}^n, \tag{6.29}$$

$$t \to \frac{\partial P}{\partial t}(t,0,\cdot)\phi^0 + \frac{d}{dt}\int_{-h}^{0} P(t,\theta,\cdot)\phi^1(\theta)\,d\theta : [0,T] \to L^2, \tag{6.30}$$

belong to $L^2(0,T;\mathbb{R}^n)$ *and* $L^2(0,T;L^2)$, *respectively.*

(vi) *For each* ϕ^0 *in* $\mathbb{R}^n$ *and* α *in* $I(-h,0)$, *the following functions are continuous*

$$t \to P(t,\alpha,0)\phi^0 : [0,T] \to \mathbb{R}^n, \tag{6.31}$$

$$t \to P(t,\cdot,0)\phi^0 : [0,T] \to L^2, \tag{6.32}$$

$$t \to DP(t,\cdot,0)\phi^0 : [0,T] \to L^2, \tag{6.33}$$

where

$$[DP(t,\cdot,0)\phi^0](\alpha) = \frac{\partial P}{\partial \alpha}(t,\alpha,0)\phi^0; \tag{6.34}$$

the function

$$t \to \frac{\partial P}{\partial t}(t,\cdot,0)\phi^0 : [0,T] \to L^2 \tag{6.35}$$

belongs to $L^2(0,T;L^2)$. *Moreover for each* ϕ *and* ψ *in* $L^2(-h,0;\mathbb{R}^n)$

$$\int_{-h}^{0} \phi(\alpha)\cdot\frac{\partial}{\partial \alpha}\int_{-h}^{0} P(t,\theta,\alpha)\psi(\theta)\,d\theta\,d\alpha = \int_{-h}^{0}\int_{-h}^{0} \frac{\partial P}{\partial \alpha}(t,\alpha,\theta)\phi(\alpha)\,d\alpha\cdot\psi(\theta)\,d\theta. \tag{6.36}$$

When ψ *belongs to* $L^2(-h,0;\mathbb{R}^n)\cap L^1(-h,0;\mathbb{R}^n)$,

$$\int_{-h}^{0} \phi(\alpha)\cdot\frac{\partial}{\partial \alpha}\int_{-h}^{0} P(t,\theta,\alpha)\psi(\theta)\,d\theta\,d\alpha = \int_{-h}^{0} \phi(\alpha)\cdot\int_{-h}^{0} \frac{\partial P}{\partial \alpha}(t,\theta,\alpha)\psi(\theta)\,d\theta\,d\alpha; \tag{6.37}$$

in particular when h *is finite,* $L^2(-h,0;\mathbb{R}^n)\cap L^1(-h,0;\mathbb{R}^n) = L^2(-h,0;\mathbb{R}^n)$ *and the above identity is true for all* ϕ *and* ψ *in* $L^2(-h,0;\mathbb{R}^n)$.

(vii) *The continuous linear operator* $\Pi(t) = jP(t)j^*$ *on* M^2 *can be decomposed as a matrix of operators*

$$\Pi(t) = \begin{bmatrix} \Pi^{00}(t) & \Pi^{01}(t) \\ \Pi^{10}(t) & \Pi^{11}(t) \end{bmatrix} \tag{6.38}$$

where

$$\Pi^{00}(t)\in\mathscr{L}(\mathbb{R}^n,\mathbb{R}^n),\qquad \Pi^{01}(t)\in\mathscr{L}(L^2,\mathbb{R}^n),$$
$$\Pi^{10}(t)\in\mathscr{L}(\mathbb{R}^n,L^2),\qquad \Pi^{11}(t)\in\mathscr{L}(L^2,L^2). \tag{6.39}$$

For each t in $[0,T]$, α in $I(-h,0)$ and $\phi=(\phi^0,\phi^1)$ in M^2

$$\Pi^{00}(t)=P(t,0,0),\qquad [\Pi^{10}(t)\phi^0](\alpha)=P(t,0,\alpha)\phi^0, \tag{6.40}$$

$$\Pi^{01}(t)\phi^1=\int_{-h}^{0}P(t,\theta,0)\phi^1(\theta)\,d\theta=\int_{-h}^{0}P(t,0,\theta)^*\phi^1(\theta)\,d\theta, \tag{6.41}$$

$$[\Pi^{11}(t)\phi^1](\alpha)=\int_{-h}^{0}P(t,\theta,\alpha)\phi^1(\theta)\,d\theta=\int_{-h}^{0}P(t,\alpha,\theta)^*\phi^1(\theta)\,d\theta. \tag{6.42}$$

Proof. Cf. Appendix to § 6. □

Remark 6.1. The case $h=+\infty$ introduces additional difficulties since the matrix functions (6.20) need not be L^2 with respect to (t,α,θ). Moreover identity (6.37) is not necessarily true and only identity (6.36) can be used in the derivation of the differential equations for $P(t,\alpha,\theta)$.

6.2. Coupled differential equations for $P(t,\alpha,\theta)$. Knowing the properties of $P(t,\alpha,\theta)$ and its relation to $P(t)$, we can now use the Riccati differential equation (5.45) of Theorem 5.5 to obtain a set of differential equations for $P(t,\alpha,\theta)$.

THEOREM 6.6. *The matrix function $P(t,\alpha,\theta)$ satisfies the following set of differential equations (in the sense of distributions):*

$$\begin{aligned}&\frac{\partial P}{\partial t}(t,0,0)+\int_{-h}^{0}d_\theta\eta^T P(t,0,\theta)+\int_{-h}^{0}P(t,\theta,0)\,d_\theta\eta,\\ &-\int_{-h}^{0}P(t,\zeta,0)\,d_\zeta\beta N^{-1}\int_{-h}^{0}d_\xi\beta^T P(t,0,\xi)+Q=0,\\ &P(T,0,0)=0,\end{aligned} \tag{6.43}$$

$$\begin{aligned}&\left(\frac{\partial}{\partial t}+\frac{\partial}{\partial\theta}\right)P(t,\theta,0)+\int_{-h}^{0}d_\alpha\eta^T P(t,\theta,\alpha)\\ &-\int_{-h}^{0}P(t,\zeta,0)\,d_\zeta\beta N^{-1}\int_{-h}^{0}d_\xi\beta^T P(t,\theta,\xi)=0,\\ &P(T,\theta,0)=0,\qquad \theta\in I(-h,0),\end{aligned} \tag{6.44}$$

$$\begin{aligned}&\left(\frac{\partial}{\partial t}+\frac{\partial}{\partial\theta}+\frac{\partial}{\partial\alpha}\right)P(t,\alpha,\theta)-\int_{-h}^{0}P(t,\zeta,\theta)\,d_\zeta\beta N^{-1}\int_{-h}^{0}d_\xi\beta^T P(t,\alpha,\xi)=0,\\ &P(T,\alpha,\theta)=0\quad\forall\alpha,\theta\ \textit{in}\ I(-h,0),\end{aligned} \tag{6.45}$$

$$P(t,\alpha,\theta)^*=P(t,\theta,\alpha)\quad\forall t\in[0,T]\quad\forall\alpha,\theta\in I(-h,0). \tag{6.46}$$

The matrix function

$$(t,\theta)\to\frac{\partial P}{\partial t}(t,\theta,0):[0,T]\times I(-h,0)\to\mathscr{L}(\mathbb{R}^n,\mathbb{R}^n) \tag{6.47}$$

is an L^2-matrix function. When h is finite the matrix function

$$(t, \alpha, \theta) \to \frac{\partial P}{\partial t}(t, \alpha, \theta) : [0, T] \times I(-h, 0) \times I(-h, 0) \to \mathscr{L}(\mathbb{R}^n, \mathbb{R}^n) \tag{6.48}$$

is an L^2-matrix function.

Proof. Cf. Appendix to § 6. □

Remark 6.2. When $h = +\infty$ the matrix function $P(t, \alpha, \theta)$ is not necessarily an L^2-matrix function.

COROLLARY 6.7. *Let $a > 0$ and $b > 0$ be two finite real numbers and $h = \max\{a, b\}$. Let B and L be of the form*

$$Bw = B_0 w(0) + B_1 w(-a), \tag{6.49}$$

$$L\phi = A_0 \phi(0) + A_1 \phi(-b). \tag{6.50}$$

Then the system of equations (6.43) *to* (6.45) *reduces to*

$$\begin{aligned}
&\frac{\partial P}{\partial t}(t, 0, 0) + A_0^T P(t, 0, 0) + A_1^T P(t, 0, -b) + P(t, 0, 0) A_0 + P(t, -b, 0) A_1, \\
&-[P(t, 0, 0) B_0 + P(t, -a, 0) B_1] N^{-1} [B_0^T P(t, 0, 0) + B_1^T P(t, 0, -a)] = 0, \qquad (6.51) \\
&P(T, 0, 0) = 0,
\end{aligned}$$

$$\begin{aligned}
&\left(\frac{\partial}{\partial t} + \frac{\partial}{\partial \theta}\right) P(t, \theta, 0) + A_0^T P(t, \theta, 0) + A_1^T P(t, \theta, -b), \\
&-[P(t, 0, 0) B_0 + P(t, -a, 0) B_1] N^{-1} [B_0^T P(t, \theta, 0) + B_1^T P(t, \theta, -a)] = 0, \qquad (6.52) \\
&P(T, \theta, 0) = 0 \quad \forall \theta \in I(-h, 0),
\end{aligned}$$

$$\begin{aligned}
&\left(\frac{\partial}{\partial t} + \frac{\partial}{\partial \alpha} + \frac{\partial}{\partial \theta}\right) P(t, \alpha, \theta) - [P(t, 0, \theta) B_0 + P(t, -a, \theta) B_1] N^{-1} \\
&\cdot [B_0^T P(t, \alpha, 0) + B_1^T P(t, \alpha, -a)] = 0, \qquad (6.53) \\
&P(T, \alpha, \theta) = 0 \quad \forall \alpha, \theta \in I(-h, 0).
\end{aligned}$$

"Modulo" an appropriate change of variables, the above system of equations coincides with the one of Koivo and Lee [1] and R. H. Kwong [3, (2.29)-(2.31)].

7. The linear-quadratic optimal control problem on $[0, \infty[$. We now associate with equation (5.3) on $[0, \infty[$

$$\begin{aligned}
&\frac{dx}{dt}(t) = Lx_t + Bu_t \text{ in } [0, \infty[, \ x \in H^1_{\text{loc}}(0, \infty; \mathbb{R}^n), \\
&(x(0), x_0, u_0) = (\phi^0, \phi^1, w) \in \mathbb{R}^n \times L^2(-h, 0; \mathbb{R}^n) \times L^2(-h, 0; \mathbb{R}^n),
\end{aligned} \tag{7.1}$$

the cost function

$$J(u, \phi^0, \phi^1, w) = \int_0^\infty [|C^0 x(t)|^2 + Nu(t) \cdot u(t)]\, dt, \tag{7.2}$$

where C^0 is an arbitrary $r \times n$ matrix ($r \geqq 1$, an integer).

Define $Q = (C^0)^T C^0$. The infinite time horizon optimal control problem consists in minimizing the functional (7.2) over all u in $L^2(0, \infty; \mathbb{R}^m)$:

$$\text{Inf}\{J(u, \phi^0, \phi^1, w) | u \in L^2(0, \infty; \mathbb{R}^m)\}. \tag{7.3}$$

7.1. Stabilizability. The fundamental hypothesis to ensure the existence of a unique solution to problem (7.3) is the stabilizability hypothesis with respect to the observation operator.

DEFINITION 7.1. The pair (L, B) is *stabilizable with respect to* C^0 if for each ϕ in M^2 and $w=0$, there exists $u \in L^2(0, \infty; \mathbb{R}^m)$ such that

$$\int_0^\infty |C^0 x(t)|^2 \, dt < \infty. \tag{7.4}$$

□

We have seen that system (7.1) can be reformulated in state space and embedded in a larger family of state equations

$$[D_t - (A_V^T)^*] l^* y = (B^T J^{-1})^* u, \text{ in } [0, \infty[, \qquad (l^* y)(0) = y^0 \text{ in } V'. \tag{7.5}$$

The cost function (7.2) becomes

$$\hat{J}(u, y^0) = \int_0^\infty |Cy(t)|^2 + Nu(t) \cdot u(t) \, dt \tag{7.6}$$

where $C: W' \to \mathbb{R}^r$ is defined as

$$C(w_0, w_1) = C^0 w_0. \tag{7.7}$$

Define $\hat{Q}: W' \to W$ as $\hat{Q}(w_0, w_1) = ((C^0)^T C^0 w_0, 0)$. So we can associate with problem (7.5)-(7.6), the following concept of stabilizability.

DEFINITION 7.2. The pair $((A_V^T)^*, (B^T)^*)$ is stabilizable with respect to the observation C if for each y^0 in V', there exists u in $L^2(0, \infty; \mathbb{R}^m)$ such that

$$\int_0^\infty |Cy(t)|^2 \, dt < \infty. \tag{7.8}$$

□

Remark 7.1. Definition 7.1 (resp. Definition 7.2) says that for each initial condition ϕ in M^2 (resp. y^0 in V'), there exists a control function u in $L^2_{\text{loc}}(0, \infty; \mathbb{R}^m)$ for which the cost function is finite or, equivalently,

$$\int_0^\infty |u(s)|^2 \, ds < \infty \quad \text{and} \quad \int_0^\infty |C^0 x(t)|^2 \, dt < \infty \left(\text{resp.} \int_0^\infty |Cy(t)|^2 \, dt < \infty \right).$$

At this stage we do not want to speculate on the properties of an eventual feedback operator and the well-posedness of the closed loop system. We shall later see (cf. Theorem 7.5) that the hypothesis of Definition 7.1 is equivalent to the existence of a closed-loop operator of the form

$$K = -N^{-1} B^T J^{-1} P: V' \to \mathbb{R}^m, \qquad P \in \mathcal{L}(V', V) \tag{7.9}$$

for which the closed-loop system

$$[D_t - (A_V^T)^* + (B^T J^{-1})^* N^{-1} B^T J^{-1} P] l^* y = 0, \qquad (l^* y)(0) = y^0 \in V' \tag{7.10}$$

is well posed and its solution is exponentially stable for all y^0 in V'. □

When h is finite we have the following extension of the result in Vinter and Kwong [1].

THEOREM 7.3. *When h is finite, Definitions* 7.1 *and* 7.2 *are equivalent.*

Proof. It is clear that Definition 7.2 implies Definition 7.1. It suffices to choose $y^0 = j^*(\phi^0, H\phi^1 + Kw)$. The converse of the theorem follows from the next lemma which is an extension of Proposition 7.2 in Vinter and Kwong [1]. □

LEMMA 7.4. *When h is finite, there exist $\bar{\phi} = (\bar{\phi}^0, \bar{\phi}^1)$ in M^2 such that*

$$(l^* y)(2h) = j^*(\bar{\phi}^0, H\bar{\phi}^1), \tag{7.11}$$

where y is the solution of (7.5) with initial condition y^0 in V' and $u(t) = 0$ in $[0, 2h]$.

7.2. The family of operators $P_T(s)$. In § 5 we have fixed the final time $T > 0$ and embedded our problem in an internal $[s, T]$, $0 \leqq s \leqq T$. More precisely we have minimized the cost function (5.5) associated with the state equation (5.6). In Theorem 5.2 we have defined a family of decoupling operators $\{P_T(s) | 0 \leqq s \leqq T\}$ which we shall now index with the superscript T. This family is completely characterized by Theorem 5.2(i)a):

$$\begin{aligned} &[D_t - (A_V^T)^*] l^* \phi = -R\psi \text{ in } [s, T], \qquad (l^*\phi)(s) = k, \\ &[D_t + A^T]\psi + \hat{Q}\phi = 0 \text{ in } [s, T], \qquad \psi(T) = 0, \\ &P_T(s)k = \psi(s). \end{aligned} \tag{7.12}$$

Denote by $(\bar{\phi}, \bar{\psi})$ the solution of (7.12) for the initial condition $\bar{k}$. Using the integration by parts formula (4.38)

$$\langle (l^*\phi)(s), \bar{\psi}(s)\rangle_V = \langle (l^*\phi)(T), \bar{\psi}(T)\rangle_V - \int_s^T [\langle \phi, [D_t + A^T]\bar{\psi}\rangle_W + \langle [D_t - (A_V^T)^*] l^*\phi, \bar{\psi}\rangle_V]\, dt$$

and

$$\langle k, P_T(s)\bar{k}\rangle_V = \int_s^T [\langle \phi, \hat{Q}\bar{\phi}\rangle_W + \langle R\psi, \bar{\psi}\rangle_V]\, dt. \tag{7.13}$$

But recall that the optimal controls are of the form

$$u = -N^{-1}B^T J^{-1}\psi, \qquad \bar{u} = -N^{-1}B^T J^{-1}\bar{\psi};$$

thus (7.13) can be rewritten in the form

$$\langle k, P_T(s)\bar{k}\rangle_V = \int_s^T [\langle \phi, \hat{Q}\bar{\phi}\rangle_W + u \cdot N\bar{u}]\, dt. \tag{7.14}$$

In particular

$$\langle k, P_T(s)k\rangle_V = \hat{J}_s^T(u, k) = \operatorname{Inf}\{\hat{J}_s^T(v, k) | v \in L^2(s, T; \mathbb{R}^m)\}. \tag{7.15}$$

It is not difficult but fundamental to notice that for all s in $[0, T[$ and all h in V'

$$\operatorname{Inf}\{\hat{J}_s^T(v, k) | v \in L^2(s, T; \mathbb{R}^m)\} = \operatorname{Inf}\{\hat{J}_0^{T-s}(w, k) | w \in L^2(0, T-s; \mathbb{R}^m)\}. \tag{7.16}$$

As a result for all $T_1 \geqq s_1 \geqq 0$ and all $T_2 \geqq s_2 \geqq 0$

$$T_2 - s_2 = T_1 - s_1 \Rightarrow P_{T_2}(s_2) = P_{T_1}(s_1), \tag{7.17}$$

$$T_2 - s_2 \geqq T_1 - s_1 \Rightarrow P_{T_2}(s_2) \geqq P_{T_1}(s_1). \tag{7.18}$$

7.3. Solution of the optimal control problem on $[0, \infty[$. We summarize our results in a single theorem.

THEOREM 7.5. *Assume that the pair (L, B) is stabilizable with respect to C^0.*

(i) *For each k in V' there exists a unique u in $L^2(0, \infty; \mathbb{R}^m)$ such that*

$$\hat{J}(u, k) = \operatorname{Inf}\{\hat{J}(v, k) | v \in L^2_{\text{loc}}(0, \infty; \mathbb{R}^m)\}. \tag{7.19}$$

(ii) *There exists a positive symmetrical operator P in $\mathscr{L}(V', V)$,*

$$\forall k, \bar{k} \in V', \quad \langle k, P\bar{k}\rangle_V = \langle \bar{k}, Pk\rangle_V, \tag{7.20}$$

$$\forall k \in V', \quad \langle k, Pk\rangle_V \geqq 0, \tag{7.21}$$

such that

$$\forall t \geqq 0, \quad \forall k \in V', \quad P_T(t)k \to Pk \text{ in } V\text{-strong as } T \text{ goes to } +\infty. \tag{7.22}$$

(iii) *The optimal control is of the form*

$$u(t) = -N^{-1}B^T J^{-1} P(l^* y)(t), \qquad t \geqq 0, \tag{7.23}$$

where y is the solution in

$$\mathscr{Z}_{\text{loc}}(0, \infty) = \{v : [0, \infty[\to W' \mid \forall T > 0,\ v|_{[0,T]} \in \mathscr{Z}(0, T)\} \tag{7.24}$$

of the equation

$$[D_t - (A_V^T)^*](l^* y) + RP(l^* y) = 0, \quad 0 \leqq t, \quad (l^* y)(0) = k. \tag{7.25}$$

Moreover for all k and $\bar{k}$ in V'

$$\langle k, P\bar{k}\rangle_V = \int_0^\infty [Cy(t) \cdot C\bar{y}(t) + u(t) \cdot N\bar{u}(t)]\, dt, \tag{7.26}$$

$$\langle k, P\bar{k}\rangle_V = \int_0^\infty [Cy(t) \cdot C\bar{y}(t) + \langle RP(l^* y)(t), P(l^* \bar{y})(t)\rangle_V\, dt, \tag{7.27}$$

where $\bar{u}$ and $\bar{y}$ are the optimal control and the solution to (7.25) *corresponding to the initial condition $\bar{k}$.*

(iv) *P is a positive symmetrical solution of the Riccati equation*

$$A^T Pj^* + jP(A^T)^* - jPRPj^* + \tilde{Q} = 0 \text{ in } \mathscr{L}(M^2, M^2), \tag{7.28}$$

where $\tilde{Q} \in \mathscr{L}(M^2, M^2)$ is defined as $\tilde{Q}(\phi^0, \phi^1) = (Q\phi^0, 0)$.

(v) *P is the minimal positive symmetrical solution of* (7.28). *That is, any other positive symmetrical solution $\bar{P}$ of* (7.28) *is such that*

$$\forall k \in V', \quad \langle k, \bar{P}k\rangle_V \geqq \langle k, Pk\rangle_V. \tag{7.29}$$

Proof. The Proof will be given in the Appendix. □

It is also possible to give conditions to ensure the uniqueness of the minimum positive solution to the Riccati equation (7.28) in terms of the stabilizability of the pair $(L^T, C^0)^T)$ with respect to the identity matrix in $\mathbb{R}^n$ (when the length h of the memory is finite). The reader is referred to Vinter and Kwong [1] and L. Pandolfi [2]. Necessary and sufficient conditions can be obtained by using techniques developed by M. Sorine [1], [2].

7.4. Equations for the kernel of the operator P. We proceed exactly as in § 6.

DEFINITION 7.6. The operator P_J in $\mathscr{L}((H^1)', H^1)$ is defined as

$$P_J = J^{-1}P(J^*)^{-1}. \tag{7.30}$$

□

Its properties are analogous to those of P:

$$\begin{aligned} &\forall \bar{k}, k \in (H^1)', \quad \langle \bar{k}, P_J k\rangle_{H^1} = \langle k, P_J \bar{k}\rangle_{H^1}, \\ &\forall k \in (H^1)', \quad \langle k, P_J k\rangle_{H^1} \geqq 0. \end{aligned} \tag{7.31}$$

DEFINITION 7.7. For each $(\alpha, \theta) \in I(-h, 0) \times I(-h, 0)$, let $P(\alpha, \theta)$ be the $n \times n$ matrix function defined as follows:

$$(e, d) \to d \cdot P(\alpha, \theta)e = \langle \delta_\theta d, P_J(\delta_\alpha e)\rangle_{H^1} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}. \tag{7.32}$$

□

THEOREM 7.8. (i) *For all* $(\alpha, \theta) \in I(-h, 0) \times I(-h, 0) \to \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$

$$P(\alpha, \theta)^* = P(\theta, \alpha). \tag{7.33}$$

(ii) *The matrix function*

$$(\alpha, \theta) \to P(\alpha, \theta) : I(-h, 0) \times I(-h, 0) \to \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n) \tag{7.34}$$

is bounded and continuous.

(iii) *There exists a constant* $c > 0$ *such that*

$$\sup\left\{\int_{-h}^{0} |P(\theta, \alpha)|^2 \, d\theta \,|\, \alpha \in I(-h, 0)\right\} \leqq c, \tag{7.35}$$

$$\sup\left\{\int_{-h}^{0} |P(\alpha, \theta)|^2 \, d\alpha \,|\, \theta \in I(-h, 0)\right\} \leqq c, \tag{7.36}$$

$$\sup\left\{\int_{-h}^{0} \left|\frac{\partial P}{\partial \alpha}(\theta, \alpha)\right|^2 d\alpha \,|\, \theta \in I(-h, 0)\right\} \leqq c, \tag{7.37}$$

$$\sup\left\{\int_{-h}^{0} \left|\frac{\partial P}{\partial \theta}(\theta, \alpha)\right|^2 d\theta \,|\, \alpha \in I(-h, 0)\right\} \leqq c. \tag{7.38}$$

When h is finite the matrix functions

$$\left.\begin{aligned} &(\alpha, \theta) \to P(\alpha, \theta), \\ &(\alpha, \theta) \to \frac{\partial P}{\partial \alpha}(\alpha, \theta), \\ &(\alpha, \theta) \to \frac{\partial P}{\partial \theta}(\alpha, \theta), \end{aligned}\right\} : I(-h, 0) \times I(-h, 0) \to \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n), \tag{7.39}$$

all belong to $L^2(I(-h, 0) \times I(-h, 0); \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n))$.

(iv) *For each* $\phi = (\phi^0, \phi^1)$ *in* M^2 *and* α *in* $I(-h, 0)$

$$(P_J J^* j^* \phi)(\alpha) = P(0, \alpha)\phi^0 + \int_{-h}^{0} P(\theta, \alpha)\phi^1(\theta) \, d\theta. \tag{7.40}$$

For all ϕ *and* ψ *in* $L^2(-h, 0; \mathbb{R}^n)$

$$\int_{-h}^{0} \psi(\alpha) \cdot \frac{\partial}{\partial \alpha} \int_{-h}^{0} P(\theta, \alpha)\phi(\theta) \, d\theta \, d\alpha = \int_{-h}^{0} \int_{-h}^{0} \frac{\partial P}{\partial \alpha}(\alpha, \theta)\psi(\alpha) \, d\alpha \cdot \phi(\theta) \, d\theta. \tag{7.41}$$

When ϕ *belongs to* $L^2(-h, 0; \mathbb{R}^n) \cap L^1(-h, 0; \mathbb{R}^n)$

$$\int_{-h}^{0} \psi(\alpha) \cdot \frac{\partial}{\partial \alpha} \int_{-h}^{0} P(\theta, \alpha)\phi(\theta) \, d\theta \, d\alpha = \int_{-h}^{0} \psi(\alpha) \cdot \int_{-h}^{0} \frac{\partial P}{\partial \alpha}(\theta, \alpha)\phi(\theta) \, d\theta \, d\alpha; \tag{7.42}$$

in particular, when h is finite, $L^2(-h, 0; \mathbb{R}^n) \cap L^1(-h, 0; \mathbb{R}^n) = L^2(-h, 0; \mathbb{R}^n)$ *and the above identity is true for all* ϕ *and* ψ *in* $L^2(-h, 0; \mathbb{R}^n)$.

(v) *The continuous linear operator* $\Pi = jPj^*$ *in* $\mathcal{L}(M^2, M^2)$ *can be decomposed as a matrix of operators as in Theorem* 6.5(vi).

Proof. Similar to the proof of Theorem 6.5. □

THEOREM 7.9. *The matrix $P(\alpha, \theta)$ satisfies the following set of differential equations*

$$(7.43)\quad \int_{-h}^{0} d_\theta \eta^T P(0,\theta) + \int_{-h}^{0} P(\theta,0)\, d_\theta \eta - \int_{-h}^{0} P(\zeta,0)\, d_\zeta \beta N^{-1} \int_{-h}^{0} d_\xi \beta^T P(0,\xi) + Q = 0,$$

$$(7.44)\quad \frac{\partial P}{\partial \theta}(\theta,0) + \int_{-h}^{0} d_\alpha \eta^T P(\theta,\alpha) - \int_{-h}^{0} P(\zeta,0)\, d_\zeta \beta N^{-1} \int_{-h}^{0} d_\xi \beta^T P(\theta,\xi) = 0,$$

$$(7.45)\quad \left(\frac{\partial}{\partial \theta} + \frac{\partial}{\partial \alpha}\right) P(\alpha,\theta) - \int_{-h}^{0} P(\zeta,\theta)\, d_\zeta \beta N^{-1} \int_{-h}^{0} d_\xi \beta^T P(\alpha,\xi) = 0,$$

$$(7.46)\quad P(\alpha,\theta)^* = P(\theta,\alpha) \quad \forall \alpha, \theta \in I(-h,0).$$

When h is finite, the elements of the matrix function $(\alpha,\theta) \to \partial P/\partial\alpha(\alpha,\theta)$ and $\partial P/\partial\theta(\alpha,\theta)$ belong to L^2.

Proof. Similar to proof of Theorem 6.6 □

Appendix to Section 2.

Proof of Lemma 2.1. In the proof we shall use results from W. Rudin [1, Chap. 6] on complex measures which includes, as a special case, real measures. We only provide a proof for B; the argument for L is the same with obvious changes.

(i) By the Riesz representation theorem (cf. W. Rudin [1, p. 131, Thm. 6.19]).

(ii) The map $t \to v_t : [0,a] \to C_0(-h,0;\mathbb{R}^m)$,

$$v_t : I(-h,0) \to \mathbb{R}^m, \qquad v_t(\theta) = v(t+\theta),$$

is continuous since the functions v in $C_c(-h,a;\mathbb{R}^m)$ are uniformly continuous. So the function $\mathcal{B}u$ is continuous since B is continuous on $C_0(-h,0;\mathbb{R}^m)$. Hence the map (2.5) is linear and continuous.

(iii) For arbitrary functions f in $C(0,a;\mathbb{R}^n)$ and u in $C_c(-h,a;\mathbb{R}^m)$ consider the integral

$$(1)\qquad \int_0^a d_t m f(t) \cdot \int_{-h}^{0} d_\theta \beta u(t+\theta) = \sum_{i=1}^{n} \sum_{j=1}^{m} \int_0^a d_t m f_i(t)\, d_\theta \beta_{ij} u_j(t+\theta),$$

where the f_i's, β_{ij}'s and u_j's are the respective components of f, β and u and m denotes the Lebesgue measure on $[0,a]$. For each pair i,j the integrand is continuous in $[0,a]\times I(-h,0)$ and hence $(m\times\beta_{ij})$-measurable. Moreover it is also integrable since

$$\int_0^a d_t m \int_{-h}^{0} |f(t)|\, d_\theta|\beta|\, |u(t+\theta)| \leqq |\beta|([-h,0]\cap\mathbb{R})\|f\|_{L^1}\|u\|_C < \infty,$$

where $|\beta|$ is the total variation of the matrix of measures β. By the Riesz representation theorem and the definition of $|\beta|$, the quantity $|\beta|(I(-h,0))$ is finite. So the conditions of Fubini's theorem (cf. W. Rudin [1, p. 140, Thm. 7.8]) are met and the order of integration can be changed in (1):

$$(f, \mathcal{B}v) = \sum_{i=1}^{n} \sum_{j=1}^{m} \int_{-h}^{0} d_\theta \beta_{ij} \int_0^a d_t m f_i(t) v_j(t+\theta).$$

Each element in the above double summation can be estimated separately:

$$\left|\int_{-h}^{0} d_\theta \beta_{ij} \int_0^a d_t m f_i(t) v_j(t+\theta)\right| \leqq \int_{-h}^{0} d_\theta |\beta_{ij}| \int_0^a d_t m |f_i(t)|\,|v_j(t+\theta)|$$

$$\leqq \int_{-h}^{0} d_\theta |\beta_{ij}|\,\|f_i\|_{L^2(0,a)} \|v_j\|_{L^2(\theta,a+\theta)}$$

$$\leqq |\beta_{ij}|(I(-h,0))\|f_i\|_{L^2(0,a)} \|v_j\|_{L^2(-h,a)}.$$

Finally

$$|(f, \mathcal{B}v)| \leqq |\beta|(I(-h,0))\|f\|_2 \|v\|_2.$$

This new estimate combined with the density of $C(0, a; \mathbb{R}^n)$ in $L^2(0, a; \mathbb{R}^n)$ (cf. W. Rudin [1, p. 68, Thm. 3.14]) yields

$$\|\mathcal{B}v\|_{L^2(0,a)} \leqq |\beta|(I(-h,0))\|v\|_{L^2(-h,a)}.$$

Note that the L^2-norm of v is with respect to the Lebesgue measure. By density of $C_c(-h, a; \mathbb{R}^m)$ in $L^2(-h, a; \mathbb{R}^m)$ (cf. W. Rudin [1, p. 68, Thm. 3.14]) the map (2.5) extends by continuity to a continuous linear map on all $L^2(-h, a; \mathbb{R}^m)$. □

Proof of Lemma 2.3. We already know that an element x of $H^1(-\infty, 0; \mathbb{R}^n)$ is almost everywhere equal to a unique bounded continuous function $\bar{x}$ in $C(-\infty, 0; \mathbb{R}^n)$. We also know that the injection $x \to \bar{x}$ of $H^1(-\infty, 0; \mathbb{R}^n)$ into $C(-\infty, 0; \mathbb{R}^n)$ is continuous (cf. Adams [1, p. 97, Thm. 5.4]).

It remains to prove that $\bar{x}$ belongs to the closed subspace $C_0(-\infty, 0; \mathbb{R}^n)$ of $C(-\infty, 0; \mathbb{R}^n)$. For each pair $-\infty < s' \leqq t \leqq 0$

$$\bar{x}(t) = \bar{x}(s') + \int_{s'}^{t} Dx(r)\, dr$$

and

$$(t-s)^{1/2}|\bar{x}(t)| \leqq \|\bar{x}\|_{L^2(s,t;\mathbb{R}^n)} + (t-s)\|Dx\|_{L^2(s,t;\mathbb{R}^n)}.$$

But $\bar{x} = x$ in $L^2(s, t; \mathbb{R}^n)$ and for $s = t-1$

$$|\bar{x}(t)| \leqq \sqrt{2}\|x\|_{H^1(t-1,t;\mathbb{R}^n)} \leqq \sqrt{2}\|x\|_{H^1(-\infty,t;\mathbb{R}^n)}.$$

But for each $\varepsilon > 0$, there exists $T < 0$ such that

$$\forall t \leqq T, \qquad \|x\|_{H^1(-\infty,t;\mathbb{R}^n)} \leqq \varepsilon.$$

In view of the above two inequalities

$$\forall \varepsilon > 0, \exists t < 0 \quad \text{such that } \forall t \leqq T, \qquad |\bar{x}(t)| \leqq \varepsilon.$$

This shows that $\bar{x}(t)$ goes to zero as t goes to $-\infty$ and proves that $\bar{x}$ belongs to $C_0(-\infty, 0; \mathbb{R}^n)$. □

Proof of Lemma 2.6. (i) Same proof as in Lemma 2.1(ii).
(ii) Choose z and x in $C_c(s, t; \mathbb{R}^n)$

$$\begin{aligned}
\int_s^t z(t-r)\cdot(\mathscr{L}e_+^s x)(r)\,dr &= \int_s^t z(t-r)\cdot\int_{-h}^0 d_\theta\eta(e_+^s x)(r+\theta)\,dr\\
&= \int_{-h}^0\int_s^t z(t-z)\cdot d_\theta\eta(e_+^s x)(r+\theta)\,dr\\
&= \int_{-h}^0\int_s^t (e_+^0 z)(t-r)\cdot d_\theta\eta(e_+^s x)(r+\theta)\,dr\\
&= \int_{-h}^0\int_{s+\theta}^{t+\theta} (e_+^0 z)(t-r+\theta)\cdot d_\theta\eta(e_+^s x)(r)\,dr\\
&= \int_{-h}^0\int_s^t d_\theta\eta^T(e_+^0 z)_{t-r}(\theta)\cdot(e_+^s x)(r)\,dr\\
&= \int_s^t L^T(e_+^0 z)_{t-r}\cdot(e_+^s x)(r)\,dr\\
&= \int_s^t (\mathscr{L}^T e_+^0 z)(t-r)\cdot(e_+^s x)(r)\,dr. \qquad \square
\end{aligned}$$

Proof of Theorem 2.7. (i) For $t=s$ the result is obvious. Fix the pair $0\leqq s<t$. Let x be the solution of (2.32):

$$\begin{aligned}
&\frac{dx}{dr}(r) = [\mathscr{L}e_+^s x + \mathscr{B}e_+^s u + f^0](r) + (e_+^{-h}\xi^1)(s-r), \qquad s\leqq r\leqq t,\\
&x(s) = \xi^0.
\end{aligned} \tag{2}$$

For an arbitrary $\psi=(\psi^0,\psi^1)$ in $V=\mathscr{D}(A^T)$, let z be the solution in $H^1_{\mathrm{loc}}(0,\infty;\mathbb{R}^n)$ of the equation

$$\frac{dz}{dr}(r) = L^T z_r, \quad r\geqq 0, \quad (z(0), z_0)=\psi. \tag{3}$$

We know that for $r\geqq 0$

$$(z(r), z_r) = S^T(r)\psi. \tag{4}$$

Inner product (2) with $z(t-r)$ on $[s,t]$:

$$\int_s^t z(t-r)\cdot[-\dot{x}(r)+(\mathscr{L}e_+^s x)(r)+(\mathscr{B}e_+^s u)(r)+f^0(r)+(e_+^{-h}\xi^1)(s-r)]\,dr=0. \tag{5}$$

Integrate by parts and use identities (2.41)

$$\begin{aligned}
0 = &\int_{-s}^t [-\dot{z}(t-r)+(\mathscr{L}^T e_+^0 z)(t-r)]\cdot x(r)\,dr\\
&+\int_s^t [(\mathscr{B}^T e_+^0 z)(t-r)\cdot u(r)+z(t-r)\cdot f^0(r)]\,dr\\
&+\int_s^t z(t-r)\cdot(e_+^{-h}\xi^1)(s-r)\,dr - z(0)\cdot x(t)+z(t-s)\cdot x(s)
\end{aligned}$$

or

$$\int_s^t [(\mathscr{L}^T e_-^0 z)(t-r)\cdot x(r)+(\mathscr{B}^T e_-^0 z)(t-r)\cdot u(r)]\,dr+\psi^0\cdot x(t)$$

$$(6)\qquad =\int_s^t [(\mathscr{B}^T z)(t-r)\cdot u(r)+z(t-r)\cdot f^0(r)\,dr$$

$$+z(t-s)\cdot\xi^0+\int_s^t z(t-r)\cdot(e_+^{-h}\xi^1)(s-r)\,dr.$$

But $e_-^0 z=\psi^1$ and after a change of variable the left-hand side of the above inequality becomes

$$\psi^0\cdot x(t)+\int_{-(t-s)}^0 [(\mathscr{L}^T e_-^0\psi^1)(-\alpha)x(t+\alpha)+(\mathscr{B}^T e_-^0\psi^1)(-\alpha)\cdot u(t+\alpha)]\,d\alpha$$

$$(7)\qquad =\psi^0\cdot x(t)+(H^T\psi^1,(e_+^s x)_t)+(K^T\psi^1,(e_+^s u)_t).$$

In addition

$$\int_s^t z(t-r)\cdot(e_+^{-h}\xi^1)(s-r)\,dr=\int_{s-t}^0 (e_+^0 z)(t-s+\alpha)\cdot(e_+^{-h}\xi^1)(\alpha)\,d\alpha$$

$$(8)\qquad =((e_+^0 z)_{t-s},\xi^1)=(z_{t-s},\xi^1)-((e_-^0\psi^1)_{t-s},\xi^1)$$

and

$$((e_-^0\psi^1)_{t-s},\xi^1)=\int_{-h}^0 (e_-^0\psi^1)(t-s+\alpha)\cdot\xi^1(\alpha)\,d\alpha$$

$$(9)\qquad =\int_{-h}^0 \psi^1(\beta)(e_+^{-h}\xi^1)(s-t+\beta)\,d\beta=(\psi^1,\mathscr{E}(\xi^1)(t-s)).$$

By combining (6)-(9) we obtain

$$\psi^0\cdot x(t)+(\psi^1,H(e_+^s x)_t+K(e_+^s u)_t+\mathscr{E}(\xi^1)(t-s)$$

$$=\int_s^t [B^T z_{t-r}\cdot u(r)+z(t-r)\cdot f^0(r)]\,dr+z(t-s)\cdot\xi^0+(z_{t-s},\xi^1)$$

or using the definition (2.34) for $\hat{x}(t)$ and (2.42) for j

$$((\psi,\hat{x}(t))=((S^T(t-s)j\psi,\xi))$$

$$(10)\qquad +\int_s^t [B^T J^{-1}S_V^T(t-r)\psi\cdot u(r)+S^T(t-r)j\psi\cdot(f^0(r),0)]\,dr.$$

By definition of j in (2.42)

$$(11)\qquad S^T(t)j\psi=jS_V^T(t)\psi\quad\forall\psi\in V=\mathscr{D}(A^T),$$

and identity (2.43) can be obtained from identity (10) which is true for all ψ in V.

(ii) We first show that $\hat{x}$ belongs to $C(s,T;M^2)$. By construction, the $\mathbb{R}^n$-component of $\hat{x}(t)$ belongs to $H^1(s,T;\mathbb{R}^n)$ and, a fortiori, to $C(s,T;\mathbb{R}^n)$. We only need to show the continuity of the L^2-component of $\hat{x}(t)$. Pick any t' and t in $[s,t]$:

$$\|\hat{x}^1(t')-\hat{x}^1(t)\|_2\leqq\|H((e_+^s(x))_{t'}-(e_+^s(x))_t)\|_2+\|K((e^s(u))_{t'}-(e_+^s(u))_t)\|_2$$

$$+\left[\int_{-h}^0 |(\mathscr{E}(\xi^1))(t'-s-\theta)-(\mathscr{E}(\xi^1))(t-s-\theta)|^2\,d\theta\right]^{1/2}.$$

The function $e_+^{-h}\xi^1$ is the extension by 0 of the function ξ^1 defined on $I(-h,0)$ to $]-\infty,0]$. For all $r\geqq 0$ we define the right shift

$$(e_+^{-h}\xi^1)_{-r}: I(-h,0)\to\mathbb{R}^n,\quad (e_+^{-h}\xi^1)_{-r}(\theta)=(e_+^{-h}\xi^1)(\theta-r),\quad \theta\in I(-h,0).$$

As a result for all $r\geqq 0$ and $\theta\in I(-h,0)$

$$\mathscr{E}(\xi^1)(r-\theta)=(e_+^{-h}\xi^1)(\theta-r).$$

By continuity of H and K, we obtain the following first upper bound:

$$c[\|(e_+^s(x))_{t'}-(e_+^s(x))_t\|_2+\|(e_+^s(u))_{t'}-(e_+^s(u))_t\|_2+\|(\xi_-^1)_{s-t'}-(\xi_-^1)_{s-t}\|_2].$$

But the square bracket only contains terms which can be considered as right or left shifts of L^2-functions. The continuity of $\hat{x}^1(t)$ now follows from the continuity of the shift operator.

Next compute the vectorial distributional derivative of $j^*\hat{x}(t)$. For all ψ in V

$$\langle\psi, j^*\hat{x}(t)\rangle_V=\langle\psi, j^*①+②\rangle_V,$$

where

$$①=\left[S^T(t-s)^*\xi+\int_s^t S^T(t-r)^*(f^0(r),0)\,dr\right]\in M^2,$$

$$②=\int_s^t S_V^T(t-r)^*(B^TJ^{-1})^*u(r)\,dr\in V'.$$

In view of the definitions and results of § 2.2, the function $t\to\langle\psi, j^*①\rangle_V=((j\psi,①))$ belongs to $H^1(s,T;\mathbb{R})$ and

$$\frac{d}{dt}\langle\psi, j^*①\rangle_V=((A^T\psi,①))+((j\psi,(f^0(t),0)))$$

for all ψ in $\mathscr{D}(A^T)=V$. For all ψ in $\mathscr{D}(A_V^T)$, the function $t\to\langle\psi,②\rangle_V$ belongs to $H^1(s,T;\mathbb{R})$ and

$$\frac{d}{dt}\langle\psi,②\rangle_V=\int_s^t\langle S_V^T(t-r)A_V^T\psi,(B^TJ^{-1})^*u(r)\rangle_V\,dr+\langle\psi,(B^TJ^{-1})^*u(t)\rangle_V$$
$$=\langle A_V^T\psi,②\rangle_V+\langle\psi,(B^TJ^{-1})^*u(t)\rangle_V.$$

Finally for all ψ in $\mathscr{D}(A_V^T)$, $A^T\psi=jA_V^T\psi$ and

$$\frac{d}{dt}\langle\psi, j^*①+②\rangle_V=((jA_V^T\psi,①))+((j\psi,(f^0(t),0)))+\langle A_V^T\psi,②\rangle_V+\langle\psi,(B^TJ^{-1})^*u(t)\rangle_V$$
$$=\langle A_V^T\psi, j^*①+②\rangle_V+\langle\psi,(B^TJ^{-1})^*u(t)+j^*(f^0(t),0)\rangle_V.$$

But

$$\langle A_V^T\psi, j^*①+②\rangle_V=\langle A_V^T\psi, j^*\hat{x}(t)\rangle_V=((jA_V^T\psi,\hat{x}(t)))=((A^T\psi,\hat{x}(t))),$$

and

$$\frac{d}{dt}\langle\psi, j^*\hat{x}(t)\rangle_V=\langle\psi,(A^T)^*\hat{x}(t)+(B^TJ^{-1})^*u(t)+j^*(f^0(t),0)\rangle_V. \tag{12}$$

By density of $\mathscr{D}(A_V^T)$ in V the above equation is true for all ψ in V and the vectorial

distributional derivative $D_t j^*\hat{x}$ of $j^*\hat{x}$ is the element of $L^2(s, T; V')$ given by

$$\frac{d}{dt}(j^*\hat{x})(t) = (A^T)^*\hat{x}(t) + (B^T J^{-1})^* u(t) + j^*(f^0(t), 0). \tag{13}$$

At this stage we have shown that $\hat{x}$ belongs to $\mathcal{W}(s, t; M^2, V')$ and that $\hat{x}$ is a solution of (13) with $\hat{x}(s) = \xi$. The proof of uniqueness follows standard arguments and will be omitted.

(iii) Equation (2.49) follows from (12) (ii).

(iv) By definition of B, the map $B^T: C_0(-h, 0; \mathbb{R}^n) \to \mathbb{R}^n$ is given by the expression

$$B^T\phi = B_0^T\phi(0) + \int_{-h}^{0} B_1(\theta)^T\phi(\theta)\, d\theta;$$

by definition of B_M^T and all ϕ in $H^1(-h, 0; \mathbb{R}^n)$

$$B^T J^{-1}\phi = B_M^T j\phi \Rightarrow B^T J^{-1} = B_M^T j.$$

Making use of the above identity, we get

$$S_V^T(t-r)^*(B^T J^{-1})^* = S_V^T(t-r)^*(B_M^T j)^* = S_V^T(t-r)^* j^*(B_M^T)^* = j^* S^T(t-r)^*(B_M^T)^*.$$

Then the substitution of the last identity in (2.46) yields

$$\hat{x}(t) = S^T(t-s)^*\xi + \int_s^t S^T(t-r)^*[(B_M^T)^* u(r) + (f^0(r), 0)]\, dr.$$

Identities (2.51) and (2.52) are obtained from identities (2.45) and (2.46) by a similar argument. □

Appendix to Section 4.

Proof of Theorem 4.5. (i) From Lemma 4.4 by transposing (4.15).

(ii) We first establish identity (4.33). For some arbitrary ψ in V, let v be the solution in $\mathcal{V}(0, T)$ of the equation

$$\frac{djv}{dt} + A^T v = 0, \quad v(T) = \psi \Rightarrow v(t) = S_V^T(T-t)\psi \quad \forall t \in [0, T].$$

Upon substitution in (4.30) we get

$$\langle z_T, \psi\rangle_V = \int_0^T \langle f(t), S_V^T(T-t)\psi\rangle_V\, dt + \langle \xi, S_V^T(T)\psi\rangle_V$$

$$= \left\langle S_V^T(T)^*\psi + \int_0^T S_V^T(T-t)^* f(t)\, dt, \psi \right\rangle_V.$$

Since ψ is arbitrary we obtain (4.33). To establish identity (4.32) fix an arbitrary g in $L^2(0, T; V)$ and construct the solution v in $\mathcal{V}(0, T)$ of the equation

$$\frac{djv}{dt} + A^T v + jg = 0, \qquad v(T) = 0 \Rightarrow v(t) = \int_t^T S_V^T(r-t)g(r)\, dr.$$

By direct substitution in (4.30)

$$\int_0^T \langle z(t), lg(t)\rangle_W\, dt = \int_0^T \left\langle f(t), \int_t^T S_V^T(r-t)g(r)\, dr \right\rangle_V dt + \left\langle \xi, \int_0^T S_V^T(r)g(r)\, dr \right\rangle_V$$

$$= \int_0^T \left\langle S_V^T(r)^*\xi + \int_0^r S_V^T(r-t)^* f(t)\, dt, g(r) \right\rangle_V dr.$$

Since the last identity is true for all g we necessarily obtain (4.32).

(iii) Denote by $\mathscr{A}$ the isomorphism (4.24) in Lemma 4.3. From parts (i) and (ii) for all (f, ξ) in $L^2(0, T; V') \times V'$ there exists a unique x in $\mathscr{Z}(0, T)$ such that

$$x(t) = l^* z(t), \quad \text{a.e. in } [0, T] \quad \text{and} \quad (f, \xi) = \mathscr{A}x = \mathscr{A}l^* z.$$

Moreover, always from parts (i) and (ii), the map

$$(f, \xi) \to z: L^2(0, T; V') \times V' \to \mathscr{Z}(0, T)$$

is linear, injective and continuous. As a result $\mathscr{A}l^*$ is surjective, linear and, from (4.31), $(\mathscr{A}l^*)^{-1}$ is continuous. Therefore $\mathscr{A}l^*$ is an isomorphism.

(iv) The integration by parts formula follows from parts (i) and (iii) by choosing

$$f = \left[\frac{d}{dt} - (A^T)^*\right](l^* z) \quad \text{and} \quad \xi = (l^* z)(0)$$

in the variational equation (4.30). □

Proof of Theorem 4.6. (i) Given v in $C(0, T; V)$, let $\tilde{v}$ be the unique solution in $\mathscr{V}(0, T) \subset C(0, T; V)$ of the differential equation

$$-\left[\frac{d}{dt} + A^T\right]\tilde{v} = lKv + g, \qquad \tilde{v}(T) = \psi$$

for arbitrary g in $L^2(0, T; W)$ and ψ in V. Denote by Γ the map

$$v \to \Gamma v = \tilde{v}; \qquad C(0, T; V) \to C(0, T; V).$$

By construction, it is linear and continuous. We show that the map Γ is a contraction and hence has a unique fixed point. Given v_1 and v_2 in $C(0, T; V)$, we obtain, by linearity, that $\tilde{v}_2 - \tilde{v}_1$ is the unique solution of the equation

$$-\left[\frac{d}{dt} + A^T\right](\tilde{v}_2 - \tilde{v}_1) = lK(v_2 - v_1), \qquad \tilde{v}_2(T) - \tilde{v}_1(T) = 0.$$

This last equation is equivalent to

$$\tilde{v}_2(t) - \tilde{v}_1(t) = \int_t^T S_V^T(s - t)K(s)[v_2(s) - v_1(s)]\, ds, \qquad 0 \leqq t \leqq T.$$

By uniform boundedness of the operators $\{K(t): 0 \leqq t \leqq T\}$ and $\{S_V^T(t): 0 \leqq t \leqq T\}$

$$\|\tilde{v}_2(t) - \tilde{v}_1(t)\|_V \leqq c \int_t^T \|v_2(s) - v_1(s)\|_V\, ds, \qquad 0 \leqq t \leqq T.$$

This is sufficient to conclude that Γ is a contraction (standard argument). In particular there exists a unique v in $C(0, T; V)$, and a fortiori in $\mathscr{V}(0, T)$, such that $\Gamma v = v$: v is the unique solution in $\mathscr{V}(0, T)$ of the equation

$$-\left[\frac{d}{dt} + A^T\right]v = lKv + g, \qquad v(T) = \psi.$$

From this we conclude that the continuous linear map (4.42) is bijective. By the Banach inverse mapping theorem, it is an isomorphism.

(ii) An argument similar to the one of part (i) can be used to prove that the map (4.44) is an isomorphism. □

Appendix to Section 5.

Proof of Lemma 5.3. The equivalence of (i) and (ii) is a direct consequence of identity (5.18). We show that (i) is equivalent to (iii). Substitute $v=\psi$ and $z=\beta$ in the integration by parts formula (4.38) of Theorem 4.5(iv) on the time interval $[s, T]$

$$\langle (l^*\beta)(s), \psi(s)\rangle_V = \langle (l^*\beta)(T), \psi(T)\rangle_V + \int_s^T \langle \beta, -[D_t + A^T]\psi\rangle_W \, dt$$

$$- \int_s^T \langle [D_t - (A_V^T)^*] l^*\beta, \psi\rangle_V \, dt.$$

In view of (5.19) and (5.21) the above identity reduces to

$$0 = \int_s^T [\langle \beta, \hat{Q}\phi\rangle_W + \langle R\gamma - f, \psi\rangle_V] \, dt. \tag{1}$$

Similarly substitute $v=\gamma$ and $z=\phi$ in (4.38) on the time interval $[s, T]$

$$\langle (l^*\phi)(s), \gamma(s)\rangle_V = \langle (l^*\phi)(T), \gamma(T)\rangle_V + \int_s^T \langle \phi, -[D_t + A^T]\gamma\rangle_W \, dt$$

$$- \int_s^T \langle [D_t - (A_V^T)^*] l^*\phi, \gamma\rangle_V \, dt$$

and

$$\langle h, r(s)\rangle_V = \int_s^T [\langle \phi, \hat{Q}\beta + q\rangle_W + \langle R\psi, \gamma\rangle_V] \, dt. \tag{2}$$

Using the symmetry of $\hat{Q}$ and R and substracting (1) from (2) we obtain

$$\langle h, r(s)\rangle_V = \int_s^T [\langle \phi, q\rangle_W + \langle f, \psi\rangle_V] \, dt \tag{3}$$

for all s in $[0, T]$ and all h in V'. This last identity proves the equivalence of (i) and (iii).

To complete the proof of the lemma, recall that ψ and ϕ are necessarily related as follows:

$$\psi(t) = P(t)(l^*\phi)(t), \qquad s \leqq t \leqq T,$$

(cf. identity (5.18) in Theorem 5.2 applied to system (5.19)-(5.20).) Thus

$$\langle f(t), \psi(t)\rangle_V = \langle f(t), P(t)(l^*\phi)(t)\rangle_V = \langle (l^*\phi)(t), P(t)f(t)\rangle_V$$

$$= \langle l^*\phi(t), P(t)f(t)\rangle_V = \langle \phi(t), lP(t)f(t)\rangle_W$$

and when $q(t) = -lP(t)f(t)$, the right-hand side of (3) is identically null for all h and s. This shows that $r=0$ and completes the proof of the lemma. □

Proof of Theorem 5.4. (i) Pick any x in $\mathscr{Z}(0, T)$ and consider the equation

$$-[D_t + A^T]p + lPRp = \hat{Q}x - lP[D_t - (A_V^T)^*]l^*x, \qquad p(T) = 0. \tag{4}$$

By the perturbation Theorem 4.6 there exists a unique solution p in $\mathscr{V}(0, T)$ since the right-hand side belongs to $L^2(0, T; W)$. Define the function f, q and the vector ξ

$$f = [D_t - (A_V^T)^*]l^*x + Rp, \quad q = -lPf, \quad \xi = (l^*x)(0).$$

Then the pair (x, p) is the solution of the system

$$-[D_t + A^T]p = \hat{Q}x + q, \quad p(T) = 0, \quad [D_t - (A_V^T)^*](l^*x) = -Rp + f, \quad (l^*x)(0) = \xi.$$

But this is precisely the Hamiltonian system where q and f are such that $q = -lPf$. By Lemma 5.3 this implies that

$$p(t) = P(t)(l^*x)(t).$$

The substitution of the right-hand side of the last identity into (4) yields (5.29).

To prove uniqueness, let Z be a family in the class $(\mathscr{P})$ such that for all x in $\mathscr{X}(0, T)$ the function Zl^*x is the unique solution in $\mathscr{V}(0, T)$ to the equation

$$(5)\quad [D_t + A^T]Zl^*x - lZ[D_t - (A_V^T)^*]l^*x - lZRZl^*x + \hat{Q}x = 0, \qquad (Zl^*x)(T) = 0.$$

By substracting (5.29) from (5) we obtain

$$(6)\quad \begin{aligned} &\{[D_t + A^T] - lPR\}(Z-P)l^*x - l(Z-P)\{[D_t - (A_V^T)^*] + RP\}l^*x \\ &\quad - l(Z-P)R(Z-P)l^*x = 0, \\ &[(Z-P)l^*x](T) = 0. \end{aligned}$$

For all x in $\mathscr{X}(0, T)$, $(Z-P)l^*x$ belongs to $\mathscr{V}(0, T)$. We use integration by parts formula (4.38) on $[t, T]$ to obtain

$$\begin{aligned} &\langle x, \{[D_t + A^T] - lPR\}(Z-P)l^*x\rangle_{L^2(t,T;W)} \\ &\quad = -\langle\{[D_t - (A_V^T)^*] + RP\}l^*x, (Z-P)l^*x\rangle_{L^2(t;T;V)} \\ &\qquad - \langle(l^*x)(t), (Z-P)(t)(l^*x)(t)\rangle_V. \end{aligned}$$

Finally from the last identity and (6)

$$(7)\quad \begin{aligned} &-2\langle\{[D_t - (A_V^T)^*] + RP\}l^*x, (Z-P)l^*x\rangle_{L^2(t,T;V)} \\ &\quad = \langle R(Z-P)l^*x, (Z-P)l^*x\rangle_{L^2(t,T;V)} \\ &\qquad + \langle(l^*x)(t), (Z-P)(t)(l^*x)(t)\rangle_V \end{aligned}$$

where we have used the fact that $Z(t)$ and $P(t)$ are symmetrical. For an arbitrary h in V', construct the solution x in $\mathscr{V}(t, T; W', V')$ to

$$\{[D_t - (A_V^T)^*] + RP\}l^*x = 0 \quad \text{in } [t, T], \qquad (l^*x)(t) = h$$

(this is possible by the perturbation Theorem 4.6). Substitute this x in (7) to obtain

$$(8)\quad \langle h, [Z(t) - P(t)]h\rangle_V \leqq 0 \quad \forall h \in V'.$$

Now repeat the same steps with P and Z interchanged. The end result will be

$$(9)\quad \langle h, [P(t) - Z(t)]h\rangle_V \leqq 0 \quad \forall h \in V'.$$

Since $P(t) - Z(t)$ is a positive symmetrical operator, for all h and k in V'

$$\begin{aligned} 4\langle h, [P(t) - Z(t)]k\rangle_V &= \langle h + k, [P(t) - Z(t)](h + k)\rangle_V - \langle h - k, [P(t) - Z(t)](h - k)\rangle_V \\ &= 0, \end{aligned}$$

since from (8) and (9)

$$\langle h, [P(t) - Z(t)]h\rangle_V = 0 \quad \forall h \in V'.$$

This shows that for all t in $[0, T]$, $Z(t) = P(t)$. Therefore P is unique in the class $(\mathscr{P})$.

(ii) The continuity of the linear map Pl^* defined by (5.28) is a direct consequence of the continuity of the solution p to (4) with respect to the data and the uniform boundedness of P in $[0, T]$.

(iii) From identity (5.18) in Theorem 5.2(i)

$$r(t) = p(t) - P(t)(l^*y)(t).$$

The function r belongs to $\mathcal{V}(0, T)$ since both p and Pl^*y do. Recall that

$$-[D_t + A^T]p = \hat{Q}y + q, \qquad p(T) = 0. \tag{10}$$

Substitute $x = y$ in (5.29)

$$-[D_t + A^T](Pl^*y) = \hat{Q}y - lPRPl^*y - lP[D_t - (A_V^T)^*]l^*y, \qquad (Pl^*y)(T) = 0. \tag{11}$$

By substracting (11) from (10)

$$-[D_t + A^T]r = lP\{[D_t - (A_V^T)^*]l^*y + RPl^*y\} + q, \qquad r(T) = 0. \tag{12}$$

But from identities (5.18) and (5.16)

$$RPl^*y = R(p - r), \qquad [D_t - (A_V^T)^*](l^*y) = -Rp + f,$$

the expression between curly brackets in (12) reduces to $f - Rr$ and (12) finally reduces to (5.30). By the perturbation Theorem 4.6, we know that the solution r to (5.30) is unique. □

Proof of Theorem 5.5. (i) When x belongs to $\mathcal{W}(0, T; M^2, V')$, $i^*x \in \mathcal{Z}(0, T)$ and $l^*(i^*x) = j^*x$ (since $j = il$). So $P(l^*i^*x) = P(j^*x)$ is the unique solution in $\mathcal{V}(0, T)$ to the equation

$$[D_t + A^T](Pj^*x) - lP[D_t - (A_V^T)^*]j^*x - lPRPj^*x + \hat{Q}i^*x = 0, \qquad (Pj^*x)(T) = 0. \tag{13}$$

Equation (5.31) is now obtained from (13) by having the map i act on each term of (13). The following identities are used

$$i\hat{Q}i^*x = \tilde{Q}x, \qquad ilP[\ ] = jP[\ ],$$

$$i[D_t + A^T](Pj^*x) = \frac{d}{dt}(jPj^*x) + A^TPj^*x,$$

$$[D_t - (A_V^T)^*]j^*x = \frac{d}{dt}j^*x - (A^T)^*x.$$

(Cf. Lemma 5.1, identities (4.14) and (4.28) and Lemma 4.3.) The uniqueness argument is similar to the one in the proof of Theorem 5.4(i).

So far we have proved that the statement of Theorem 5.4(i) implies part (i) of the theorem. Conversely we can go back from (5.31) to (13) and use the density of $\{i^*x | x \in \mathcal{W}(0, T; M^2, V')\}$ in $\mathcal{Z}(0, T)$ to recover (5.29). We just sketch the proof. The linear subspace $i^*C(0, T; M^2)$ of $L^2(0, T; W')$ is dense in $L^2(0, T; W')$. So the linear subspace

$$S = \{x \in \mathcal{Z}(0, T) | \exists z \in C(0, T; M^2) \text{ such that } i^*z = x\}$$

is dense in

$$E = \{x \in \mathcal{Z}(0, T) | \exists z \in L^2(0, T; W') \text{ such that } x = z\} = \mathcal{Z}(0, T).$$

But it is not too difficult to see that

$$S = i^*\mathcal{W}(0, T; M^2, V') \qquad \text{(as a topological vector space)}.$$

This is sufficient to prove the equivalence of part (i) and Theorem 5.4(i).

(ii) Equation (5.32) is obtained from (5.31) by choosing functions x of the form

$$x(t)=h, \quad 0\leqq t\leqq T, \quad h\in M^2.$$

We now show that the property

(I) $\qquad \forall h\in M^2, \qquad t\to P(t)j^*h$ belongs to $\mathcal{V}(0, T)$

implies the property

(II) $\quad \forall x\in \mathcal{W}(0, T; M^2, V'), \qquad t\to P(t)j^*x(t)$ belongs to $\mathcal{V}(0, T)$.

For each $x\in \mathcal{W}(0, T; M^2, V')\subset C(0, T; M^2)$ and from property (I) the function $t\to P(t)j^*x(t)$ belongs to $C(0, T; V)$. Moreover the function

$$t\to \frac{d}{dt}j^*x(t)$$

belongs to $L^2(0, T; V')$.

Therefore for almost all t in $[0, T]$

$$\frac{d}{dt}[jP(t)j^*x(t)]=\frac{d}{dt}[jP(t)j^*h]_{h=x(t)}+jP(t)\frac{d}{dt}(j^*x(t)). \tag{14}$$

But from (5.32)

$$-\frac{d}{dt}(jP(t)j^*h)|_{h=x(t)}=[A^TP(t)j^*+jP(t)(A^T)^*-jP(t)RP(t)j^*+\tilde{Q}]x(t)$$

and necessarily $d/dt(jPj^*x)$ belongs to $L^2(0, T; M^2)$. Furthermore

$$\frac{d}{dt}(jP(t)j^*x(t))+A^TP(t)j^*x(t)$$

$$=-[jP(t)(A^T)^*-jP(t)RP(t)j^*+\tilde{Q}]x(t)+jP(t)\frac{d}{dt}j^*x(t).$$

But $j=il$, $\tilde{Q}=i\hat{Q}i^*$ and

$$\frac{d}{dt}(jPj^*x)+A^TPj^*x=iF$$

where $F\in L^2(0, T; W)$ is given by the expression

$$F=l\left\{P\frac{d}{dt}(j^*x)-P[(A^T)^*-RPj^*]x\right\}-\hat{Q}i^*x.$$

There exists a unique $y\in \mathcal{V}(0, T; V, W)$ such that

$$[D_t+A^T]y=F, \qquad y(T)=0.$$

Therefore

$$\frac{d}{dt}(jPj^*x)+A^TPj^*x=i[D_t+A^T]y=\frac{d}{dt}jy+A^Ty, \qquad (Pj^*x)(T)=0=y(T).$$

By uniqueness of solution in $\mathcal{W}(0, T; V, M^2)$, $y=Pj^*x\in \mathcal{V}(0, T)$. This proves that (I) implies (II). The converse is obvious. Now to complete the proof we must show that for an arbitrary x in $\mathcal{W}(0, T; V, M^2)$ we can go from (5.32) to (5.31). Set $h=x(t)$ in

(5.32) and recall from (14) that

$$\frac{d}{dt}(jP(t)j^*h)|_{h=x(t)}=\frac{d}{dt}(jP(t)j^*x(t))-jP(t)\frac{d}{dt}[j^*x(t)].$$

This yields (5.31) and completes the equivalence of property (II) and (5.31) with property (I) and (5.32). The uniqueness property of P for (ii) follows from the uniqueness property for (i). This completes the proof of the theorem. □

Appendix to Section 6.

Proof of Theorem 6.5. (i) From the first property (6.8)

$$\forall e, d\in\mathbb{R}^n, \qquad \langle\delta_\theta d, P_J(t)\delta_\alpha e\rangle_{H^1}=\langle\delta_\alpha e, P_J(t)\delta_\theta d\rangle_{H^1},$$
$$d\cdot P(t,\alpha,\theta)e=e\cdot P(t,\theta,\alpha)d.$$

This proves identity (6.14).

(ii) For all $\bar t$ and t in $[0, T]$, $\bar h$, h, k and $\bar k$ in $(H^1)'$

$$\langle\bar h, P_J(\bar t)\bar k\rangle_{H^1}-\langle h, P_J(t)k\rangle_{H^1}=\langle\bar h-h, P_J(\bar t)[\bar k-k]\rangle_{H^1}+\langle h, [P_J(\bar t)-P_J(t)]k\rangle_{H^1}$$
$$+\langle\bar h-h, P_J(\bar t)k\rangle_{H^1}+\langle h, P_J(t)(\bar k-k)\rangle_{H^1}.$$

In particular for $\bar h=\delta_{\bar\theta}d$, $h=\delta_\theta d$, $\bar k=\delta_{\bar\alpha}e$, $k=\delta_\alpha e$,

$$|d\cdot P(\bar t,\bar\alpha,\bar\theta)e-d\cdot P(t,\alpha,\theta)e|$$
$$\leqq c\|(\delta_{\bar\theta}-\delta_\theta)d\|_{(H^1)'}\|(\delta_{\bar\alpha}-\delta_\alpha)e\|_{(H^1)'}+|\langle\delta_\theta d, P_J(\bar t)\delta_\alpha e\rangle-\langle\delta_\theta d, P_J(t)\delta_\alpha e\rangle|$$
$$+c\|(\delta_{\bar\theta}-\delta_\theta)d\|_{(H^1)'}\|\delta_\alpha e\|_{(H^1)'}+c\|\delta_\theta d\|_{(H^1)'}\|(\delta_{\bar\alpha}-\delta_\alpha)e\|_{(H^1)'}$$

where we have used the boundedness of $P_J(t)$ (cf. third property (6.8)). In view of Lemma 6.4,

$$\|(\delta_{\bar\theta}-\delta_\theta)d\|_{(H^1)'}\leqq|d||\bar\theta-\theta|^{1/2},$$
$$\|(\delta_{\bar\alpha}-\delta_\alpha)e\|_{(H^1)'}\leqq|e||\bar\alpha-\alpha|^{1/2},$$
$$\exists c'>0, \quad \forall\alpha\in I(-h,0), \quad \|\delta_\alpha e\|_{(H^1)'}\leqq c'|e|.$$

As a result

$$|d\cdot P(\bar t,\bar\alpha,\bar\theta)e-d\cdot P(t,\alpha,\theta)e|$$
$$\leqq|\langle\delta_\theta d, P_J(\bar t)\delta_\alpha e\rangle-\langle\delta_\theta d, P_J(t)\delta_\alpha e\rangle|+c|d||e||\bar\theta-\theta|^{1/2}|\bar\alpha-\alpha|^{1/2}$$
$$+c|d||e||\bar\theta-\theta|^{1/2}+c|d||e||\bar\alpha-\alpha|^{1/2},$$

where $c>0$ stands for a generic constant. From the last property (6.8) the function $t\to\langle\delta_\theta d, P_J(t)\delta_\alpha c\rangle_{H^1}$ is continuous. As $(\bar t,\bar\alpha,\bar\theta)$ goes to (t,α,θ) the right-hand side of the above inequality goes to zero. This shows that each component of the matrix $P(t,\alpha,\theta)$ and a fortiori the matrix $P(t,\alpha,\theta)$ itself is jointly continuous with respect to its arguments. To show it is bounded, consider

$$|d\cdot P(t,\alpha,\theta)e|=|\langle\delta_\theta d, P_J(t)\delta_\alpha e\rangle_{H^1}|.$$

We know that the norm of the operator $P_J(t)$ is uniformly bounded with respect to t (cf. third property (6.8)); we also know that

$$\exists c'>0 \text{ such that } \|\delta_\theta d\|_{(H^1)'}\leqq c'|d|, \qquad \|\delta_\alpha e\|_{(H^1)'}\leqq c'|e|$$

(cf. inequality (6.12) in Lemma 6.4). As a result there exists a constant $c>0$ (independent of t, α, θ, d and e) such that

$$\forall d, e \in \mathbb{R}^n, \qquad |d \cdot P(t, \alpha, \theta)e| \leqq c|d||e|.$$

(iii) In view of the symmetry (6.14) it is sufficient to prove (6.16) and (6.18): the estimates (6.17) and (6.19) will follow from (6.16) and (6.18), respectively. For each e in $\mathbb{R}^n$, t in $[0, T]$ and θ in $I(-h, 0)$ the function $P_J(t)\delta_\theta e$ belongs to $H^1 = H^1(-h, 0; \mathbb{R}^n)$. Moreover

$$\|P_J(t)\delta_\theta e\|_{H^1} \leqq \|P_J(t)\|_{\mathscr{L}((H^1)', H^1)} \|\delta_\theta e\|_{(H^1)'} \leqq cc'|e|$$

(from the fourth property (6.8) and inequality (6.12) in Lemma 6.4). But for all d in $\mathbb{R}^n$ and α in $I(-h, 0)$

$$d \cdot P(t, \theta, \alpha)e = \langle \delta_\alpha d, P_J(t)\delta_\theta e\rangle_{H^1} = d \cdot (P_J(t)\delta_\theta e)(\alpha)$$

and

$$(P_J(t)\delta_\theta e)(\alpha) = P(t, \theta, \alpha)e. \tag{1}$$

In particular

$$\|P_J(t)\delta_\theta e\|_{L^2}^2 + \|DP_J(t)\delta_\theta e\|_{L^2}^2 = \|P_J(t)\delta_\theta e\|_{H^1}^2 \leqq c''|e|^2$$

or in view of (1)

$$\int_{-h}^{0} |P(t, \theta, \alpha)e|^2 \, d\alpha + \int_{-h}^{0} \left|\frac{\partial P}{\partial \alpha}(t, \theta, \alpha)e\right|^2 d\alpha \leqq c''|e|^2. \tag{2}$$

Estimates (6.16) and (6.18) now follow through a chain of standard estimates

(iv) Recall that the following sequence of maps and injection is linear and continuous (cf. Theorem 5.4(ii))

$$\mathscr{W}(0, T; M^2, V') \xrightarrow{i^*} \mathscr{Z}(0, T) \xrightarrow{Pl^*} \mathscr{V}(0, T) \hookrightarrow \mathscr{W}(0, T; V, M^2).$$

In particular all functions x of the form

$$x(t) = (\phi^0, \phi^1), \qquad 0 \leqq t \leqq M^2,$$

for some $\phi = (\phi^0, \phi^1)$ in M^2, belong to $\mathscr{W}(0, T; M^2, V')$ and the function $Pj^*x = Pl^*i^*x$ belongs to $\mathscr{W}(0, T; V, M^2)$, that is

$$Pj^*x \in C(0, T; V), \qquad jPj^*x \in H^1(0, T; M^2).$$

By definition of $P_J(t)$, $P_J(t) = J^{-1}P(t)(J^*)^{-1}$,

$$P_J J^* j^* x = J^{-1} Pj^* x \in C(0, T; H^1), \tag{3}$$

$$jJP_J J^* j^* x \in H^1(0, T; M^2). \tag{4}$$

But for each t in $[0, T]$

$$jJP_J(t)J^*j^*\phi = ([P_J(t)J^*j^*\phi](0), P_J(t)J^*j^*\phi).$$

Therefore

$$P_J J^* j^* x \in H^1(0, T; L^2) \tag{5}$$

and the function

$$t \to [P_J(t)J^*j^*\phi](0) : [0, T] \to \mathbb{R}^n \tag{6}$$

belongs to $H^1(0, T; \mathbb{R}^n)$. The properties of the functions (6.21), (6.22) and (6.23) follow from (3), (5) and (6).

(v) We prove (6.24) and apply the results of part (iv) to obtain identities (6.25) to (6.30). For all d in $\mathbb{R}^n$ and α in $I(-h, 0)$

$$\begin{aligned} d \cdot (P_J(t)J^*j^*\phi)(\alpha) &= \langle \delta_\alpha d, P_J(t)J^*j^*\phi\rangle_{H^1} \\ &= \langle J^*j^*\phi, P_J(t)\delta_\alpha d\rangle_{H^1} \\ &= ((\phi, jJP_J(t)\delta_\alpha d)) \\ &= \phi^0 \cdot [P_J(t)\delta_\alpha d](0) + \int_{-h}^{0} \phi^1(\theta) \cdot [P_J(t)\delta_\alpha d](\theta)\, d\theta \\ &= \langle \delta_0\phi^0, P_J(t)\delta_\alpha d\rangle_{H^1} + \int_{-h}^{0} \langle \delta_\theta\phi^1(\theta), P_J(t)\delta_\alpha d\rangle_{H^1}\, d\theta \\ &= \phi^0 \cdot P(t, \alpha, 0)d + \int_{-h}^{0} \phi^1(\theta) \cdot P(t, \alpha, \theta)\, d\, d\theta \\ &= [P(t, 0, \alpha)\phi^0 + \int_{-h}^{0} P(t, \theta, \alpha)\phi^1(\theta)\, d\theta] \cdot d, \end{aligned}$$

where we have used Lemma 6.4 and the symmetry relationship (6.14). This proves (6.24). The continuity of the functions (6.25) to (6.27) is a consequence of the continuity of the function (6.21), and for (6.25), the continuity of the injection of $H^1(-h, 0; \mathbb{R}^n)$ into $C_0(-h, 0; \mathbb{R}^n)$.

The properties of the functions (6.29) and (6.30) are obtained from those of the functions (6.22) and (6.23).

(vi) The properties of the functions (6.31) to (6.33) and (6.35) are obtained from the properties of the functions (6.25) to (6.27) and (6.30) with $\phi^1 = 0$ and the symmetry property (6.14). To prove (6.36), we use (6.24) with $(\phi^0, \phi^1) = (0, \psi)$, the property of the function (6.27) and identity (6.28):

$$(7) \qquad (\phi, DP_J(t)J^*j^*(0, \psi))_2 = \int_{-h}^{0} \phi(\alpha) \cdot \frac{\partial}{\partial \alpha} \int_{-h}^{0} P(t, \theta, \alpha)\psi(\theta)\, d\theta\, d\alpha.$$

But from identity (6.13) in Lemma 6.4

$$J^*j^*(0, \psi) = \int_{-h}^{0} \delta_\theta\psi(\theta)\, d\theta$$

and by continuity of D and $P_J(t)$

$$DP_J(t)J^*j^*(0, \psi) = \int_{-h}^{0} DP_J(t)\delta_\theta\psi(\theta)\, d\theta.$$

Moreover, for ϕ in $L^2(-h, 0; \mathbb{R}^n)$

$$\begin{aligned} (\phi, DP_J(t)J^*j^*(0, \psi))_2 &= \int_{-h}^{0} (\phi, DP_J(t)\delta_\theta\psi(\theta))_2\, d\theta \\ &= \int_{-h}^{0} \int_{-h}^{0} \phi(\alpha) \cdot \frac{\partial P}{\partial \alpha}(t, \theta, \alpha)\psi(\theta)\, d\alpha\, d\theta \end{aligned}$$

and using (6.14)

$$(8)\qquad (\phi, DP_J(t)J^*j^*(0,\psi))_2 = \int_{-h}^{0}\left[\int_{-h}^{0}\frac{\partial P}{\partial\alpha}(t,\alpha,\theta)\phi(\alpha)\,d\alpha\right]\cdot\psi(\theta)\,d\theta.$$

Comparing (7) and (8) we obtain (6.36).

The proof of identity (6.37) amounts to justifying a change in the order of integration on the right-hand side of (6.36). In view of inequality (6.19) it is easy to check that the integrand

$$(\alpha,\theta)\to\frac{\partial P}{\partial\alpha}(t,\alpha,\theta)\phi(\alpha)\cdot\psi(\theta)$$

is an L^1-function when $\phi\in L^2(-h,0;\mathbb{R}^n)$ and $\psi\in L^2(-h,0;\mathbb{R}^n)\cap L^1(-h,0;\mathbb{R}^n)$. This proves (6.37). When h is finite $L^2(-h,0;\mathbb{R}^n)\subset L^1(-h,0;\mathbb{R}^n)$ and this justifies the remark following identity (6.37).

(vii) Recall that for all $\phi=(\phi^0,\phi^1)$ in M^2

$$\Pi(t)\phi = jP(t)j^*\phi = jJP_J(t)J^*j^*\phi = ([P_J(t)J^*j^*\phi](0), P_J(t)J^*j^*\phi).$$

Now from identity (6.24)

$$\Pi^{00}(t)\phi^0+\Pi^{01}(t)\phi^1 = [P_J(t)J^*j^*\phi](0) = P(t,0,0)\phi^0+\int_{-h}^{0}P(t,\theta,0)\phi^1(\theta)\,d\theta,$$

$$[\Pi^{10}(t)\phi^0+\Pi^{11}(t)\phi^1](\alpha) = [P_J(t)J^*j^*\phi](\alpha) = P(t,0,\alpha)\phi^0+\int_{-h}^{0}P(t,\theta,\alpha)\phi^1(\theta)\,d\theta.$$

Since ϕ^0 and ϕ^1 are independent, this is sufficient to obtain (6.40) to (6.42). The last part of identities (6.41) and (6.42) is again a direct consequence of (6.14). □

Proof of Theorem 6.6. Start with the Riccati equation (5.45) of Theorem 5.5. Take the inner product (5.45) with k in M^2. The first term is the derivative of

$$①=((jP(t)j^*h,k))=\left[P(t,0,0)h^0+\int_{-h}^{0}P(t,\alpha,0)h^1(\alpha)\,d\alpha\right]\cdot k^0$$
$$+\int_{-h}^{0}\left[P(t,0,\theta)h^0+\int_{-h}^{0}P(t,\alpha,\theta)h^1(\alpha)\,d\alpha\right]\cdot k^1(\theta)\,d\theta.$$

The second term is

$$②=((A^TP(t)j^*h,k))=((A^TJP_J(t)J^*j^*h,k))$$
$$=L^TP_J(t)J^*j^*h,k^0+(D_\theta P_J(t)J^*j^*h,k^1)$$
$$=\int_{-h}^{0}d_\theta\eta^T\left[P(t,0,\theta)h^0+\int_{-h}^{0}P(t,\alpha,\theta)h^1(\alpha)\,d\alpha\right]\cdot k^0$$
$$+\int_{-h}^{0}\frac{\partial}{\partial\theta}\left[P(t,0,\theta)h^0+\int_{-h}^{0}P(t,\alpha,\theta)h^1(\alpha)\,d\alpha\right]\cdot k^1(\theta)\,d\theta.$$

The third term is

$$③ = ((jP(t)(A^T)^*h, k)) = ((h, A^TP(t)j^*k))$$
$$= h^0 \cdot \int_{-h}^{0} d_\alpha \eta^T \left[P(t, 0, \alpha)k^0 + \int_{-h}^{0} P(t, \theta, \alpha)k^1(\theta)\, d\theta \right]$$
$$+ \int_{-h}^{0} h^1(\alpha) \cdot \frac{\partial}{\partial \alpha} \left[P(t, 0, \alpha)k^0 + \int_{-h}^{0} P(t, \theta, \alpha)k^1(\theta)\, d\theta \right] d\alpha.$$

In view of (6.36)

$$\int_{-h}^{0} h^1(\alpha) \cdot \frac{\partial}{\partial \alpha} \int_{-h}^{0} P(t, \theta, \alpha)k^1(\theta)\, d\theta\, d\alpha = \int_{-h}^{0} \left[\int_{-h}^{0} \frac{\partial P}{\partial \alpha}(t, \alpha, \theta)h^1(\alpha)\, d\alpha \right] \cdot k^1(\theta)\, d\theta.$$

Using property (6.16) we can change the order of integration in the term

$$h^0 \cdot \int_{-h}^{0} d_\alpha \eta^T \int_{-h}^{0} P(t, \theta, \alpha)k^1(\theta)\, d\theta = \int_{-h}^{0} d\theta \left[\int_{-h}^{0} P(t, \alpha, \theta) d_\alpha \eta h^0 \right] \cdot k^1(\theta).$$

The final result is

$$③ = \int_{-h}^{0} P(t, \alpha, 0)\, d_\alpha \eta h^0, k^0 + \int_{-h}^{0} d\theta \left[\int_{-h}^{0} P(t, \alpha, \theta)\, d_\alpha \eta h^0 \right] \cdot k^1(\theta)$$
$$+ \int_{-h}^{0} \frac{\partial P}{\partial \alpha}(t, \alpha, 0)h^1(\alpha)\, d\alpha \cdot k^0$$
$$+ \int_{-h}^{0} d\theta \left[\int_{-h}^{0} d\alpha \frac{\partial P}{\partial \alpha}(t, \alpha, \theta)h^1(\alpha) \right] \cdot k^1(\theta).$$

The fourth term is

$$④ = ((jP(t)RP(t)j^*h, k)) = N^{-1}B^TJ^{-1}P(t)j^*h \cdot B^TJ^{-1}P(t)j^*k.$$

But

$$B^TJ^{-1}P(t)j^*k = B^TJ^{-1}JP_J(t)J^*j^*k = B^TP_J(t)J^*j^*k$$
$$= \int_{-h}^{0} d_\zeta \beta^T \left[P(t, 0, \zeta)k^0 + \int_{-h}^{0} P(t, \alpha, \zeta)k^1(\alpha)\, d\alpha \right]$$

and in view of (6.16) the order of integration can be changed in the last term. Finally,

$$④ = N^{-1} \left\{ \int_{-h}^{0} d_\xi \beta^T P(t, 0, \xi)h^0 + \int_{-h}^{0} d\alpha \int_{-h}^{0} d_\xi \beta^T P(t, \alpha, \xi)h^1(\alpha) \right\}$$
$$\cdot \left\{ \int_{-h}^{0} d_\zeta \beta^T P(t, 0, \zeta)k^0 + \int_{-h}^{0} d\theta \int_{-h}^{0} d_\zeta \beta^T P(t, \theta, \zeta)k^1(\theta) \right\}.$$

The last and simplest term is

$$⑤ = ((\tilde{Q}h, k)) = Qh^0 \cdot k^0.$$

Equations (6.43), (6.44) and (6.45) are now obtained by special choice of the pair (h, k) in the equation

$$\frac{d}{dt}① + ② + ③ + ④ + ⑤ = 0. \tag{9}$$

For $h=(h^0,0)$ and $k=(k^0,0)$ we obtain (6.41). For $h=(0,h^1)$ and $k=(k^0,0)$ equation (9) reduces to

$$\begin{aligned}(10)\qquad &\frac{d}{dt}\int_{-h}^{0} P(t,\alpha,0)h^1(\alpha)\,d\alpha+\int_{-h}^{0} d_\theta\eta^T\int_{-h}^{0} d\alpha\,P(t,\alpha,\theta)h^1(\alpha)\\ &+\int_{-h}^{0}\frac{\partial P}{\partial\alpha}(t,\alpha,0)h^1(\alpha)\,d\alpha\\ &-\int_{-h}^{0} P(t,\zeta,0)\,d_\zeta\beta N^{-1}\int_{-h}^{0} d\alpha\int_{-h}^{0} d_\xi\beta^T P(t,\alpha,\xi)h^1(\alpha)=0.\end{aligned}$$

From Theorem 6.5(vi) the map $\alpha\to\partial P/\partial t(t,\alpha,0)$ belongs to L^2 and

$$\frac{d}{dt}\int_{-h}^{0} P(t,\alpha,0)h^1(\alpha)\,d\alpha=\int_{-h}^{0}\frac{\partial P}{\partial t}(t,\alpha,0)h^1(\alpha)\,d\alpha.$$

In view of (6.16), the order of integration can be changed in the second term of (10) and the map

$$\alpha\to\int_{-h}^{0} d_\theta\eta^T P(t,\alpha,\theta)$$

belongs to L^2. The same remark applies to the last part of the last term in (10). As a result we can regroup terms in (8) to obtain

$$\begin{aligned}\int_{-h}^{0}\Big\{&\frac{\partial P}{\partial t}(t,\alpha,0)+\int_{-h}^{0} d_\theta\eta^T P(t,\alpha,\theta)+\frac{\partial P}{\partial\alpha}(t,\alpha,0)\\ &-\int_{-h}^{0} P(t,\zeta,0)d_\zeta\beta N^{-1}\int_{-h}^{0} d_\xi\beta^T P(t,\alpha,\xi)\Big\}h^1(\alpha)\,d\alpha=0.\end{aligned}$$

As a function of α, all terms in the curly bracket belong to L^2 and are continuous with respect to t in $[0,T]$ with values in $L^2(-h,0;\mathscr{L}(\mathbb{R}^n,\mathbb{R}^n))$. Therefore the curly bracket is null almost everywhere. This establishes (6.44).

To obtain (6.45), we set $h=(0,\phi)$ and $k=(0,\psi)$ in (9)

$$\begin{aligned}&\frac{d}{dt}\int_{-h}^{0}\int_{-h}^{0} P(t,\alpha,\theta)\phi(\alpha)\,d\alpha\cdot\psi(\theta)\,d\theta+\int_{-h}^{0}\frac{\partial}{\partial\theta}\Big[\int_{-h}^{0} P(t,\alpha,\theta)\phi(\alpha)\,d\alpha\Big]\cdot\psi(\theta)\,d\theta\\ &+\int_{-h}^{0}\Big[\int_{-h}^{0}\frac{\partial P}{\partial\alpha}(t,\alpha,\theta)\phi(\alpha)\,d\alpha\Big]\cdot\psi(\theta)\,d\theta\\ &-N^{-1}\int_{-h}^{0}\Big[\int_{-h}^{0} d_\xi\beta^T P(t,\alpha,\xi)\Big]\phi(\alpha)\,d\alpha\int_{-h}^{0}\Big[\int_{-h}^{0} d_\zeta\beta^T P(t,\theta,\zeta)\Big]\psi(\theta)\,d\theta=0.\end{aligned}$$

But in view of (6.30) with $(\phi^0,\phi^1)=(0,\phi)$

$$\frac{d}{dt}\int_{-h}^{0}\int_{-h}^{0} P(t,\alpha,\theta)\phi(\alpha)\,d\alpha\cdot\psi(\theta)\,d\theta=\int_{-h}^{0}\Big[\frac{\partial}{\partial t}\int_{-h}^{0} P(t,\alpha,\theta)\phi(\alpha)\,d\alpha\Big]\cdot\psi(\theta)\,d\theta.$$

Moreover for all ϕ in $L^2(-h,0;\mathbb{R}^n)\cap L^1(-h,0;\mathbb{R}^n)$ and ψ in $L^2(-h,0;\mathbb{R}^n)$ (cf. (6.37))

$$\int_{-h}^{0}\frac{\partial}{\partial\theta}\Big[\int_{-h}^{0} P(t,\alpha,\theta)\phi(\alpha)\,d\alpha\Big]\cdot\psi(\theta)\,d\theta=\int_{-h}^{0}\Big[\int_{-h}^{0}\frac{\partial P}{\partial\theta}(t,\alpha,\theta)\phi(\alpha)\,d\alpha\Big]\cdot\psi(\theta)\,d\theta.$$

By factoring ϕ and ψ

$$(11)\qquad \int_{-h}^{0}\left\{\frac{\partial}{\partial t}\int_{-h}^{0} P(t,\alpha,\theta)\phi(\alpha)\,d\alpha+\int_{-h}^{0} Z(t,\alpha,\theta)\phi(\alpha)\,d\alpha\right\}\cdot\psi(\theta)\,d\theta$$

where

$$(12)\quad Z(t,\alpha,\theta)=\left(\frac{\partial}{\partial\theta}+\frac{\partial}{\partial\alpha}\right)P(t,\alpha,\theta)-\int_{-h}^{0} P(t,\zeta,\theta)\,d_{\zeta}\beta N^{-1}\int_{-h}^{0} d_{\xi}\beta^{T}P(t,\alpha,\xi).$$

Since the above identity is true for all ϕ in $L^2(-h,0;\mathbb{R}^n)\cap L^1(-h,0;\mathbb{R}^n)$ and ψ in $L^2(-h,0;\mathbb{R}^n)$, the time distributional derivative of $P(t,\alpha,\theta)$ is given by

$$\frac{\partial P}{\partial t}(t,\alpha,\theta)+Z(t,\alpha,\theta)=0\quad\text{a.e.}$$

When h is finite the matrix function Z is L^2 with respect to its arguments hence $\partial P/\partial t$ also belongs to L^2. □

Appendix to Section 7.

Proof of Theorem 7.5. The proof of parts (i) and (ii) follows standard arguments given for instance in R. Datko [1]. Part (iii) also uses standard arguments.

(iv) Start with the Riccati differential equation (5.45) of Theorem 5.5: for all k in M^2

$$(1)\qquad -\frac{d}{dt}(jP_T(t)j^*k)=A^TP_T(t)j^*k+jP_T(t)(A^T)^*k-jP_T(t)RP_T(t)j^*k+\tilde{Q}k.$$

Denote by $Z_T(t)$ the right-hand side of (1). For all $t>0$ the operator $Z_T(t)$ strongly converges to some operator

$$(2)\qquad Z=A^TPj^*+jP(A^T)^*-jPRPj^*+\tilde{Q}$$

in $\mathcal{L}(M^2,M^2)$ as T goes to $+\infty$ and

$$\forall k\in V,\qquad \lim_{t\leqq T\to\infty}\frac{d}{dt}(jP_T(t)j^*k)=-Zk\quad\text{in } M^2.$$

For each arbitrary fixed $\tau>0$, define the sequence

$$f_n(t)=jP_n(t)j^*k,\quad 0\leqq t\leqq\tau,\quad n\geqq 1.$$

The sequence $\{f_n\}$ belongs to $H^1(0,\tau;M^2)$,

$$\lim_{n\to\infty} f_n=f,\quad f(t)=jPj^*k,\qquad 0\leqq t\leqq\tau,$$

$$\lim_{n\to\infty}\frac{d}{dt}f_n(t)=g(t)=-Zk,\qquad 0\leqq t\leqq\tau.$$

But $(d/dt)f_n$ and g belong to $L^2(0,\tau;M^2)$ and there exists $c>0$ such that for all $n\geqq 1$ and t in $[0,\tau]$

$$\left|\frac{d}{dt}f_n(t)\right|\leqq c\|k\|_{M^2},\qquad |g(t)|\leqq c\|k\|_{M^2}.$$

By the Lebesgue Dominated Convergence Theorem

$$\frac{df_n}{dt}\to g\quad\text{in } L^2(0,\tau;M^2).$$

Finally we have shown that $\{f_n\}$ converges in the H^1-norm, and that

$$f_n \to f, \qquad \frac{df_n}{dt} \to g.$$

By completeness of $H^1(0, \tau; M^2)$

$$\frac{df}{dt} = g \Rightarrow 0 = \frac{d}{dt}(jPj^*k) = -Zk.$$

Since the above is true for all k in M^2, we conclude that $Z = 0$. Equation (7.28) is finally obtained from (2) with $Z = 0$.

(v) Let $\bar{P}$ in $\mathscr{L}(V', V)$ be another positive symmetrical solution of (7.28). Rearrange (7.28) for $\bar{P}$:

$$[A^T - j\bar{P}R]\bar{P}j^* + j\bar{P}[A^T - j\bar{P}R]^* + j\bar{P}R\bar{P}j^* + \tilde{Q} = 0. \tag{3}$$

Let z in $\mathscr{W}_{\text{loc}}(0, \infty; M^2, V')$ be the solution of

$$\frac{d}{dt}j^*z(t) - (A^T)^*z(t) + R\bar{P}j^*z(t) = 0, \qquad z(0) = k \in M^2 \tag{4}$$

(by Theorem 4.6(ii) and Theorem 2.7(ii) by choosing $u = -B^TJ^{-1}Pj^*z$). From (3)

$$((z(t), \{[A^T - j\bar{P}R]\bar{P}j^* + j\bar{P}[A^T - j\bar{P}R]^* + j\bar{P}R\bar{P}j^* + \tilde{Q}\}z(t))) = 0 \tag{5}$$

and if w is defined as

$$w(t) = -N^{-1}B^TJ^{-1}\bar{P}z(t), \qquad t \geqq 0, \tag{6}$$

we obtain from (5)

$$2\langle [A^T - j\bar{P}R]^*z(t), \bar{P}j^*z(t)\rangle_V + Nw(t) \cdot w(t) + ((z(t), \tilde{Q}z(t))) = 0.$$

From (4) the above equation reduces to

$$\frac{d}{dt}\{\langle j^*z(t), \bar{P}j^*z(t)\rangle_V + Nw(t) \cdot w(t) + \langle i^*z(t), \hat{Q}i^*z(t)\rangle_W\} = 0$$

and for all $t > 0$

$$\langle j^*k, \bar{P}j^*k\rangle_V = \langle j^*z(t), \bar{P}j^*z(t)\rangle_V + \int_0^t [|Ci^*z(s)|^2 + Nw(s) \cdot w(s)]\, ds \geqq J_0^t(w, j^*k)$$

$$\geqq J_0^t(u_t, j^*k) = \langle j^*k, P_t(0)j^*k\rangle_V,$$

where u_t is the minimizing control on $[0, t]$. By letting t go to $+\infty$, we obtain

$$\langle j^*k, \bar{P}j^*k\rangle \geqq \langle j^*k, Pj^*k\rangle. \tag{7}$$

By density of j^*M^2 in V' and continuity of the symmetrical operator $\bar{P} - P$, inequality (7) extends to all of V'. □

REFERENCES

R. A. ADAMS [1], *Sobolev Spaces*, Academic Press, New York, 1975.

R. DATKO [1], *A linear control problem in an abstract Hilbert space*, J. Differential Equations, 9 (1971), pp. 346-359.

J. G. BORISOVIC AND A. S. TURBABIN [1], *On the Cauchy problem for linear non-homogeneous differential equation with retarded arguments*, Soviet Math. Dokl, 10 (1969), pp. 401-405.

M. C. DELFOUR [1], *Status of the state space theory of linear hereditary differential systems with delays in state and control variables*, in Analysis and Optimization of Systems, A. Bensoussan and J. L. Lions, eds., Springer-Verlag, Berlin, Heidelberg, New York, 1980, pp. 83-96.

——— [2], *The linear quadratic optimal control theory for systems with delays in state and control variables*, in Control Science and Technology for the Progress of Society, H. Akashi, ed., Vol. I, Pergamon Press, Oxford, 1981, pp. 361-366.

——— [3], *The largest class of hereditary systems defining a C_0 semigroup on the product space*, Canad. J. of Math., 32 (1980), pp. 969-978.

——— [4], *The linear quadratic optimal control problem for hereditary differential systems; theory and numerical solution*, J. Appl. Math. Optim., 3 (1977), pp. 101-162.

——— [5] *Linear optimal control of systems with state and control variable delays*, Automatica, 20 (1984), pp. 69-77.

M. C. DELFOUR, E. B. LEE AND A. MANITIUS [1], *F-reduction of the operator Riccati equation for hereditary differential systems*, Automatica, 14 (1978), pp. 385-395.

M. C. DELFOUR AND A. MANITIUS [1], *The structural operator F and its role in the theory of retarded systems* I, J. Math. Anal. Appl., 73 (1980), pp. 466-490.

M. C. DELFOUR AND M. SORINE [1], *The linear-quadratic optimal control problem for parabolic systems with boundary control through a Dirichlet condition*, in Proc. 3rd IFAC Symposium on Control of Distributed Parameter Systems, J. P. Babary and L. Le Letty, eds., IFAC Publications, Toulouse, France, 1982, pp. I.13-I.16.

N. DUNFORD AND J. T. SCHWARTZ [1], *Linear Operators, Part* I: *General Theory*, Interscience, New York, 1967.

A. ICHIKAWA [1], *Generation of a semigroup on some product space with applications to evolution equations with delays*, Control Theory Centre Report no. 52, Univ. Warwick, Coventry, 1977.

——— [2], *Optimal quadratic control and filtering for evolution equations with delay in control and observation*, Control Theory Centre Report no. 53, Univ. Warwick, Coventry, 1977.

——— [3], *Quadratic control of evolution equations with delay in control*, this Journal, 20 (1982), pp. 645-668.

L. V. KANTOROVICH AND G. P. AKILOV [1], *Functional Analysis in Normed Spaces*, Pergamon Press, New York, 1964.

H. KOIVO AND E. B. LEE [1], *Controller synthesis for linear systems with retarded state and control variables and quadratic cost*, Automatica, 8 (1972), pp. 203-208.

R. H. KWONG [1], *A linear-quadratic Gaussian theory for systems with delays in the state, control and observations*, Report 7714, Systems Control Group, Univ. Toronto, Canada, September 1977.

——— [2], *The linear quadratic Gaussian problem for systems with delays in the state, control, and observations*, Proc. 14th Allerton Conference on Circuit and System Theory, Univ. Illinois, Urbana, 1976, pp. 545-549.

——— [3], *A stability theory of the linear-quadratic-Gaussian problem for systems with delays in the state, control, and observations*, this Journal, 18 (1980), pp. 49-75.

——— [4], *Characterization of kernel functions associated with operator algebraic Riccati equations for linear delay systems*, Systems Control Report no. 7906, Univ. Toronto, Toronto, Canada, June 1979.

R. KWONG AND A. D. WILLSKY [1], *Optimal filtering and filter stability of linear stochastic delay systems*, IEEE Trans. Automat. Control, AC-22 (1977), pp. 196-201.

——— [2], *Estimation and filter stability of stochastic delay systems*, this Journal, 16 (1978), pp. 660-681.

J. L. LIONS [1], *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, Heidelberg, Berlin, 1971.

A. W. OLBROT [1], *Stabilizability, detectability, and spectrum assignment for linear autonomous systems with general time delays*, IEEE Trans. Automat. Control, AC-23 (1978), pp. 887-890.

L. PANDOLFI [1], *Canonical realization of systems with delayed controls*, Ricerche di Automatica, 10 (1979), pp. 27-37.

——— [2], *On feedback stabilization of functional differential equations*, Bolletino UMI 4, 11, Supplemento al fasciculo 3, Giugno 1975, Serie IV, vol. XI, pp. 626-635.

F. RIESZ AND B. SZ. NAGY [1], *Functional Analysis*, seventh printing, Frederick Ungar, New York, 1978.

W. RUDIN [1], *Real and Complex-Analysis*, McGraw-Hill, New York, 1966.

R. B. VINTER [1], *Stabilizability and semigroups with discrete generators*, J. Inst. Math. Appl., 20 (1977), pp. 371-378.

M. SORINE [1], *Un résultat d'existence et d'unicité pour l'équation de Riccati stationnaire*, Report CRMA-984, Université de Montréal, September 1980.

——— [2], *Sur le semigroupe non-linéaire associé à l'équation de Riccati*, Report CRMA-1055, Université de Montréal, Montreal, Canada, September 1981.

R. B. VINTER AND R. H. KWONG [1], *The infinite time quadratic control problem for linear systems with state and control delays: an evolution equation approach*, this Journal, 19 (1981), pp. 139-153.

K. YOSIDA [1], *Functional Analysis*, Springer-Verlag, New York, 1966.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 5, September 1986

RICCATI EQUATIONS FOR HYPERBOLIC PARTIAL DIFFERENTIAL EQUATIONS WITH $L_2(0, T; L_2(\Gamma))$—DIRICHLET BOUNDARY TERMS*

I. LASIECKA† AND R. TRIGGIANI†

Abstract. This paper studies the quadratic optimal control problem for second order (linear) hyperbolic partial differential equations defined on a bounded domain $\Omega \subset R^n$ with boundary Γ. Both the finite interval case $[0, T]$, $T < \infty$, and the infinite interval case $T = \infty$ (regulator problem) are considered. The distinguishing feature of the paper, which differentiates it from previous (scarce!) literature on the subject, is that the controls are only $L_2(0, T; L_2(\Gamma))$-functions which act in the Dirichlet B.C. and that the corresponding solutions are penalized in the $L_2(0, T; L_2(\Omega))$-norm (smoother controls, particularly in space, were taken in the few previous works on this subject). The well-posedness of this formulation stems from recent results by the authors about regularity of second order hyperbolic mixed problems [L-T.1], [L-T.3]. Under minimal assumptions, the optimal control is synthesized, in a pointwise feedback form, through an operator which is shown to satisfy in a suitable sense a Riccati differential equation for $T < \infty$ and a Riccati algebraic equation for $T = \infty$. Unlike most, if not all, of the literature on quadratic control problems (for different dynamics!), the algebraic Riccati equation is *not* derived as a limit process as $T \uparrow \infty$ of the Riccati differential (or integral) equation on $[0, T]$. This has a special advantage in the case of hyperbolic dynamics. Rather, the approach followed for the control problems is direct, in the sense that first an operator is defined in terms of the hyperbolic dynamics and only subsequently shown to satisfy a Riccati equation (differential for $T < \infty$, algebraic for $T = \infty$). Regularity results of the optimal pair are also included. A functional analytic model, based on cosine operator theory and introduced by the authors in [L-T.1], [L-T.3], is used throughout to describe the hyperbolic dynamics.

Key words. Riccati equations, boundary control, hyperbolic partial differential equations

AMS(MOS) subject classification. 35

1. Introduction, problem formulation and statement of main results.

1.1. Introduction and problem formulation. The aim of the present paper is a study of the quadratic optimal control problem on both a finite, fixed interval $[0, T]$, $T < \infty$, (§§ 2–4) and also on the infinite interval $T = \infty$ (§ 5) for second order (linear) hyperbolic partial differential equations, where L_2-boundary controls act in the Dirichlet Boundary Conditions (B.C.). In this introductory section, we shall first formulate the problems and point out their relationship to (scarce!) existing literature on the subject, and then present a general orientative statement of our main results. Complete formal statements and proofs are given in §§ 2–5.

Let Ω be an open bounded domain in R^n with boundary Γ, say (piecewise) of class C^2. Let $-A(\xi, \partial)$ be a partial differential operator of order two in Ω with smooth real coefficients

$$-A(\xi, \partial) = \sum_{i,j=1}^{n} \frac{\partial}{\partial \xi_i}\left(a_{ij}(\xi)\frac{\partial}{\partial \xi_j}\right) + \sum_{j=1}^{n} b_j(\xi)\frac{\partial}{\partial \xi_j} + c_0(\xi)$$

with principal part uniformly strongly elliptic in Ω

$$\sum_{i,j=1}^{n} a_{ij}(\xi)\eta_i\eta_j \geqq \alpha \sum_{j=1}^{n} \eta_j^2, \qquad a_{ij} \equiv a_{ji}, \quad \alpha > 0.$$

* Received by the editors September 6, 1983, and in revised form April 16, 1985. This research was partially supported by the National Science Foundation under DMS-8301668.

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

We consider the mixed hyperbolic problem

$$\begin{aligned} &\frac{\partial^2 y}{\partial t^2}(t,\xi) = -A(\xi,\partial)y(t,\xi) && \text{in } (0,T]\times\Omega \equiv Q, \\ &y(0,\xi) = y_0(\xi), \quad \frac{\partial y}{\partial t}(0,\xi) = y_1(\xi) && \xi\in\Omega, \\ &y(t,\sigma) = u(t,\sigma) && \text{in } (0,T]\times\Gamma \equiv \Sigma, \end{aligned} \tag{1.1}$$

where the control function $u(t,\sigma)$ acts in the Dirichlet B.C. and is assumed to belong to $L_2(0,T;L_2(\Gamma)) \equiv L_2(\Sigma)$. Let $L_2(0,T;L_2(\Omega)) \equiv L_2(Q)$. The *homogeneous problem* (i.e. (1.1) with $u\equiv 0$) is *uniformly well-posed in* $L_2(\Omega)$; equivalently [F1], the operator $-A$ consisting of $-A(\xi,\partial)$ plus homogeneous Dirichlet B.C. *is the generator of a strongly continuous* (s.c.) cosine operator $C(t)$ on $L_2(\Omega)$, $t\in R$. Thus, $Ah = A(\xi,\partial)h$, $h\in\mathcal{D}(A) = \{x\in L_2(\Omega)\colon A(\xi,\partial)h\in L_2(\Omega),\ h|_\Gamma = 0\}$. The self-adjoint principal part of $-A(\xi,\partial)$ generates a s.c. (self-adjoint) cosine operator on $L_2(\Omega)$, while the first order part of $-A(\xi,\partial)$ is a perturbation which preserves generation of a s.c. cosine operator [F.3, Thm. 2.1], [T-W.1, Prop. 4.1]. Similarly, $-A^*$ generates a s.c. cosine operator on $L_2(\Omega)$, given precisely by $C^*(t)$; see [N.2] for the general case. For the necessary background on cosine operators, we refer e.g. to [F.1, I, II], [F.3], [N.2], [S.1], [T-W.1] plus bibliography cited therein. Of this theory we shall recall below only a few facts that will be frequently needed in the sequel. The optimal control problem of penalizing on $[0,T]$ both the boundary input $u\in L_2(\Sigma)$ and the corresponding solution y in a suitable norm is, of course, intimately dependent on the problem of *regularity* of the mixed problem (1.1). In particular, does $u\in L_2(\Sigma)$ imply that the corresponding solution $y\in L_2(Q)$? Only very recently was this seemingly natural question given an affirmative answer [L-T.1]. (For instance, the regularity results of the fundamental treatise [L-M.1] do not answer the above question.) Even more recently, this $L_2(Q)$-regularity result was markedly improved to conclude that, in fact, for $u\in L_2(\Sigma)$ the corresponding solution $y\in C([0,T];L_2(\Omega))$ [L-T.3].[1] See also the rather comprehensive study on regularity of hyperbolic mixed problems in [L-L-T.1].

THEOREM 1.1 (Regularity for (1.1)). *Let* Ω *either have* C^1*-boundary* Γ *or else be a parallelepiped. Let* $y_0\in L_2(\Omega)$ *and* $y_1\in H^{-1}(\Omega)$. *Then*:

(i) *The map from the input* $u\to$ *corresponding solution* $[y,\dot y]$ *is a continuous operator*: $L_2(\Sigma)\to L_2(Q)\otimes L_2(0,T;H^{-1}(\Omega))$ [L-T.1].

(ii) *In fact, the map from the input* $u\to$ *corresponding solution* $[y,\dot y]$ *of* (1.1) *is a continuous operator* $L_2(\Sigma)\to C([0,T];L_2(\Omega))\otimes C([0,T];H^{-1}(\Omega))$ [L-T.3].

On the basis of the (weaker) regularity result (i), one can then associate with (1.1) the quadratic cost functional[2] on $[0,T]$,

$$J(u,y) \equiv \int_0^T \{(Ry(t),y(t))_\Omega + |u(t)|_\Gamma^2\}\,dt = |R^{1/2}y|_Q^2 + |u|_\Sigma^2 \tag{1.2a}$$

with R satisfying:

(1.3) (H.1) R is a nonnegative self-adjoint, bounded operator on $L_2(\Omega)$

so that R may be the identity on $L_2(\Omega)$.

[1] In January, 1983 Professor J. L. Lions has kindly informed us that he also had proved this result in 1982 with a proof "100% different" from ours. Paper [L-L-T.1] is the result of correspondence that followed.

[2] The norms are all L_2-norms over the indicated domains, unless otherwise stated. Also, our treatment can be directly extended as to include quadratic penalization of the velocity $\dot y(t)$ in $H^{-1}(\Omega)$.

The *optimal control problem* (O.C.P.) on the finite interval $[0, T]$ is now:

(O.C.P.) Minimize $J(u, y(u))$ over all $u \in L_2(\Sigma)$, where $y(u)$ is the solution to (1.1) corresponding to u.

In light of Theorem 1.1(i), the functional $J(u, y(u))$ is continuous on $L_2(\Sigma)$; since it is strictly convex, it follows by standard arguments in optimization theory, that the (O.C.P.) admits a unique solution, which we shall denote by u^0. The corresponding optimal solution is then denoted by y^0. Actually, in view of the stronger regularity statement (ii) in Theorem 1.1, there is no essential *extra difficulty* in penalizing also the final state $|y(T)|_\Omega$; i.e. in replacing J in (1.2*a*) with

$$J(u, y) = |y(T)|^2_\Omega + |y|^2_Q + |u|^2_\Sigma, \tag{1.2b}$$

since the operator $u \in L_2(\Sigma) \to y(T) \in L_2(\Omega)$ is bounded.[3] Thus:

THEOREM 1.2. *Under the conclusion of Theorem* 1.1, *the* (O.C.P.) *corresponding to* (1.2a) *or* (1.2b) *admits a unique solution*: $u^0 \in L_2(\Sigma)$ *and* $y^0 \in C([0, T]; L_2(\Omega))$.

In § 5, we study also the quadratic optimal control problem (1.2a) with $T = \infty$ (regulator problem), under the minimal assumption—verified in [L-T.5]—that for each initial data there is a pair of u and y such that the corresponding cost with $T = \infty$ be finite. Before turning to the statement and proof of our results, we review (the scarce) existing literature on the quadratic optimal problem for hyperbolic partial differential equations, under Dirichlet boundary control. As already explained, a basic preliminary difficulty encountered in the study of the (O.C.P.) is a question of regularity of the solutions to the mixed problem (1.1); in particular, whether the cost J in (1.2a) is well-set. Thus, the crux of the case that we study here is that we penalize both the Dirichlet boundary control and the corresponding solution in the L_2-norms; i.e., in $L_2(0, T; L_2(\Gamma))$ and $L_2(0, T; L_2(\Omega))$, respectively. This is the *distinguishing feature*, which differentiates our present results from those already existing in the literature, e.g., [C-P.1], [C-P.2], [L.1][4]; where, in face of this, smoother boundary controls were considered, e.g. $u \in H^2_0([0, T] \times \Gamma)$ as in [L.1, p. 325], or $u \in L_2(0, T; H^{1/2}(\Gamma))$ as in [C-P.1] and [C-P.2]. Moreover, Riccati's synthesis is not investigated in [L.1] in the hyperbolic boundary case, only in the distributed case, see [L.1, p. 348]. We also stress that our approach here to the Riccati's synthesis is "explicit" and "constructive" (in the style of [L-T.2] in the parabolic case) in the sense that an operator is first defined by an explicit formula in terms of the given dynamics (see (2.22a) or (2.24) below) and only subsequently proved to be a solution of a Riccati differential and integral equation. This way, the usual difficulty—encountered in all "indirect" approaches to Riccati synthesis of much of the literature (for different dynamics!)—of proving existence of a solution to the operator Riccati differential equation is automatically taken care of. As to the infinite interval case, $T = \infty$, we are not aware of any derivation of the algebraic Riccati equation for the hyperbolic dynamics (1.1) in the literature.

1.2. Statement of main results. While we refer to §§ 2–5 for the complete, technical statements of our results, we wish to give here a preliminary general description of them, which will help orient the reader. To do this, some preliminary background

[3] This fact is surprising in light of known results for parabolic equations, where the solution y to an $L_2(\Sigma)$-control acting in the Dirichlet B.C. may not have a well defined point $y(T)$ [L.1, p. 202]. This pathology adds extra difficulty in the corresponding quadratic cost problem with final state penalization in $L_2(\Omega)$ see [L-T.2].

[4] See also [V-J.1].

material is needed. It is well known that the operator

$$\begin{vmatrix} 0 & I \\ -A & 0 \end{vmatrix}, \quad \text{with domain} = \mathcal{D}(A) \otimes \mathcal{D}(A^{1/2}) \tag{1.4}$$

generates a strongly continuous (s.c.) group on $H_0^1(\Omega) \otimes L_2(\Omega)$, which is, in fact, unitary if A is self-adjoint. (We may assume without loss of generality that the fractional powers of A are well-defined.) With $H_0^1(\Omega) \equiv \mathcal{D}(A^{1/2})$ [F.2], [L.2] it follows quickly from Theorem 1.1 that the operator $\mathcal{A}$,

$$\mathcal{A} = \begin{vmatrix} 0 & I \\ -A & 0 \end{vmatrix}, \quad \text{with domain } \mathcal{D}(\mathcal{A}) = \mathcal{D}(A^{1/2}) \otimes L_2(\Omega) \tag{1.5}$$

generates a s.c. group, which will be denoted by $e^{\mathcal{A}t}$, also on the space

$$E \equiv L_2(\Omega) \times [\mathcal{D}(A^{1/2})]' \equiv L_2(\Omega) \times H^{-1}(\Omega), \tag{1.6}$$

$$\left\| \begin{vmatrix} x_1 \\ x_2 \end{vmatrix} \right\|_E^2 = \|x_1\|_{L_2(\Omega)}^2 + \|x_2\|_{H^{-1}(\Omega)}^2, \qquad \|z\|_{H^{-1}(\Omega)} = \|z\|_{[\mathcal{D}(A^{1/2})]'} = \|A^{-1/2} z\|_{L_2(\Omega)}. \tag{1.7}$$

We next introduce the Dirichlet map D (natural "harmonic" extension of boundary data on Γ into the interior Ω), defined by

$$Du = y \quad \text{where} \begin{cases} -A(\xi, \partial) y = 0 & \text{in } \Omega, \\ y|_\Gamma = u & \text{in } \Gamma. \end{cases} \tag{1.8}$$

It is a well-known result of elliptic theory [L-M.1, I], [N.1] that

(1.9a) D is a continuous operator $H^\sigma(\Gamma) \to H^{\sigma+1/2}(\Omega)$, σ real.

We shall in particular use

$$\begin{aligned} &D\text{: continuous } L_2(\Gamma) \to H^{1/2-2\varepsilon}(\Omega) \equiv \mathcal{D}(A^{1/4-\varepsilon}), \qquad \varepsilon > 0, \\ &\|z\|_{H^{1/2-2\varepsilon}(\Omega)} = \|A^{1/4-\varepsilon} z\|_{L_2(\Omega)}, \end{aligned} \tag{1.9b}$$

where for the identification on the right of (1.9b) as well as for the identification

$$H_0^{3/2-2\varepsilon}(\Omega) \equiv \mathcal{D}(A^{3/4-\varepsilon}), \qquad \varepsilon > 0, \quad \varepsilon \neq \tfrac{1}{2}, \tag{1.9c}$$

we refer to [[F.2], [L.2, Appendix]. We shall also use freely that [L-M.1, p. 196 with Remark 2.6, p. 121], [L.3, Remark 5.1, p. 238]

$$\mathcal{D}(A) = \mathcal{D}(A^*), \quad \text{hence } \mathcal{D}(A^\theta) = \mathcal{D}(A^{*\theta}), \qquad 0 \leqq \theta < 1. \tag{1.9d}$$

Regarding cosine operator theory, we briefly recall that $C(t)$ is even and $C(0) = I$, thus $S(t)x = \int_0^t C(\tau)x\, d\tau$ is odd. Moreover $d^2C(t)x/dt^2 = -AC(t)x$, $x \in \mathcal{D}(A)$ and $dC(t)x/dt = -AS(t)x$, $x \in \mathcal{D}(A^{1/2})$. Also, [F.1, I, II]

(1.10) the map $t \to A^{1/2}S(t)x$, is well-defined and continuous for all $x \in L_2(\Omega)$ and $\|A^{1/2}S(t)\| \leqq M_\gamma e^{\gamma t}$.

It is important to note that [S.1] (see also [L-T.1, Prop. 2.1]

$$C(t)x - x = -A \int_0^t \tau C(t-\tau)x\, d\tau, \qquad x \in L_2(\Omega) \tag{1.11a}$$

which after integration by parts can be written as[5]

$$(1.11b)\qquad C(t)x - x = -A\int_0^t S(t-\tau)x\,d\tau = -A\int_0^t S(\sigma)x\,d\sigma, \qquad x \in L_2(\Omega).$$

A few other results on the cosine operator theory (e.g. identities (3.6) below) will be quoted when needed. Our main results on the Riccati's feedback synthesis of the optimal control are as follows (see §§ 2-4 for more complete results).

THEOREM 1.3. *Case* $T<\infty$: (i) *For* R *as in* (H.1), *see* (1.3), *the unique control* u^0 *of the optimal control problem* (O.C.P.) *on* $[0, T]$, $T<\infty$ *can be expressed in feedback form as* (*see* (2.28) *below*)

$$u^0(t) = u^0(t, t_0=0; y_0, y_1) = -\mathscr{B}^*\mathscr{P}(t)\begin{vmatrix} y^0(t) \\ \dot{y}^0(t)\end{vmatrix}$$

$$= -D^*A^*\int_t^T S^*(\tau - t)R\Phi_1(\tau, 0)\begin{vmatrix} y_0 \\ y_1\end{vmatrix} d\tau \quad \text{in } t \in [0, T].$$

$\Phi(\tau, t) = [\Phi_1(\ ,\ ), \Phi_2(\ ,\ )]$ *is the evolution operator of the optimal feedback system* (*see* (2.13)-(2.14) *below*), *while* $\mathscr{P}(t)$ *is defined* ((2.22a) *below*)

$$\mathscr{P}(t)x = \int_t^T e^{\mathscr{A}^*(\tau - t)}\begin{vmatrix} R & 0 \\ 0 & 0\end{vmatrix}\Phi(\tau, t)x\,d\tau.$$

(ii) *For* R *satisfying, in addition* (*see* (3.43) *below*) *the assumption*

$$(H.2)\qquad R\text{: continuous } H^{1/2-\delta}(\Omega) = \mathscr{D}(A^{1/4-\delta/2}) \to H_0^{1/2+\delta}(\Omega) \equiv \mathscr{D}(A^{1/4+\delta/2})$$

for some arbitrarily small $\delta > 0$, *henceforth kept fixed, then*:

(ii.1) *For initial data* $[y_0, y_1] \in Y_r \equiv H^{1/2-\delta}(\Omega) \times H^{-1/2-\delta}(\Omega)$ *we have* (*see more precise statement in Theorem* 3.11 *below*)[6]

$$u^0 \in H^{1/2+\delta, 1/2+\delta}(\Sigma), \text{ a fortiori } u^0 \in C([0, T]; L_2(\Gamma)),$$

$$y^0 \in C([0, T]; H^{1/2-\delta}(\Omega)) \cap H^{1/2-\delta}(0, T; L_2(\Omega)),$$

$$\dot{y}^0 \in C([0, T]; H^{-1/2-\delta}(\Omega)).$$

(ii.2) *The operator*

$$\mathscr{B}^*\mathscr{P}(t)x = D^*A^*\int_t^T S^*(\tau - t)R\Phi_1(\tau, t)x\,d\tau$$

is continuous $Y_r \to C([0, T]; L_2(\Gamma))$ (*Theorem* 4.1 *below*).

(ii.3) *The operator* $\mathscr{P}(t)$ *is a self-adjoint nonnegative definite operator on* E *and satisfies the following Riccati differential equation*:

$$(R.D.E.)\qquad \frac{d}{dt}(\mathscr{P}(t)x, y)_E = -(Rx_1, y_1)_\Omega - (\mathscr{P}(t)x, \mathscr{A}y)_E - (\mathscr{P}(t)\mathscr{A}x, y)_E$$

$$+(\mathscr{B}^*\mathscr{P}(t)x, \mathscr{B}^*\mathscr{P}(t)y)_\Gamma$$

[5] This last result (1.11) rigorously justifies for all $x \in L_2(\Omega)$ the following procedure—which however is only formal for $x \notin \mathscr{D}(A^{1/2})$:

$$-A\int_0^t S(\sigma)x\,d\sigma = \int_0^t -AS(\sigma)x\,d\sigma = \int_0^t \frac{dC(\sigma)}{d\sigma}x\,d\sigma = C(t)x - x.$$

[6] $H^{r,s}(\Sigma) = L_2(0, T; H^r(\Gamma)) \cap H^s(0, T; L_2(\Gamma))$, as usual [L-M.1, II] and similarly for $H^{r,s}(Q)$.

for all $x, y \in Y_r$ *and in* $t \in [0, T)$ *with terminal condition* $\mathscr{P}(T) = 0$ (*see Theorem* 4.3 *below*), *as well as the corresponding Riccati integral equation.*

Remark 1.1. A *direct* approach (i.e. with no reference to a control problem, in particular *without* using the preliminary optimality conditions available in the control problem) was carried out in [DaP-L-T.1] to study the well-posedness of the R.D.E. for hyperbolic dynamics. In particular, in the case of the present paper, this approach shows also *uniqueness* of a nonnegative self-adjoint solution of the R.D.E. (satisfying further properties), provided a stronger assumption is made on R, namely that $RA^{1/4+\varepsilon}$ is a bounded operator on $L_2(\Omega)$.

In § 5, we study the regulator problem, i.e. problem (1.2a) (or the corresponding version which penalizes quadratically also the velocity $\dot{y}$ in $H^{-1}(\Omega)$) for $T = \infty$, under the sole assumption (H.1) for R (which allows R to be, in particular, the identity on $L_2(\Omega)$) and under the usual minimal assumption of the finite cost (see Theorem 1.2 below). We then derive that:

(i) the optimal control of the regulator problem is given in feedback form by

$$-u^0(t) = \mathscr{B}^*\mathscr{P}\begin{vmatrix} y^0(t) \\ \dot{y}^0(t) \end{vmatrix} = \mathscr{B}^*\mathscr{P}\Phi(t)\begin{vmatrix} y_0 \\ y_1 \end{vmatrix};$$

(ii) here $\mathscr{P}$ is a nonnegative, self-adjoint operator on E satisfying the algebraic Riccati equation

$$(Rx_1, y_1)_\Omega + (\mathscr{P}x, \mathscr{A}y)_E + (\mathscr{P}\mathscr{A}x, y)_E = (\mathscr{B}^*\mathscr{P}x, \mathscr{B}^*\mathscr{P}y)_\Gamma, \qquad x, y \in \mathscr{D}(\mathscr{A});$$

(iii) moreover, the optimal solution $\Phi(t)|y_0/y_1|$ defines a s.c. semigroup on E generated by $\mathscr{A} - \mathscr{B}\mathscr{B}^*\mathscr{P}$. Uniqueness of the algebraic Riccati equation for the problem which also penalizes $\dot{y}$ is discussed in Theorem 5.11.

The approach presented in § 5 is particularly simple, when applied to a special but physically important case, which includes, in particular, the canonical wave equation, see Remark 5.2. A main feature of this approach is that—contrary to much, if not all, of the existing literature (see e.g. [F3] for a recent contribution on the *parabolic* case with Dirichlet boundary control)—the algebraic Riccati equation is *not* derived as a limit of the differential Riccati equation on $[0, T]$ via the limit process $T \uparrow \infty$: this would have required the study of the limit $\mathscr{B}^*\mathscr{P}_T(t)$ to $\mathscr{B}^*\mathscr{P}(\mathscr{P}_T(t) =$ Riccati operator of the problem on $[0, T]$) a particularly delicate question in the case of hyperbolic dynamics. This way we can dispense of assumption (H.2) for R, used for the Riccati differential equation, and derive the algebraic Riccati equation with R satisfying only assumption (H.1).

2. Optimality, the operators $\mathscr{P}(t)$ and $\Phi(\cdot, \cdot)$ and their preliminary properties. As an abstract version of the mixed problem (1.1), we can take the input-solution formula (see [L-T.1]) with $y_0 \in L_2(\Omega)$, $y_1 \in H^{-1}(\Omega)$:

$$y(t) = y(t, t_0 = 0; y_0, y_1) = C(t)y_0 + S(t)y_1 + (Lu)(t), \tag{2.1a}$$

$$\frac{dy}{dt}(t) = \frac{dy}{dt}(t, t_0 = 0; y_0, y_1) = -AS(t)y_0 + C(t)y_1 + \frac{d}{dt}(Lu)(t), \tag{2.1b}$$

where $C(t)$ and $S(t)x = \int_0^t C(\tau)x\, d\tau$ are, respectively, the cosine and "sine" operators on $L_2(\Omega)$, $t \in R$, generated by $-A$, and where the linear operators L and dL/dt are defined by

$$(Lu)(t) = A\int_0^t S(t-\tau)Du(\tau)\, d\tau, \tag{2.2a}$$

$$\frac{d}{dt}(Lu)(t) = A\int_0^t C(t-\tau)Du(\tau)\,d\tau. \tag{2.2b}$$

In view of Theorem 1.1 (ii), we have that

$$L\text{: continuous } L_2(\Sigma)\to C([0,T];L_2(\Omega)), \tag{2.3a}$$

$$\frac{dL}{dt}\text{: continuous } L_2(\Sigma)\to C([0,T];H^{-1}(\Omega)). \tag{2.3b}$$

Thus, by duality of Theorem 1.1 (i), the operator L^* adjoint of L, defined by $(Lu, v)_Q = (u, L^*v)_\Sigma$, and hence given by

$$\begin{aligned}(L^*v)(t) &= D^*A^*\int_t^T S^*(\tau-t)v(\tau)\,d\tau\\ &= D^*A^*\int_T^t S^*(t-\tau)v(\tau)\,d\tau, \qquad 0\leqq t\leqq T\end{aligned} \tag{2.4}$$

satisfies the property

$$L^*\text{: continuous } L_2(Q)\to L_2(\Sigma) \tag{2.5a}$$

and, in fact, even the property

$$L^*\text{: continuous } L_1(0,T;L_2(\Omega))\to L_2(\Sigma) \tag{2.5b}$$

by duality of Theorem 1.1 (ii). (Here and throughout the paper, we extend an inner product by continuity into a corresponding duality pairing, without changing notation and with no explicit mention necessarily made).

In order to treat the (O.C.P.) corresponding to the cost $J(u, y)$ in (1.2a), we introduce the Lagrangian

$$\mathscr{L}(u,y,p)\equiv\tfrac{1}{2}\{|u|_\Sigma^2+(Ry,y)_Q\}+(p,y-C(\cdot)y_0-S(\cdot)y_1-Lu)_Q. \tag{2.6}$$

The optimality conditions: $\mathscr{L}_y(u^0,y^0,p^0)=\mathscr{L}_u(u^0,y^0,p^0)=0$ yield, respectively e.g. as in [L-T.2, below (2.7), p. 47]

$$p^0=-Ry^0, \qquad u^0=L^*p^0, \quad \text{hence } u^0=-L^*Ry^0. \tag{2.7}$$

By eliminating y^0 between (2.1a) and (2.7), we obtain

$$u^0=-[I+L^*RL]^{-1}L^*R\{C(\cdot)y_0+S(\cdot)y_1\}\in L_2(\Sigma) \tag{2.8a}$$

and hence, using this in (2.1a) and (H.1) = (1.3)

$$y^0=\{I-L[I+L^*RL]^{-1}L^*R\}\{C(\cdot)y_0+S(\cdot)y_1\}. \tag{2.8b}$$

Note that, in (2.8), the (self-adjoint) inverse operator is well-defined and bounded on $L_2(\Sigma)$. On the other hand, if we eliminate u^0 instead between (2.1a) and (2.7), we obtain

$$y^0=[I+LL^*R]^{-1}\{C(\cdot)y_0+S(\cdot)y_1\}, \tag{2.8c}$$

$$u^0=-L^*R[I+LL^*R]^{-1}\{C(\cdot)y_0+S(\cdot)y_1\}, \tag{2.8d}$$

where we have to show the existence and boundedness of the new inverse operator. Indeed

$$[I+LL^*R]^{-1}=I-L[I+L^*RL]^{-1}L^*R \tag{2.8e}$$

---

[7] Thus, $L^*v=\partial z/\partial\eta_{A^*}|_\Gamma$, where: $z_{tt}=-A^*(\xi,\partial)z+v$, in $(0,T]\times\Omega$, $z(T,\xi)=z_t(T,\xi)=0$, and $z(t,\zeta)\equiv 0$, in $(0,T]\times\Gamma$.

well-defined and bounded on $L_2(Q)$. In fact, first $[I+LL^*R]$ is (i) *injective* on $L_2(Q)$ (since $x+LL^*Rx=0$, $x\in L_2(Q)$, after $L_2(Q)$-inner product with Rx, yields $L^*Rx=0$ by (1.3), hence $x=0$) and (ii) has *range dense* in $L_2(Q)$, since its adjoint $[I+RLL^*]$ has trivial null-space here ($z+RLL^*z=0$, $z\in L_2(Q)$, after $L_2(Q)$-inner product with LL^*z, yields $L^*z=0$ by (1.3), hence $z=0$). Secondly, the identity $I=I+LL^*R-L[I+L^*RL]^{-1}[I+L^*RL]L^*R$, where $[I+L^*RL]L^*R=L^*R[I+LL^*R]$, yields that the two sides in (2.8e) coincide on the range of $[I+LL^*R]$ on $L_2(Q)$ and so, by (ii) and the boundedness of the right-hand side operator, on all of $L_2(Q)$. Equations (2.8c-e) are justified. Notice that (2.8) provides the optimal solution (u^0, y^0) as $L_2(0, T)$-*trajectories* with values in $L_2(\Gamma)$ and $L_2(\Omega)$, respectively, in terms of the initial data. Our goal, however, is to express the optimal control u^0 in "feedback form;" i.e. as an operator acting pointwise in time (or a.e. in t) on the "measured" solution $[y^0(t), \dot{y}^0(t)]$ (so called "on line," or real time implementation in the engineering literature), precisely as described in Theorem 1.3. To accomplish this, an evolution operator will be introduced to describe the dynamics of the feedback system. Let s be an arbitrary time $0\leqq s<T$. Henceforth we take s as the new initial time of our optimal control problem with corresponding initial datum $y_s=[y_{0s}, y_{1s}]\in E$; i.e. we consider the optimal control problem of the introduction, but over the time interval $[s, T]$ rather than over $[0, T]$. We shall denote the corresponding optimal solution by $y^0(t, s; y_s)$ and $u^0(t, s; y_s)$.[8] The same procedure leading to the expressions (2.8c-d), once applied to the new problem, gives then

$$-u^0(\cdot, s; y_s)=L_s^*R\{y^0(\cdot, s; y_s)\}, \tag{2.9a}$$

$$u^0(t, s; y_s)=-L_s^*R[I_s+L_sL_s^*R]^{-1}\{C(\cdot-s)y_{0s}+S(\cdot-s)y_{1s}\}, \tag{2.9b}$$

$$y^0(t, s; y_s)=[I_s+L_sL_s^*R]^{-1}\{C(\cdot-s)y_{0s}+S(\cdot-s)y_{1s}\}, \tag{2.9c}$$

as elements of $L_2(s, T; L_2(\Gamma))$ and $L_2(s, T; L_2(\Omega))$, respectively. Here (compare with (2.2a) and (2.4)), we have

$$(L_su)(t)\equiv A\int_s^t S(t-\tau)Du(\tau)\,d\tau, \qquad s\leqq t\leqq T, \tag{2.10}$$

$$(L_s^*v)(t)\equiv\begin{cases}(L^*v)(t), & s\leqq t\leqq T,\\ 0, & 0\leqq t<s,\end{cases} \quad \text{a.e.} \tag{2.11}$$

The optimal dynamics is

$$\begin{aligned} y^0(t, s; y_s)&=C(t-s)y_{0s}+S(t-s)y_{1s}+\{L_s[u^0(\cdot, s; y_s)]\}(t),\\ \dot{y}^0(t, s; y_s)&=-AS(t-s)y_{0s}+C(t-s)y_{1s}+A\int_s^t C(t-\tau)Du^0(\tau, s; y_s)\,d\tau, \end{aligned} \tag{2.12}$$

(in $C([s, T];L_2(\Omega))$ and $C([s, T]; H^{-1}(\Omega))$, respectively). We now wish to obtain an explicit expression for the operator $\Phi(t, s)$ defined by

$$\begin{vmatrix} y^0(t, s; y_s)\\ \dot{y}^0(t, s; y_s)\end{vmatrix}=\Phi(t, s)\begin{vmatrix} y_{0s}\\ y_{1s}\end{vmatrix}, \qquad 0\leqq s\leqq t\leqq T, \tag{2.13}$$

which describes the evolution of the optimal solution originating at the starting point y_s at the initial time s. Using (2.9c) and its time derivative, we arrive at the explicit

---

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

expression of $\Phi(t, s)$, $(0 \leqq s \leqq t \leqq T)$

$$\Phi(t, s) = \begin{vmatrix} V_s C(\cdot - s) & V_s S(\cdot - s) \\ \cdot & \cdot \end{vmatrix} \equiv \begin{vmatrix} \Phi_1(t, s) \\ \Phi_2(t, s) \end{vmatrix} \tag{2.14}$$

where we have set $V_s \equiv [I_s + L_s L_s^* R]^{-1}$. For fixed s, by (2.13) and Theorem 1.1, $\Phi(\cdot, s)$ is a bounded operator $E \to C([s, T]; E)$. Note that L_s acts on functions defined after s (see (2.10)) and after s the actions of the operators L_s^* and L^* coincide (a.e.) (see (2.11)), thus

$$L_s L_s^* = L_s L^*. \tag{2.15}$$

Moreover, since obviously on $L_2(Q)$: $\|I_s + L_s^* R L_s\| \geqq 1$ and hence $\|[I_s + L_s^* R L_s]^{-1}\| \leqq 1$, uniformly in s, the version of (2.8e) corresponding to "s" gives $\|V_s\|_{\mathscr{L}(L_2(Q))} \leqq \text{const}_T$, for all $s \in [0, T]$ and hence from (2.14)

$$\|\Phi(\cdot, s)\|_{\mathscr{L}(E \to L_2(0, T; E))} \leqq M_T \quad \text{uniformly in } s \in [0, T]. \tag{2.16}$$

Actually, a stronger version of (2.16) is true. From (2.11) and (2.5a) we obtain $\|L_s^*\|_E = \|L^*\|_E$ and along with (2.9b) and the uniform bound on $\|V_s\|$ in the $\mathscr{L}(L_2(Q))$-norm we deduce

$$\|u^0(\cdot, s; x)\|_{L_2(s, T; L_2(\Gamma))} \leqq \text{const}_T \|x\|_E, \tag{2.17}$$

where the constant const_T depends on T but not on s. Moreover, if $u^0_{\text{ext}}(\tau, s; x)$ denotes the extension to $0 \leqq \tau < s$ by zero of the function $u^0(\tau, s; x)$, we have from (2.10)

$$\begin{aligned} \|\{L_s u^0(\cdot, s; x)\}(t)\|_{L_2(\Omega)} &\leqq \sup_{0 \leqq t \leqq T} \left\| A \int_0^t S(t-\tau) D u^0_{\text{ext}}(\tau, s; x)\, d\tau \right\|_{L_2(\Omega)}, \\ &\leqq C_T \|u^0_{\text{ext}}(\cdot, s; x)\|_{L_2(0, T; L_2(\Gamma))}, \qquad \text{(by (2.3a))} \\ &= C_T \|u^0(\cdot, s; x)\|_{L_2(s, T; L_2(\Gamma))} \leqq \text{const}_T \|x\|_E. \qquad \text{(by (2.17))} \end{aligned} \tag{2.18}$$

From (2.12)-(2.13), we then plainly obtain (by (2.18)):

$$\|\Phi_1(t, s)x\|_{L_2(\Omega)} = \|y^0(t, s; x)\|_{L_2(\Omega)} \leqq \text{const}_T \|x\|_E \tag{2.19a}$$

uniformly in s and t, $0 \leqq s \leqq t \leqq T$, which is the sought after improvement over (2.16). A similar argument gives

$$\|\Phi_2(t, s)x\|_{H^{-1}(\Omega)} \leqq \text{const}_T \|x\|_E \tag{2.19b}$$

uniformly in s and t, $0 \leqq s \leqq t \leqq T$. We can then collect some preliminary properties of $\Phi(t, s)$ in the next lemma.

LEMMA 2.1. *For the operator $\Phi(t, s)$, defined by* (2.13), $0 \leqq s \leqq t \leqq T$, *as a bounded operator from E into itself, the following properties hold true*:

(i) $\Phi(t, t) = I$ (*identity on E*), $0 \leqq t \leqq T$.

(ii) $\Phi(t, s)\Phi(s, \tau) = \Phi(t, \tau)$ (*transition*), $0 \leqq \tau \leqq s \leqq t \leqq T$.

(iii) *For each fixed s, the operator $\Phi(\cdot, s)$ is continuous $E \to C([s, T]; E)$ (strong continuity in the first variable).*

(iv) *The following uniform bound attains*:

$$\|\Phi(t, s)\|_{\mathscr{L}(E \to E)} \leqq \text{const}_T \quad \textit{for all } 0 \leqq s \leqq t \leqq T,$$

where const_T *depends only on T, but not on s.*

(v) *For t fixed, $0 < t \leqq T$, the operator $\Phi(t, \cdot)$ is continuous $E \to C([0, t]; E)$ (strong continuity in the second variable).*

Proof. (i) and (ii) above are obvious; (iii) and, in fact, (iv) were proved above (see (2.19a-b)). Moreover, strong continuity in the first variable (property (iii)) combined with the uniform bound in property (iv) yields strong continuity in the second variable in the usual way (see e.g. [L-T.2, p. 58]. □

In order to express the optimal control u^0 in a (pointwise, or a.e.) feedback form, as claimed in Theorem 1.3, we compute via (2.9a), (2.13), (2.11), and (2.4):

$$-u^0(t, s; y_s) = \{L_s^* R y^0(\cdot, s; y_s)\}(t) = \{L_s^* R \Phi_1(\cdot, s) y_s\}(t), \qquad s \leqq t \leqq T$$

$$= D^* A^* \int_t^T S^*(\tau - t) R \Phi_1(\tau, s) y_s \, d\tau.$$

If we now take the initial time $s = t$ with initial data $y_s = y_t = [y^0(t), \dot{y}^0(t)]$, we obtain the desired pointwise relation

$$-u^0(t) = D^* A^* \int_t^T S^*(\tau - t) R \Phi_1(\tau, t) \begin{vmatrix} y^0(t) \\ \dot{y}^0(t) \end{vmatrix} d\tau \tag{2.20}$$

$$= D^* A^* \int_t^T S^*(\tau - t) R \Phi_1(\tau, 0) y \, d\tau \tag{2.21}$$

with initial point $y = [y_1, y_2] \in E$ at $t = 0$, where in going from (2.20) to (2.21) we have used $[y^0(t), \dot{y}^0(t)] = \Phi(t, 0) y$, (see (2.13)), as well as $\Phi_1(\tau, t)\Phi(t, 0) y = [\Phi(\tau, t)\Phi(t, 0) y]_1 = [\Phi(\tau, 0) y]_1 = \Phi_1(\tau, 0) y$ by Lemma 2.1(ii). Indeed, we can see also directly the following.

PROPOSITION 2.2. *The expression* (2.21) *is well-defined as an* $L_2(\Sigma)$*-function for all* $y \in E$.

Proof. The proof will follow quickly from certain results to be established in § 3. It is therefore deferred to Appendix 1. □

Motivated by this, we now define an operator $\mathscr{P}(t)$ on E by

$$\mathscr{P}(t) x \equiv \int_t^T e^{\mathscr{A}^*(\tau - t)} \begin{bmatrix} R & 0 \\ 0 & 0 \end{bmatrix} \Phi(\tau, t) x \, d\tau, \qquad x \in E \tag{2.22a}$$

$$\equiv \int_t^T e^{\mathscr{A}^*(\tau - t)} \begin{vmatrix} R \Phi_1(\tau, t) x \\ 0 \end{vmatrix} d\tau.$$

For $y = [y_1, y_2]$ also in E, we compute

$$(\mathscr{P}(t) x, y)_E = \int_t^T \left(\begin{vmatrix} R \Phi_1(\tau, t) x \\ 0 \end{vmatrix}, e^{\mathscr{A}(\tau - t)} y \right)_E d\tau \tag{2.22b}$$

where, of course, from (2.1)

$$e^{\mathscr{A} t} = \begin{vmatrix} C(t) & S(t) \\ -AS(t) & C(t) \end{vmatrix} \quad \text{on } E, \tag{2.23}$$

to get that

$$(\mathscr{P}(t) x, y)_E = \int_t^T \{(C^*(\tau - t) R \Phi_1(\tau, t) x, y_1)_\Omega$$

$$+ (A^{*1/2} S^*(\tau - t) R \Phi_1(\tau, t) x, A^{-1/2} y_2)_\Omega\} \, d\tau.$$

We have written the last term so that, by (1.7b)

$$(A^{*1/2} S^*(\tau - t) R \Phi_1(\tau, t) x, A^{-1/2} y_2)_\Omega = (A^{1/2} A^{*1/2} S^*(\tau - t) R \Phi_1(\tau, t) x, y_2)_{H^{-1}(\Omega)}.$$

Thus, we obtain that $\mathscr{P}(t): E \to E$ is given alternatively by the more illuminating expression

$$\mathscr{P}(t)x = \int_t^T \begin{vmatrix} C^*(\tau - t)R\Phi_1(\tau, t)x \\ A^{1/2}A^{*1/2}S^*(\tau - t)R\Phi_1(\tau, t)x \end{vmatrix} d\tau, \qquad x \in E. \tag{2.24}$$

Now, by (1.9) with $\sigma = 0$, the Dirichlet map D is continuous: $L_2(\Gamma) \to H^{1/2}(\Omega) = H_0^{1/2}(\Omega)$ [L-M.1, I, p. 55]. Thus, D^* is continuous $H^{-1/2}(\Omega) \to L_2(\Gamma)$, ($L_2(\Gamma) =$ pivot space) where $(Du, y)_\Omega = (u, D^*y)_\Gamma$, for all $u \in L_2(\Gamma)$, and $y \in H^{-1/2}(\Omega)$. Next, we define an (unbounded) operator $\mathscr{B}^*: E \supset \mathscr{D}(\mathscr{B}^*) \to L_2(\Gamma)$ by

$$\mathscr{B}^*v = D^*A^{*1/2}A^{-1/2}v_2, \qquad v = [v_1, v_2] \in \mathscr{D}(\mathscr{B}^*). \tag{2.25}$$

Since $\mathscr{D}(\mathscr{B}^*) \supset L_2(\Omega) \otimes H^{-1/2}(\Omega)$, we conclude that $\mathscr{B}^*$ has domain dense in E. Also, we obtain from (2.24)-(2.25):

$$\mathscr{B}^*\mathscr{P}(t)x = D^*A^* \int_t^T S^*(\tau - t)R\Phi_1(\tau, t)x \, d\tau \tag{2.26}$$

$$= \{L^*R\Phi_1(\cdot, t)x\}(t) \tag{2.27}$$

for any $x \in E$ for which (2.26) is well-defined a.e. in t in $L_2(\Gamma)$. By comparison with (2.20), we see that this is certainly the case on the optimal points $[y^0(t), \dot{y}^0(t)]$ and, indeed, the optimal control u^0 in (2.21) can then be rewritten as an $L_2(\Sigma)$-function as

$$-u^0(t) = \mathscr{B}^*\mathscr{P}(t) \begin{vmatrix} y^0(t) \\ \dot{y}^0(t) \end{vmatrix} = \mathscr{B}^*\mathscr{P}(t)\Phi(t, 0) \begin{vmatrix} y_0 \\ y_1 \end{vmatrix}, \qquad [y_0, y_1] \in E \tag{2.28}$$

more desirable than (2.20), since now $\mathscr{P}(t)$ has range in, and $\mathscr{B}^*$ acts from, the basic space E. $\mathscr{P}(t)$ will be the Riccati operator. This completes the proof of part (i) in Theorem 1.3.

Some (to be expected) properties of $\mathscr{P}(t)$ are collected next.

LEMMA 2.3. *The operator $\mathscr{P}(t)$ defined by* (2.22) *or* (2.24) *satisfies the following properties*:

(i) *For each fixed $t \in [0, T]$, $\mathscr{P}(t)$ is a bounded linear operator on E, and, in fact, $\mathscr{P}(t)$ is continuous $E \to C([0, T]; E)$.*

(ii) *For each fixed $t \in [0, T]$, the following identity—which is symmetric in x and y (both in E)—holds*:

$$(\mathscr{P}(t)x, y)_E = \int_t^T (R\Phi_1(\tau, t)x, \Phi_1(\tau, t)y)_\Omega \, d\tau + \int_t^T (\mathscr{B}^*\mathscr{P}(\tau)\Phi(\tau, t)x, \mathscr{B}^*\mathscr{P}(\tau)\Phi(\tau, t)y)_\Gamma \, d\tau. \tag{2.29}$$

Thus,

1) *$\mathscr{P}(t) = \mathscr{P}^*(t)$ and $\mathscr{P}(t)$ is self-adjoint on E;*

2) *$\mathscr{P}(t)$ is nonnegative definite.*

(iii) *The minimal* (optimal) *value of the performance index J of the optimal control problem on $[s, t]$, $s < T$, that initiates at y_s at time s is*

$$J^0(u^0(\cdot, s; y_s), y^0(\cdot, s; y_s)) = (\mathscr{P}(s)y_s, y_s)_E. \tag{2.30}$$

Hence, for any $x \in E$, the map $t \to (\mathscr{P}(t)x, x)_E$ is monotone decreasing.

Proof. (i) That $\mathscr{P}(t)$ is bounded on E for fixed t, or that $\mathscr{P}(t)x \in L_\infty(0, T; E)$, $x \in E$, follows immediately from (2.24) using: (1) the uniform bound on Φ_1 in Lemma 2.1

(iv); (2) the norm (1.7c) on $H^{-1}(\Omega)$; and (3) the property (1.10) for $A^{*1/2}S^*(t)z$, $z \in L_2(\Omega)$. To show that, in fact, $\mathscr{P}(t)x \in C([0, T]; E)$, one adds and subtracts a same quantity and uses, in addition to the above, the Lebesgue's dominated convergence theorem [H-P.1, p. 83]. Details are omitted.

(ii) By combining (2.1a, b) and recalling the definition of $\Phi(\cdot, \cdot)$ in (2.12) and also (2.23), we can write for $x \in E$

$$\Phi(\tau, t)x = e^{\mathscr{A}(\tau-t)}x + A\int_t^\tau \begin{vmatrix} S(\tau-\sigma)Du^0(\sigma, \tau; x) \\ C(\tau-\sigma)Du^0(\sigma, \tau; x) \end{vmatrix} d\sigma,$$

i.e. in view of (2.28)

$$e^{\mathscr{A}(\tau-t)}x = \Phi(\tau, t)x + A\int_t^\tau \begin{vmatrix} S(\tau-\sigma)D\mathscr{B}^*\mathscr{P}(\sigma)\Phi(\sigma, t)x \\ C(\tau-\sigma)D\mathscr{B}^*\mathscr{P}(\sigma)\Phi(\sigma, t)x \end{vmatrix} d\sigma. \tag{2.31}$$

We next substitute the expression of $e^{\mathscr{A}(\tau-t)}$ given by (2.31) into the right-hand side of (2.22b) to get

$$(\mathscr{P}(t)x, y)_E = \int_t^T (R\Phi_1(\tau, t)x, \Phi_1(\tau, t)y)_\Omega \, d\tau + I \tag{2.32}$$

where, changing the order of integration on I (see Appendix 2), we can write

$$I = \int_t^T \int_t^\tau (\Phi_1(\tau, t)x, AS(\tau-\sigma)D\mathscr{B}^*\mathscr{P}(\sigma)\Phi(\sigma, t)y)_\Omega \, d\sigma \, d\tau$$

$$= \int_t^T \left(\int_\sigma^T D^*A^*S^*(\tau-\sigma)\Phi_1(\tau, t)x \, d\tau, \mathscr{B}^*\mathscr{P}(\sigma)\Phi(\sigma, t)y \right)_\Omega d\sigma$$

(since $\Phi(\tau, \sigma)\Phi(\sigma, t)x = \Phi(\tau, t)x$, by Lemma 2.1(ii), and thus $\Phi_1(\tau, \sigma)\Phi(\sigma, t)x = \Phi_1(\tau, t)x$)

$$= \int_t^T (\mathscr{B}^*\mathscr{P}(\sigma)\Phi(\sigma, t)x, \mathscr{B}^*\mathscr{P}(\sigma)\Phi(\sigma, t)y)_\Gamma \, d\sigma \tag{2.33}$$

where, in the last step, we have made use of (2.26). Setting $x = y$ in (2.32)-(2.33) yields (ii.1)-(ii.2). Property (iii) follows from (ii) via (2.12) and (2.28). □

3. Regularity results.

3.1. Further regularity results for the mixed problem (1.1) needed for the (O.C.P.). In order to establish appropriate regularity for the optimal pair u^0, y^0 of the (O.C.P.) and to derive a Riccati equation, it is necessary to obtain preliminary *certain* regularity results of the mixed problem (1.1). This will be done in the present section, since we cannot merely quote these results from our own, or others' previous work. The technique used is in the style of [L-T.1], [L-T.3]. To this end, we begin with an abstract lemma, which will be used repeatedly.

LEMMA 3.1. *With X and Y two given separable Hilbert spaces, let $F(t) \in \mathscr{L}(X, Y)$ a.e. in $t \in [0, T]$. Define a (linear) operator $F(\cdot)$ by: $F(\cdot)x = F(t)x \in Y$, a.e. in $t \in [0, T]$, $x \in X$ and assume that $F(\cdot)$ is continuous $X \to L_2(0, T; Y)$; i.e.*

$$\|F(\cdot)x\|^2_{L_2(0,T;Y)} \equiv \int_0^T \|F(t)x\|^2_Y \, dt \leqq C_T\|x\|^2_X. \tag{3.1}$$

Then, for any $x(t) \in L_1(0, T; X)$, the operator $\bar{F}$ defined by: $(\bar{F}x(\cdot))(t) \equiv F(t)\int_0^t x(\tau) \, d\tau$

is continuous $L_1(0, T; X) \to L_2(0, T; Y)$:

$$\int_0^T \left\| F(t) \int_0^t x(\tau)\, d\tau \right\|_Y^2 dt \leqq C_T \int_0^T \|x(t)\|_X\, dt. \tag{3.2}$$

Proof of Lemma 3.1. If $F^*(t) \in \mathscr{L}(Y, X)$ a.e. in t is the adjoint of $F(t)$: $(F(t)x, y)_Y = (x, F^*(t)y)_X$ a.e., $x \in X$, $y \in Y$, we compute for $g \in L_2(0, T; Y)$, through a change of the order of integration and Schwarz inequality

$$\begin{aligned}
(\bar{F}x(\cdot), g)_{L_2(0, T; Y)} &= \int_0^T (F(t) \int_0^t x(\tau)\, d\tau, g(t))_Y\, dt \\
&= \int_0^T \int_0^t (x(\tau), F^*(t)g(t))_X\, d\tau\, dt \\
&= \int_0^T \int_\tau^T (F(t)x(\tau), g(t))_Y\, dt\, d\tau \qquad (3.3) \\
&\leqq \|g\|_{L_2(0, T; Y)} \int_0^T \left\{ \int_0^T \|F(t)x(\tau)\|_Y^2\, dt \right\}^{1/2} d\tau \\
&\leqq C_T \|g\| \int_0^T \|x(\tau)\|_X\, d\tau < \infty \qquad \text{(by (3.1))}
\end{aligned}$$

and $\bar{F}$ is well-defined from all of $L_1(0, T; Y)$ into $L_2(0, T; Y)$. Moreover, from (3.3) the adjoint $\bar{F}^*$ of $\bar{F}$, given by

$$(\bar{F}^* g)(\tau) \equiv \textstyle\int_\tau^T F^*(t)g(t)\, dt \tag{3.4}$$

is certainly well-defined on a dense set (e.g. step functions) of $L_2(0, T; Y)$ into $L_1(0, T; X)$. Thus, $\bar{F}$ is closable [K.1, p. 168] and by the closed graph theorem, $\bar{F}$ is bounded as claimed. □

The next theorem was proved in [L-T.3] and has a trace theory interpretation for the mixed problem (1.1). It played a crucial role there in our argument that succeeded in strengthening the regularity of the operator L in Theorem 1.1 from statement (i) into statement (ii). For completeness, we shall also include a sketch of its proof.

THEOREM 3.2 [L-T.3, "only if" part of Theorem 2.1]. *The operators* J_1 *and* J_2

$$(J_1 x)(t) \equiv D^* A^* S^*(t)x;\ (J_2 x)(t) \equiv D^* A^{*1/2} C^*(t)x \tag{3.5}$$

are bounded linear operators $L_2(\Omega) \to L_2(\Sigma)$.

Proof (*sketch*). It makes crucial use of the boundedness of $L^*: L_2(Q) \to L_2(\Sigma)$ (statement (2.5a)), thus of statement (i) in Theorem 1.1. The following identities hold for cosine and sine operators [T-W.1], [F.1].

$$S(s+t) + S(s-t) = 2S(s)C(t), \tag{3.6a}$$

$$S(s+t) = S(s)C(t) + S(t)C(s), \tag{3.6b}$$

$$C(t+s) - C(t-s) = -2AS(t)S(s). \tag{3.6c}$$

For $x \in L_2(\Omega)$, we now compute directly $L^*C(\cdot)x$ via (2.4) and (3.6a) (see also (1.11))

$$
\begin{aligned}
L^*C^*(\cdot)x &= D^*A^* \int_t^T S^*(\tau - t)C^*(\tau)x \, d\tau \\
&= \frac{1}{2} D^*A^* \int_t^T [S^*(2\tau - t) - S^*(t)]x \, d\tau \\
&= -\frac{1}{2}\left[\frac{1}{2} D^*C^*(2\tau - t)x\right]_{\tau=t}^{\tau=T} - \frac{(T-t)}{2} D^*A^*S^*(t)x \\
&= -\frac{1}{4} D^*C^*(2T - t)x - \frac{1}{4} D^*C^*(t)x - \frac{(T-t)}{2} D^*A^*S^*(t)x.
\end{aligned} \tag{3.7}
$$

Since $L^*C^*(\cdot)x \in L_2(\Sigma)$, and the first and second terms on the right of (3.7) are in $C([0, T]; L_2(\Gamma))$, we deduce that $(T-t)D^*A^*S^*(t)x \in L_2(0, T; L_2(\Gamma))$ for *any* T, and hence that $D^*A^*S^*(t)x \in L_2(0, T; L_2(\Gamma))$. Next, one proves that J_1 is closable, since its dual J_1^* (given explicitly below in (3.8)) is densely defined. By the closed graph theorem J_1 is bounded $L_2(\Omega) \to L_2(\Sigma)$.

The proof for J_2 is similar, starting now with $L^*A^{*1/2}S^*(\cdot)x$. See [L-T.3] for details. (We also refer to [L-T.3] for the converse of this result, which is not needed here: i.e. if J_1 and J_2 are bounded $L_2(\Omega) \to L_2(\Sigma)$, then L^* is bounded $L_2(Q) \to L_2(\Sigma)$.)

Remark 3.1. Theorem 3.2 admits a trace theory interpretation of problem (1.1) with homogeneous B.C. $u \equiv 0$, since D^*A^* is nothing but the operator of the co-normal derivative on Γ

$$
\begin{aligned}
&-D^*A^* = \frac{\partial}{\partial \nu_{A^*}}\bigg|_\Gamma, \quad \text{in particular } -D^*A^* = \frac{\partial}{\partial \nu}\bigg|_\Gamma \\
&\qquad\qquad \text{in the Laplacian case } -A(\xi, \partial) = \Delta.
\end{aligned} \tag{3.8}
$$

See in [L-T.3, Remark 2.1] or in [L-L-T.1, (3.4)] for more details. □

By taking the dual operator to J_1 and J_2: $(J_i x, u)_\Sigma = (x, J_i^* u)_\Omega$, we obtain the following.

COROLLARY 3.3. *For* $u \in L_2(\Sigma)$, *the integrals*

$$
A \int_0^t S(\tau)Du(\tau) \, d\tau \quad \text{and} \quad A^{1/2} \int_0^t C(\tau)Du(\tau) \, d\tau, \qquad 0 \leqq t \leqq T \tag{3.9}
$$

are well-defined as $L_2(\Omega)$*-functions, which moreover, are continuous in* $t \in [0, T]$.

THEOREM 3.4. *For* $0 \leqq \theta \leqq 1$ *and with reference to* (1.9c) *the operators* J_i *of* (3.5)

$$
(J_1 x)(t) = D^*A^*S^*(t)x, \qquad (J_2 x)(t) = D^*A^{*1/2}C^*(t)x
$$

are continuous $\mathscr{D}(A^\theta) \equiv \mathscr{D}(A^{*\theta}) \to H^{2\theta, 2\theta}(\Sigma)$ (*footnote* 6).

Proof. It suffices to prove the case $\theta = 1$, and interpolate with Theorem 3.2 ($\theta = 0$). With $x \in \mathscr{D}(A^*)$, we compute

$$
\left.
\begin{aligned}
&\frac{d^2}{dt^2} D^*A^*S^*(t)x = \frac{d}{dt} D^*A^*C^*(t)x = -D^*A^*S^*(t)A^*x, \\
&\frac{d^2}{dt^2} D^*A^{*1/2}C^*(t)x = -\frac{d}{dt} D^*A^{*1/2}S^*(t)A^*x = -D^*A^{*1/2}C^*(t)A^*x,
\end{aligned}
\right\} \in L_2(\Sigma) \tag{3.10}
$$

continuously, by Theorem 3.2 and $J_i x \to H^2(0, T; L_2(\Gamma))$. To prove that

$$
J_i x \in L_2(0, T; H^2(\Gamma)), \qquad x \in \mathscr{D}(A^*), \tag{3.11}
$$

we introduce, as in [L-L-T.1], a second order operator

$$B=\sum_{i,j} b_{ij}\frac{\partial^2}{\partial\xi_i\,\partial\xi_j}$$

on $\bar{\Omega}$, *tangential* to Γ (i.e. without transversal derivatives to Γ, when expressed in local coordinates) and with smooth coefficients b_{ij} on $\bar{\Omega}$. For simplicity of notation (only), we take A self-adjoint. We argue only for J_1. The solution of the problem

$$\begin{aligned} &\phi_{tt}=-A(\xi,\partial)\phi && \text{in } Q,\\ &\phi|_{t=0}=0, \qquad \phi_t|_{t=0}=x\in\mathscr{D}(A)\subset H^2(\Omega) && \text{in } \Omega,\\ &\phi|_\Gamma=0 && \text{in } \Sigma, \end{aligned} \tag{3.12}$$

is $\phi(t)=S(t)x$ and by Remark 3.1 with $\partial/\partial\nu_A$ written as $\partial/\partial\nu$

$$(J_1x)(t)=D^*AS(t)x=\frac{\partial}{\partial\nu}S(t)x\bigg|_\Gamma=\frac{\partial\phi(t)}{\partial\nu}\bigg|_\Gamma. \tag{3.13}$$

If we introduce a new variable

$$w=B\phi, \tag{3.14}$$

then proving (3.11) for $i=1$ is equivalent to proving that

$$\frac{\partial}{\partial\nu}B\phi\bigg|_\Gamma=\frac{\partial w}{\partial\nu}\bigg|_\Gamma\in L_2(\Sigma). \tag{3.15}$$

Notice that, as is well known (e.g. [K-N.1]), the commutator

$$K\equiv -BA(\cdot,\cdot)+A(\cdot,\cdot)B \tag{3.16}$$

is an operator of order $2+2-1=3$ in ξ, with smooth coefficients in $\bar{\Omega}$. From (3.14) and (3.12), we obtain the problem for w

$$\begin{aligned} &w_{tt}=-A(\xi,\partial)w+f && \text{in } Q,\\ &w|_{t=0}=0, \qquad w_t|_{t=0}=Bx\in L_2(\Omega) && \text{in } \Omega\\ &w|_\Gamma=0 && \text{in } \Sigma, \end{aligned} \tag{3.17}$$

where we have set $f\equiv K\phi$. Thus, since $\phi(t)=S(t)x\in\mathscr{D}(A^{3/2})\subset H^3(\Omega)$ for $x\in\mathscr{D}(A)$, we have

$$f\equiv K\phi\in L_2(Q). \tag{3.18}$$

The solution of (3.17) is

$$w(t)=S(t)(Bx)+\int_0^t S(t-\tau)f(\tau)\,d\tau. \tag{3.19}$$

By Theorem 3.2, since $Bx\in L_2(\Omega)$, then

$$D^*AS(t)(Bx)\in L_2(\Sigma). \tag{3.20}$$

Using identity (3.6b), we can write

$$\begin{aligned} D^*A\int_0^t S(t-\tau)f(\tau)\,d\tau=D^*AS(t)\int_0^t C(\tau)f(\tau)\,d\tau\\ -D^*A^{1/2}C(t)\int_0^t A^{1/2}S(\tau)f(\tau)\,d\tau. \end{aligned} \tag{3.21}$$

Now, both integrands in (3.21) are in $L_2(0, T; L_2(\Omega))$ by (3.18) and (1.10). Thus, Theorem 3.2 allows us to apply the abstract Lemma 3.1 with $X = L_2(\Omega)$, $Y = L_2(\Gamma)$ and obtain that each of the two terms on the right-hand side of (3.21) is in $L_2(\Sigma)$. Thus, returning to (3.19), and using (3.20)-(3.21), we conclude that

$$\frac{\partial w}{\partial \nu} = D^*AS(t)(Bx) + D^*A\int_0^t S(t-\tau)f(\tau)\,d\tau \in L_2(\Sigma) \tag{3.22}$$

and (3.15) is proved. Thus, (3.11) holds true if $i = 1$. A similar argument proves (3.11) for $i = 2$. An application of the closed graph theorem completes the proof. □

The next theorem complements statement (2.5a) on the regularity of the operator L^*; in its proof, we shall use Theorem 3.4 for $\theta = 1$.

THEOREM 3.5. *With* $0 \leqq \theta \leqq 1$, *we have for the operator* L^* *in* (2.4)

$$L^*\text{: continuous } L_2(0, T; \mathscr{D}(A^\theta)) \to H^{2\theta,2\theta}(\Sigma); \tag{3.23a}$$

equivalently,

$$L^*A^{-\theta}\text{: continuous } L_2(Q) \to H^{2\theta,2\theta}(\Sigma). \tag{3.23b}$$

Proof. It suffices to prove the case $\theta = 1$ and interpolate with (2.5a). To prove

$$v \in L_2(0, T; \mathscr{D}(A) \equiv \mathscr{D}(A^*)) \to L^*v \in L_2(0, T; H^2(\Gamma)) \tag{3.24}$$

we use (2.4) and indentity (3.6b) to get

$$\begin{aligned}(L^*v)(t) &= D^*A^*\int_t^T S^*(\tau - t)v(\tau)\,d\tau \\ &= D^*C^*(t)A^{*-1/2}\int_t^T A^{*1/2}S^*(\tau)A^*v(\tau)\,d\tau \\ &\quad - D^*S^*(t)\int_t^T C^*(\tau)A^*v(\tau)\,d\tau.\end{aligned} \tag{3.25}$$

A fortiori from Theorem 3.4 with $\theta = 1$, we are authorized to apply the abstract Lemma 3.1 with $F(t)$ either $D^*C^*(t)A^{*-1/2}$ or $D^*S^*(t)$, while $X = L_2(\Omega)$ and $Y = H^2(\Gamma)$. We thus obtain that each term in (3.25) (right) is in $L_2(0, T; H^2(\Gamma))$, i.e. (3.24). To prove the remaining part

$$v \in L_2(0, T; \mathscr{D}(A^*)) \to \frac{d^2(L^*v)}{dt^2}(t) \in L_2(\Sigma) \tag{3.26}$$

we compute from (2.4)

$$\frac{d^2(L^*v)}{dt^2}(t) = -\frac{d}{dt}D^*A^*\int_t^T C^*(\tau - t)v(\tau)d\tau = D^*A^*v(t) - (L^*(A^*v))(t) \tag{3.27}$$

and (3.27) implies (3.26) via (2.5a). □

The following consequences, particularly Theorem 3.8 below, will be needed in § 3.2, in the study of the regularity of the optimal solutions of the O.C.P.

THEOREM 3.6.

(i)

$$\left.\begin{array}{l} u \in C([0, T]; H^{1/2}(\Gamma)) \cap H^1(0, T; L_2(\Gamma)) \\ u(0) = 0 \text{ (}\textit{compatibility relation}\text{)} \end{array}\right\} \to Lu \in C([0, T]; H^1(\Omega)). \tag{3.28}$$

(ii) *Moreover,*

$$\left.\begin{array}{l} u \in H^1(0, T; L_2(\Gamma)) \\ u(0) = 0 \end{array}\right\} \to \frac{dLu}{dt} \in C([0, T]; L_2(\Omega)). \tag{3.29}$$

Proof. (i) Integrating (2.2a) by parts with $\dot{u} \in L_2(\Sigma)$ and using the *compatibility relation* gives

$$\begin{aligned} (Lu)(t) &= A \int_0^t AS(t-\tau)A^{-1}Du(\tau)\, d\tau = A \int_0^t \frac{dC(t-\tau)}{d\tau} A^{-1}Du(\tau)\, d\tau \\ &= C(0)Du(t) - C(t)Du(0) - \int_0^t C(t-\tau)D\dot{u}(\tau)\, d\tau \\ &= Du(t) - \left(A^{-1}\frac{dL\dot{u}}{dt}\right)(t) \end{aligned} \tag{3.30}$$

where by (2.2b) and the identity

$$C(t+s) = C(t)C(s) + AS(t)S(s) \qquad s, t \in R \tag{3.31}$$

(obtained summing up identities (2.4) and (2.28) in [T-W.1]),

$$\begin{aligned} \left(A^{-1}\frac{dL\dot{u}}{dt}\right)(t) &= \int_0^t C(t-\tau)D\dot{u}(\tau)\, d\tau \\ &= C(t)\int_0^t C(\tau)D\dot{u}(\tau)\, d\tau - S(t)A^{1/2}\int_0^t A^{1/2}S(\tau)D\dot{u}(\tau)\, d\tau. \end{aligned} \tag{3.32}$$

By Corollary 3.3 and (1.10), each of the two terms on (3.32) (right) is in $C([0, T]; \mathcal{D}(A^{1/2})) = H_0^1(\Omega))$. Returning to (3.30), we have $Du \in C([0, T]; H^1(\Omega))$, by elliptic theory with $u \in C([0, T]; H^{1/2}(\Gamma))$, and (i) follows. (ii) Differentiating (3.30) (via the left-hand side of (3.32)) yields for $\dot{u} \in L_2(\Sigma)$

$$\left(\frac{dLu}{dt}\right)(t) = A\int_0^t S(t-\tau)D\dot{u}(\tau)\, d\tau = (L\dot{u})(t) \in C([0, T]; L_2(\Omega)) \tag{3.33}$$

by statement (2.3a). □

Theorem 3.6 states what is needed in the sequel: a more complete result is given in [L-L-T.1]. A fortiori, since [L-M.1, I, Thm. 3.1, p. 19]

$$u \in H^{1,1}(\Sigma) \Rightarrow u \in C([0, T]; H^{1/2}(\Gamma)) \tag{3.34}$$

we obtain part (i) of the following corollary.

COROLLARY 3.7. *We have*

(i)

$$\left.\begin{array}{l} u \in H^{1,1}(\Sigma) \\ u(0) = 0 \end{array}\right\} \to \begin{array}{l} Lu \in C([0, T]; H^1(\Omega)) \\ \dfrac{dLu}{dt} \in C([0, T]; L_2(\Omega)) \end{array} \to Lu \in H^{1,1}(\Omega), \tag{3.35}$$

(ii) *For* $0 \leqq \theta < \frac{1}{2}$

$$u \in H^{\theta,\theta}(\Sigma) \to \left\{\begin{array}{l} Lu \in C([0, T]; H^{\theta}(\Omega)) \\ \dfrac{dLu}{dt} \in C([0, T]; H^{\theta-1}(\Omega)) \end{array}\right. \to Lu \in H^{\theta,\theta}(Q). \tag{3.36}$$

(iii) *Indeed for* $0 \leqq \theta < \frac{1}{2}$

$$u \in H^{\theta}(0, T; L_2(\Gamma)) \to \frac{dLu}{dt} \in C([0, T]; H^{\theta-1}(\Omega)). \tag{3.37}$$

Proof. We interpolate with $0 \leqq \theta < \frac{1}{2}$, so that the C.R. (compatibility relation) $u(0) = 0$ is irrelevant, between (2.3) and part (i) to get part (ii); and between (2.3b) and Theorem 3.6(ii) to get part (iii), see [L-M.1, Thm. 14.2, p. 95]. □

THEOREM 3.8. *With reference to the operators* L *and* L^* *in* (2.2) *and* (2.4), *the following regularity results hold*:

(i) *For* $0 \leqq \beta < \frac{1}{4}$, *we have*

$$A^{\beta} LL^* A^{-\beta}\text{: continuous } L_2(Q) \to C([0, T]; L_2(Q)); \tag{3.38a}$$

equivalently, with $\mathcal{D}(A^{\beta}) = H^{2\beta}(\Omega)$ (*see* (1.9b))

$$LL^*\text{: continuous } L_2(0, T; \mathcal{D}(A^{\beta})) \to C([0, T]; \mathcal{D}(A^{\beta})). \tag{3.38b}$$

(ii) *Moreover, for any* $\varepsilon > 0$

$$\frac{d}{dt}(LL^*) = \frac{dL}{dt} L^*\text{: continuous } L_2(0, T; H^{1/2-\varepsilon}(\Omega)) \to C([0, T]; H^{-1/2-\varepsilon}(\Omega)). \tag{3.39}$$

Proof. (i) We have

$$L^* A^{-\beta}\text{: continuous } L_2(Q) \to H^{2\beta, 2\beta}(\Sigma) \tag{3.40}$$

by (3.23b) in Theorem 3.5, and restricting β to $0 \leqq \beta < \frac{1}{4}$ so that $H^{2\beta}(\Omega) = \mathcal{D}(A^{\beta})$, we have

$$A^{\beta} L\text{: continuous } H^{2\beta, 2\beta}(\Sigma) \to C([0, T]; L_2(\Omega)) \tag{3.41}$$

a fortiori from Corollary 3.7(ii).

(ii) Plainly,

$$\frac{d(LL^*)}{dt} = \frac{dL}{dt} L^* \tag{3.42}$$

from (2.2a) and (2.4). But, by (3.23) in Theorem 3.5

$$L^*\text{: continuous } L_2(0, T; \mathcal{D}(A^{1/4-\varepsilon/2}) = H^{1/2-\varepsilon}(\Omega)) \to H^{1/2-\varepsilon, 1/2-\varepsilon}(\Sigma)$$

and part (ii) follows a fortiori from Corollary 3.7(iii) with $\theta = \frac{1}{2} - \varepsilon$. □

3.2. Regularity of optimal pair u^0, y^0 of the (O.C.P.). Once in possession of the regularity results of the previous subsection for the general mixed problem (1.1), we can now proceed to obtain regularity results for the optimal pair (u^0, y^0) of the (O.C.P.). While throughout § 2, the penalization operator R was subject only to the general assumption (H.1), so that R could also be the identity I on $L_2(\Omega)$, it turns out that the regularity theory for the optimal pair u^0, y^0 is particularly rich, if R is subject to any one of various possible "rather minimal" assumptions of ε-regularity type for an arbitrary $\varepsilon > 0$. This will also help in the derivation of the Riccati equation. Of said possible choices, we opt here for the following two versions. In addition to the standing hypothesis (H.1), R is assumed to satisfy either

$$\text{(H.2) } R\text{: continuous } H^{1/2-\delta}(\Omega) \equiv \mathcal{D}(A^{1/4-\delta/2}) \to H_0^{1/2+\delta}(\Omega) \equiv \mathcal{D}(A^{1/4+\delta/2}) \tag{3.43}$$

or else, less restrictively,

$$\text{(H.2')} \; R\text{: continuous } H^{1/2-\delta}(\Omega) \equiv \mathcal{D}(A^{1/4-\delta/2}) \to \mathcal{D}(A^{1/4}) \tag{3.44}$$

for some *arbitrarily small* number $\delta > 0$. Henceforth kept *fixed*: as a result of (H.2) or (H.2'), we shall obtain two slightly different regularity results for u^0 in Theorem 3.11 below. As the injection $\mathscr{D}(A^{1/4})$ or $(\mathscr{D}(A^{1/4+\delta/2})) \to H^{1/2-\delta}(\Omega)$ is compact, we have that

$$\left.\begin{array}{r}\text{under either assumption (3.43)}\\ \text{or assumption (3.44)}\end{array}\right\} \quad R \text{ is compact } H^{1/2-\delta}(\Omega) \to \text{itself} \tag{3.45}$$

a property needed later. With T and $\delta > 0$ fixed once and for all, we next introduce a function space

$$W = \left\{ f : f \in L_2(0, T; H^{1/2-\delta}(\Omega)), \frac{df}{dt} \in L_2(0, T; H^{-1/2-\delta}(\Omega)) \right\}. \tag{3.46}$$

By interpolation, $f \in W$ implies $D_t^{1/2-\delta} f \in L_2(Q)$, i.e.

$$f \in W \Rightarrow f \in H^{1/2-\delta, 1/2-\delta}(Q). \tag{3.47}$$

We shall need below a well-known compactness result, which we find convenient to state explicitly for easy use in various situations.

LEMMA 3.9 [A.1]. *Let B_0, B, B_1 be three Banach spaces with* (i) $B_0 \subset B \subset B_1$, $B_i =$ *reflexive*, $i = 0, 1$ (*where* $\subset$ *means that the inclusion is algebraic and topological*) *and with* (ii) *injection $B_0 \to B$ compact. Define the Banach space*

$$X \equiv \left\{ v : v \in L^{p_0}(0, T; B_0),\ v' = \frac{dv}{dt} \in L^{p_1}(0, T; B_1) \right\}$$

$0 < T < \infty$, $1 < p_i < \infty$, $i = 0, 1$ *with norm*

$$\|v\|_{L^{p_0}(0, T; B_0)} + \|v'\|_{L^{p_1}(0, T; B_1)}.$$

Then, the injection $X \to L^{p_0}(0, T; B)$ is compact.

A first application of this Lemma 3.9 is in the following theorem, a first step toward the regularity of y^0 (refer to (2.8c)).

THEOREM 3.10. *Let R satisfy* (H.1) *and* (H.2') = (3.44). *Then*

(i) *the operator LL^*R is compact $W \to W$ (see* (3.46));

(ii) *the operator $[I + LL^*R]$ is invertible on W, with bounded inverse on W.*

Proof. (i) With reference to (3.46) and Y below, we have:

$$W \xrightarrow[\text{continuous}]{R} Y \xrightarrow[\text{compact}]{\text{injection}} L_2(0, T; H^{1/2-\delta}(\Omega)) \xrightarrow[\text{continuous}]{LL^*} W, \tag{3.48}$$

by Lemma 3.9 — by Theorem (3.8)(i), $\beta = \frac{1}{4} - \delta/2$

$$Y = \left\{ v : v \in L_2(0, T; \mathscr{D}(A^{1/4})), \frac{dv}{dt} \in L_2(0, T; H^{-1/2-\delta}(\Omega)) \right\}, \tag{3.49}$$

where in the second step we apply Lemma 3.9 with $B_0 = \mathscr{D}(A^{1/4})$, $B = H^{1/2-\delta}(\Omega)$, $B_1 = H^{-1/2-\delta}(\Omega)$, $p_0 = p_1 = 2$ and in the third step Theorem 3.8, (3.38b) with $\beta = \frac{1}{4} - \delta/2$. Diagram (3.48) proves (i).

(ii) Because of the compactness of part (i), it suffices to see that $\lambda = 1$ is not an eigenvalue of LL^*R as an operator $W \to W$, finite otherwise $\lambda = 1$ would be a fortiori an eigenvalue of LL^*R: $L_2(Q) \to L_2(Q)$, which we already know to be impossible (see injectivity below (2.8e)). □

THEOREM 3.11 (regularity of optimal control u^0 and optimal solution y^0). *Let the initial data be*

$$[y_0, y_1] \in H^{1/2-\delta}(\Omega) \times H^{-1/2-\delta}(\Omega) \equiv \mathscr{D}(A^{1/4-\delta/2}) \times [\mathscr{D}(A^{1/4+\delta/2})]' \tag{3.50}$$

and let R satisfy ((H.1) *and*) (H.2′) = (3.44). *Then*:

$$y^0 \in C([0, T]; H^{1/2-\delta}(\Omega)) \cap H^{1/2-\delta}(0, T; L_2(\Omega)), \tag{3.51 (i)}$$

$$\frac{dy^0}{dt} \in C([0, T]; H^{-1/2-\delta}(\Omega)) \tag{3.52}$$

and

$$u^0 \in H^{1/2,1/2}(\Sigma). \tag{3.53 (ii)}$$

Moreover, if R satisfies instead (H.2) = (3.43), *then*

$$u^0 \in H^{1/2+\delta,1/2+\delta}(\Sigma), \tag{3.54a (ii′)}$$

$$\textit{a fortiori } u^0 \in C([0, T]; L_2(\Gamma)). \tag{3.54b}$$

Proof. (i) By (2.8c) and (3.50), as a consequence of Theorem 3.10(ii), we have $y^0 \in W$ in (3.46) and by (3.47) $y^0 \in H^{1/2-\delta,1/2-\delta}(Q)$. From here, a bootstrap argument gives the full statement (3.51)-(3.52): we return to (2.8c), now rewritten as

$$y^0 = -LL^*Ry^0 + C(\cdot)y_0 + S(\cdot)y_1 \tag{3.55}$$

where, a fortiori, $Ry^0 \in L_2(0, T; H^{1/2-\delta}(\Omega))$. Thus, we can apply Theorem 3.8(i) [(3.38b), with $\beta = \frac{1}{4} - \delta/2$], and (ii), to obtain the regularity of LL^*Ry^0, while (3.50) (and (1.10)) determines the regularity of $C(\cdot)y_0 + S(\cdot)y_1$. We obtain

$$\begin{aligned} &LL^*Ry^0,\ C(\cdot)y_0 + S(\cdot)y_1 \in C([0, T]; H^{1/2-\delta}(\Omega)), \\ &\frac{d}{dt}(LL^*Ry^0), \frac{d}{dt}[C(\cdot)y_0 + S(\cdot)y_1] \in C([0, T], H^{-1/2-\delta}(\Omega)), \end{aligned} \tag{3.56}$$

and (3.51)-(3.52) now follow from (3.56) via (3.55).

(ii) To prove (3.53-(3.54), we use $u^0 = -L^*Ry^0$, (2.7), where, a fortiori from part (i),

$$Ry^0 \in \begin{cases} C([0, T]; \mathscr{D}(A^{1/4})) & \text{if } R \text{ satisfies (H.2′) = (3.44),} \\ C([0, T]; \mathscr{D}(A^{1/4+\delta/2})) & \text{if } R \text{ satisfies (H.2) = (3.43).} \end{cases} \tag{3.57}$$

The desired conclusions on u^0 now follow from (3.57), by invoking Theorem 3.5, (3.23a). □

3.3. Further regularity properties of $\Phi(t, s)$. In Lemma 2.1 we collected some preliminary properties of the evolution operator $\Phi(t, s)$ on E, which were valid for any penalization operator R satisfying only the general assumption (H.1). In § 3.2, we introduced some "minimal" regularity assumptions (H.2) or (H.2′) on R and deduced, in Theorem 3.11, corresponding regularity properties of y^0. These, via the definition (2.13), translate into analogous properties for $\Phi(t, s)$, with s fixed. The goal of the present subsection, which culminates in Theorem 3.16 below, is to deduce further regularity properties for $\Phi(t, s)$ in the case where R satisfies also assumption (H.2) = (3.43) or (H.2′) = (3.44): these will be *uniformly* on s and either on a space smoother than E or else on the space of distributions $H^{-1/2-\delta}(\Omega)$ for the first coordinate $\Phi_1(t, s)$. These properties will be needed in the derivation of the Riccati equation. To accomplish

our aim, we see through (2.14) that we must therefore study properties of the family of operators $I_s + L_s L_s^* R \equiv I_s + L_s L^* R$ (see (2.15)), $s \in [0, T]$. First, we shall extend the statement of Theorem 3.11(ii), uniformly in s. This is accomplished via the following lemma, which we give in a generality larger than our needs in the sequel will call for.

LEMMA 3.12. *Given a Banach space Z, $\|\cdot\|$, assume that:*

(a) *$\{K_t\}$ is a family of compact operators on Z, for each $t \in [0, T]$, $T < \infty$;*

(b) *the map $t \to K_t x$ is continuous for each $x \in Z$;*

(c) *for each fixed $t \in [0, T]$, the inverse $[I + K_t]^{-1}$ exists as a well-defined bounded operator on Z.*

Then, in fact

$$\|[I + K_t]^{-1}\|_{\mathscr{L}(Z)} \leqq C_T, \quad \text{uniformly in } t \in [0, T]. \tag{3.58}$$

Proof. From (a) and (b) it follows that the set

$$\mathscr{K} \equiv \bigcup_{0 \leqq t \leqq T} K_t \quad [\text{unit ball in } Z]$$

is precompact.[9] Next, by contradiction, let conclusion (3.58) be false so that there are sequences $\{t_n\}$, $\{x_n\}$, $t_n \in [0, T]$, $\|x_n\| = 1$ such that

$$[I + K_{t_n}]x_n \to 0 \quad \text{as } n \to \infty. \tag{3.59}$$

But plainly $\{K_{t_n} x_n\} \in K$ and thus, there is a convergent subsequence

$$K_{t_{n_k}} x_{n_k} \to y \in \bar{\mathscr{K}}. \tag{3.60}$$

Then, by (3.59)-(3.60)

$$x_{n_k} = (I + K_{t_{n_k}})x_{n_k} - K_{t_{n_k}} x_{n_k} \to -y \tag{3.61}$$

so that $\|y\| = 1$. Also, by (b), the Principle of Uniform Boundedness gives $\|K_t\|_{\mathscr{L}(Z)} \leqq C_T$, for all $t \in [0, T]$. Thus, by (3.61), $K_{t_{n_k}}(x_{n_k} + y) \to 0$. Also, at the price of extracting a further subsequence (denoted by the same symbol), we have $t_{n_k} \to t_0 \in [0, T]$. These last two conclusions, along with (b) give

$$K_{t_{n_k}} x_{n_k} = K_{t_{n_k}}(x_{n_k} + y) - K_{t_{n_k}} y \to -K_{t_0} y. \tag{3.62}$$

From (3.59), using (3.61) and (3.62), we then obtain

$$-(y + K_{t_0} y) = 0, \qquad \|y\| = 1$$

which contradicts assumption (c) at $t = t_0$. □

Our first application of Lemma 3.12 is a version of Theorem 3.10(ii), uniform in s.

THEOREM 3.13. *Let the operator R in* (H.1) *satisfy* (H.2′) = (3.44). *Then, with reference to the space W in* (3.46), *we have*[10]

$$\|I_s + L_s L_s^* R]^{-1}\|_{\mathscr{L}(W)} \leqq C_T \quad \text{uniformly in } 0 \leqq s \leqq T. \tag{3.63}$$

Proof. We shall apply Lemma 3.12 with $Z = W$; with $K_s = L_s L_s^* R = L_s L^* R$ (see (2.15)) compact on W and with $I_s + L_s L_s^* R$ boundedly invertible on W for each fixed s, as guaranteed by Theorem 3.10. To verify the remaining assumption (b) of Lemma

[9] We omit the details of this topological result: it is closely related to the concept of "a collectively compact family of operators" as in P. M. Anselone's *Collectively Compact Operator Approximation Theory*, Prentice-Hall, Englewood Cliffs, NJ, 1971. Our proof here is a modification of that of Thm. 1.6 in this reference.

[10] We embed $L_2(s, T; \cdot)$ into $L_2(0, T; \cdot)$ by extension by zero on $[0, s)$ uniformly in s, where, in the last step, a subset of diagram (3.66) was used.

3.12 that the map: $s \to L_s L^* Rf$ continuous in $W, f \in W$, it suffices to show that

$$f \in W \to L_s L^* Rf \underset{(\text{in } s)}{\in} L_\infty(0, T; W) \tag{3.64}$$

since then (3.64) combined with a standard approximating argument (to be made more precise at the end of the proof)

$$\begin{matrix} f_n \text{ smooth} \in W \\ f_n \to f \text{ in } W \end{matrix} \quad \text{while } s \to L_s L^* Rf_n \text{ continuous in } W, \quad \text{for each } n \tag{3.65}$$

will yield the required continuity in s of $L_s L^* Rf$. Let $Q_s = (s, T] \times \Omega$, and for $g \in L_2(s, T; \cdot)$, let g_{ext} be its extension by zero on $0 \leqq t \leqq s$, and for simplicity $H^r(\cdot) \equiv H^{r,r}(\cdot)$ with $\cdot$ either Q or Σ. We complement diagram (3.48) by

$$\begin{aligned} f \in W &\xrightarrow[(3.44)]{} Rf \in L_2(0, T; \mathcal{D}(A^{1/2})) \\ &\xrightarrow[(3.23a)]{} L^* Rf \in H^{1/2-\delta}(\Sigma) \to LL^* Rf \in H^{1/2-\delta}(Q), \end{aligned} \tag{3.66}$$

continuously at each step, where Corollary 3.7(ii), (3.36) was used in the last step. With $f \in W$, using diagram (3.66)

$$\|L_s L^* Rf\|_{H^{1/2-\delta}(Q_s)} = \|L[L^* Rf]_{\text{ext}}\|_{H^{1/2-\delta}(Q)} \leqq C_T \|L^* Rf\|_{H^{1/2-\delta}(\Sigma)} \leqq C_T \|f\|_W \tag{3.67}$$

uniformly in $0 \leqq s \leqq T$. Similarly, using $(d/dt)(L_s L^* Rf) = (dL_s/dt) L^* Rf$, [(3.42)], we have for $f \in W$, and invoking Corollary 3.7(iii) ((3.37)) with $\theta = \frac{1}{2} - \delta$:

$$\begin{aligned} \left\| \frac{d}{dt} L_s L^* Rf \right\|_{L_2(s,T;H^{-1/2-\delta}(\Omega))} &= \left\| \frac{dL}{dt} [L^* Rf]_{\text{ext}} \right\|_{L_2(0,T;H^{-1/2-\delta}(\Omega))} \\ &\leqq C_T \|L^* Rf\|_{H^{1/2-\delta}(0,T;L_2(\Gamma))} \leqq C_T \|f\|_W \end{aligned} \tag{3.68}$$

uniformly in s, where in the last step, a subset of diagram (3.66) was used. Conclusions (3.67)–(3.68) combined are stronger than the desired property (3.64). To make (3.65) more precise, note first that if, say, $g \in C([0, T]; L_2(\Gamma))$ with $\dot{g} \in L_2(\Sigma)$, then the same steps carried out in (3.30) give, after integration by parts

$$\begin{aligned} (L_s g)(t) &= A \int_s^t S(t-\tau) Dg(\tau)\, d\tau \\ &= Dg(t) - C(t-s) Dg(s) - \int_s^t C(t-\tau) D\dot{g}(\tau)\, d\tau \\ &\underset{(\text{in } s)}{\in} C([0, T]; H^{1/2-2\varepsilon}(\Omega)). \end{aligned}$$

Differentiating in t

$$\begin{aligned} \frac{d(L_s g)(t)}{dt} &= AS(t-s) Dg(s) + A \int_s^t S(t-\tau) D\dot{g}(\tau)\, d\tau \\ &\underset{(\text{in } s)}{\in} C([0, T]; H^{-1/2-2\varepsilon}(\Omega)), \qquad \varepsilon > 0 \end{aligned}$$

by (1.9b–c); i.e. taking $2\varepsilon = \delta$, we conclude that $s \to L_s g$ continuous in W. Next, with $f \in W$, choose $f_n \in W, f_n \to f$ in W with, say, $f_n \in L_2(0, T; \mathcal{D}(A))$. Then from (3.23a), we obtain $L^* Rf_n \in H^{2,2}(\Sigma)$ and the above argument with $g = L^* Rf_n$ works a fortiori, and (3.65) holds. The proof of Theorem 3.13 is complete. □

We need two more results, before collecting the properties of the evolution operator $\Phi(t, s)$ crucial in the sequel. They are a counterpart of Theorem 3.10 and, respectively, of its uniform version, Theorem 3.13 this time, however, on the space of distributions $H^{-1/2-\delta}(\Omega) = [H_0^{1/2+\delta}(\Omega)]' = [\mathscr{D}(A^{1/4+\delta/2})]'$, and for R satisfying (H.2).

THEOREM 3.14. *Let the operator R in* (H.1) *satisfy* (H.2) = (3.43). *Then*

(i) *The operator LL^*R is compact: $L_2(0, T; H^{-1/2-\delta}(\Omega)) \to$ itself.*

(ii) *The operator $[I + LL^*R]$ is invertible on $L_2(0, T; H^{-1/2-\delta}(\Omega))$, with bounded inverse here.*

Proof. For notational convenience, we set $\beta = \frac{1}{4} + \delta/2$ in this proof. The operator LL^*R on $L_2(0, T; H^{-1/2-\delta}(\Omega))$ is equivalent to the operator

$$A^{*-\beta} LL^* R A^{*\beta} \quad \text{on } L_2(Q). \tag{3.69}$$

(i) Compactness of the latter in (3.69) is equivalent to compactness of its adjoint, i.e. of

$$A^{\beta} R L L^* A^{-\beta} \quad \text{on } L_2(Q) \tag{3.70}$$

(as LL^* is self-adjoint on $L_2(Q)$ and R is self-adjoint on $L_2(\Omega)$). Now for $2\alpha = \frac{1}{2} - \delta < 2\beta = \frac{1}{2} + \delta$, we have the diagram

$$\begin{array}{ccccccc} L_2(Q) & \xrightarrow[\substack{\text{continuous}\\ \text{by (3.23b)}}]{L^*A^{-\beta}} & H^{2\beta,2\beta}(\Sigma) & \xrightarrow[\text{injection}]{\text{compact}} & H^{2\alpha,2\alpha}(\Sigma) & \xrightarrow[(3.36)]{L} & H^{2\alpha,2\alpha}(Q) \\ & & & & & & \Big\downarrow{\scriptstyle R\ (3.43)} \\ L_2(Q) \equiv L_2(0, T; L_2(\Omega)) & & \xleftarrow[\text{continuous}]{A^{\beta}} & & L_2(0, T; H_0^{1/2+\delta}(\Omega)) \equiv \mathscr{D}(A^{\beta})). & & \end{array} \tag{3.71}$$

From here, we deduce that the operator in (3.70) is compact, and hence so is the operator on (3.69).

(ii) Because of the compactness of part (i), it is enough to show that $\lambda = 1$ is not an eigenvalue of LL^*R on $L_2(0, T; H^{-1/2-\delta}(\Omega))$ equivalently, by (3.70), of RLL^* on $L_2(0, T; \mathscr{D}(A^{\beta}))$. Indeed, the latter statement is true, for otherwise $\lambda = 1$ would be a fortiori an eigenvalue of RLL^* on $L_2(Q)$, i.e. $x + RLL^*x = 0$, impossible, as it is seen by taking $L_2(Q)$-inner product on the right with LL^*x. □

The uniform version (in s) of Theorem 3.14(ii), in the style of Theorem 3.13 is now the following.

THEOREM 3.15. *Let the operator R in* (H.1) *satisfy* (H.2) = (3.43). *Then* (*see footnote* 10):

$$\|I_s + L_s L_s^* R]^{-1}\|_{\mathscr{L}(L_2(0, T; H^{-1/2-\delta}(\Omega)))} \leqq C_T, \quad \textit{uniformly in } s \in [0, T].$$

Proof. As in the proof of Theorem 3.13, we shall apply Lemma 3.12, this time with $Z = L_2(0, T; H^{-1/2-\delta}(\Omega))$, $K_s = L_s L_s^* R = L_s L^* R$, [(2.15)], compact on Z and $I_s + L_s L^* R$ invertible on Z for each fixed s, as guaranteed by Theorem 3.14. To verify the remaining assumption (b) of Lemma 3.12 that the map

$$s \to L_s L^* R g \text{ is continuous in } Z, \text{ for each } g \in Z \tag{3.72}$$

it will suffice to show, as is the proof of Theorem 3.13, that $L_s L^* R g \in L_\infty(0, T; Z)$ in s, indeed that

$$\|L_s L^* R g\|_Z \leqq C_T \|g\|_Z \quad \text{uniformly in } s \in [0, T], \tag{3.73}$$

equivalent to (see the paragraph containing (3.69) and (3.70))

$$\|A^{\beta} R L_s L^* A^{-\beta} v\|_{L_2(s, T; L_2(\Omega))} \leqq C_T \|v\|_{L_2(Q)} \quad \text{uniformly in } s \in [0, T] \tag{3.74}$$

with $\beta = \frac{1}{4} + \delta/2$. But (3.74) holds true: in fact, in the notation of g_{ext} used above (3.66), we have by diagram (3.71) ($2\alpha = \frac{1}{2} - \delta$)

$$\begin{aligned} \|A^{\beta} R L_s L^* A^{-\beta} v\|_{L_2(s,T;L_2(\Omega))} &= \|A^{\beta} R L [L^* A^{-\beta} v]_{\text{ext}}\|_{L_2(Q)} \\ &\leqq C_T \|L^* A^{-\beta} v\|_{H^{2\alpha,2\alpha}(\Sigma)} \leqq C_T \|v\|_{L_2(Q)} \end{aligned} \tag{3.75}$$

as desired. □

Remark 3.2. We sketch another proof of conclusion (3.72), i.e. of the statement that with $\beta = \frac{1}{4} + \delta/2$

$$\|[I_s + A^{\beta} R L_s L_s^* A^{-\beta}]^{-1}\|_{(L_2(0,T;L_2(\Gamma))} \leqq C_T \quad \text{uniformly in } s \in [0, T].$$

First, using the full strength of (3.36) in the diagram (corresponding to the initial time s) in the proof of Theorem 3.14, we have that the operator

(*) $G_s \equiv A^{\beta} R L_s L_s^* A^{-\beta}$: continuous $L_2(s, T; L_2(\Gamma)) \to C([s, T]; L_2(\Omega))$ with bound independent of s

due to the continuity in s. Next, we consider the interval $s \leqq t \leqq T$, where $T - s < h$, for some $h > 0$, and prove that $[I_s + G_s]^{-1}$: continuous $L_2(s, T; L_2(\Gamma)) \to$ itself with a bound *independent* of h. Indeed

$$\begin{aligned} \|G_s x\|^2_{L_2(s,T;L_2(\Omega))} &= \int_s^T |G_s x(\cdot, s)(t)|^2_{L_2(\Omega)}\, dt \\ &\leqq \{\sup_{T-h \leqq s \leqq t \leqq T} |G_s x(\cdot, s)(t)|^2_{L_2(\Omega)}\}(T - s) \\ &\leqq C_T h \int_s^T |x(t, s)|^2_{L_2(\Omega)}\, dt \qquad \text{(by (*))} \end{aligned}$$

where C_T is independent of h by (*). We then choose h so that, say, $C_T h < \frac{1}{2}$ and repeat the procedure. After a finite number of steps we obtain the desired conclusion (3.72). □

The following is the main result of the present subsection and complements Lemma 2.1. It is obtained as a corollary of the result given above, via (2.13)–(2.14). First, let Y_r denote for convenience the space of "regular" initial data, as in (3.50):

$$Y_r \equiv H^{1/2-\delta}(\Omega) \times H^{-1/2-\delta}(\Omega) = \mathcal{D}(A^{1/4-\delta/2}) \times [\mathcal{D}(A^{1/4+\delta/2})]'. \tag{3.76}$$

THEOREM 3.16. (i) *Let the operator R satisfy* ((H.1) *and*) (H.2)′ = (3.44). *Then*:

(a) *For each s fixed, $s < T$, the operator $\Phi(\cdot, s)$ is strongly continuous $Y_r \to C([s, T]; Y_r)$.*

(b) *Moreover, with W as in* (3.46)

$$\|\Phi(\cdot, s)\|_{\mathcal{L}(Y_r \to W)} \leqq C_T \quad \textit{uniformly in } s \in [0, T] \quad \text{i.e.,} \tag{3.77a}$$

$$\|\Phi_1(\cdot, s) y\|^2_{L_2(s,T;H^{1/2-\delta}(\Omega))} + \|\Phi_2(\cdot, s) y\|^2_{L_2(s,T;H^{-1/2-\delta}(\Omega))} \leqq C_T \|y\|^2_{Y_r}. \tag{3.77b}$$

(ii) *If R satisfies* (H.2) = (3.43), *then*

$$\|A^{1/4+\delta/2} R \Phi_1(\cdot, s) y\|_{L_2(s,T;L_2(\Omega))} \leqq C_T \|y\|_{Y_r} \quad \textit{uniformly in } s \in [0, T]. \tag{3.78 (a)}$$

(b) *for $y = [y_0, y_1] \in [\mathcal{D}(A^{1/4+\delta/2})]' \times [\mathcal{D}(A^{5/4+\delta/2})]'$, whose norm we indicate by $|||\cdot|||$, then*

$$\|\Phi_1(\cdot, s) y\|_{L_2(s,T;H^{-1/2-\delta}(\Omega))} \leqq C_T |||y||| \quad \textit{uniformly in } s \in [0, T]. \tag{3.79}$$

(*We shall use* (3.79) *with much smoother y, however.*)

Proof. Property (ia) is a restatement of Theorem 3.11(i), via (2.13). Property (ib) follows from Theorem 3.13 [(3.63)], via (2.14) (and (1.10)). Property (iia) follows a fortiori from (3.77b), with $A^{1/4+\delta/2}R$ continuous $H^{1/2-\delta}(\Omega) \to L_2(\Omega)$. Property (iib) follows a fortiori from Theorem 3.15 [(3.72)], via (2.14) (and (1.10)). □

4. Further properties of $\mathscr{P}(t)$ and derivation of a Riccati differential equation.

4.1. A fundamental result for $\mathscr{B}^*\mathscr{P}(t)$. We now proceed to show that the operator $\mathscr{P}(t)$ satisfies the Riccati differential equation, as claimed in part (ii) of Theorem 1.3. To this end, it is necessary, as a preliminary step, to give a meaning to the operator $\mathscr{B}^*\mathscr{P}(t)$ as a function of t, which appears there in the quadratic term. This point is the major *difficulty* that one encounters in deriving the Riccati equation (whether differential or integral) for the problem under study. A fundamental role in this direction is played by the following theorem.

THEOREM 4.1. *Let R satisfy* ((H.1) *and*) (H.2) = (3.43) *and recall the space Y_r of "regular" initial data from* (3.76). *Then, for the operator $\mathscr{B}^*\mathscr{P}(t)$ defined by* (2.26), *the following property in t holds*:

(4.1) *$\mathscr{B}^*\mathscr{P}(t)$ is a bounded linear operator*

$$Y_r \to C([0, T]; L_2(\Gamma)), \quad \text{i.e.} \quad \max_{0 \leq t \leq T} \|\mathscr{B}^*\mathscr{P}(t)x\|_\Gamma \leq \text{const}_T \|x\|_{Y_r}, \qquad x \in Y_r.$$

Remark 4.1. Most of the effort required to prove Theorem 4.1 has already been made at the level of proving certain properties of Φ (Theorem 3.16). Observe first that the weaker statement

$$\mathscr{B}^*\mathscr{P}(t) \text{ continuous operator } Y_r \to L_\infty(0, T; L_2(\Gamma)) \tag{4.2}$$

can be quickly obtained from Theorem 3.16: with x and R as assumed, we can write (2.26) as

$$\mathscr{B}^*\mathscr{P}(t)x = D^*A^{*1/4-\delta/2} \int_t^T A^{*1/2}S^*(\tau - t)A^{*1/4+\delta/2}R\Phi_1(\tau, t)x \, d\tau \tag{4.3}$$

from which, invoking (1.9b), (1.10) we obtain

$$\begin{aligned} \|\mathscr{B}^*\mathscr{P}(t)x\|_{L_2(\Gamma)} &\leq C_T \int_t^T \|A^{*1/4+\delta/2}R\Phi_1(\tau, t)x\|_{L_2(\Omega)} \, d\tau \\ &\leq C_T \int_t^T \|A^{*1/4+\delta/2}R\Phi_1(\tau, t)x\|^2_{L_2(\Omega)} \, d\tau \\ &\leq C_T \|x\|_{Y_r} \quad \text{uniformly in } t \in [0, T] \end{aligned} \tag{4.4}$$

where in the last step we have used Theorem 3.16, statement (iia), ((3.78)). Thus, (4.2) is proved. □

Proof of Theorem 4.1. *Right continuity.* Let $t_1 \in [0, T)$ and let $t > t_1$. From (2.26), we compute after a change of variable

$$\begin{aligned} \mathscr{B}^*\mathscr{P}(t)x - \mathscr{B}^*\mathscr{P}(t_1)x &= D^*A^* \int_0^{T-t} S^*(\sigma)R\Phi_1(t+\sigma, t)x \, d\sigma \\ &\quad - D^*A^* \int_0^{T-t_1} S^*(\sigma)R\Phi_1(t_1+\sigma, t_1)x \, d\sigma = I_1(t) - I_2(t), \end{aligned} \tag{4.5}$$

$$
\begin{aligned}
I_1(t) &= D^*A^* \int_0^{T-t} S^*(\sigma)R[\Phi_1(t+\sigma, t)x - \Phi_1(t_1+\sigma, t_1)x]\, d\sigma, \\
I_2(t) &= D^*A^{*1/4-\delta/2} \int_{T-t}^{T-t_1} A^{*1/2}S^*(\sigma)A^{*1/4+\delta/2}R\Phi_1(t_1+\sigma, t_1)x\, d\sigma.
\end{aligned} \tag{4.6}
$$

As to $I_2(t)$, proceeding as in going from (4.3) to (4.4), we have that its integrand is in $L_2(t_1, T; L_2(\Omega))$. Thus, as $t \downarrow t_1$, $I_2(t) \to \theta$. As to $I_1(t)$, in a similar fashion we obtain

$$
\|I_1(t)\|^2_{L_2(\Gamma)} \leqq C_T \int_0^{T-t} \|A^{*1/4+\delta/2}R[\Phi_1(t+\sigma, t)x - \Phi_1(t_1+\sigma, t_1)x]\|^2_{L_2(\Omega)}\, d\sigma
$$

(adding and subtracting $\Phi_1(t+\sigma, t_1)x$)

$$
\leqq C_T\left\{\int_0^{T-t} A(t, \sigma) + B(t, \sigma)\, d\sigma\right\}, \tag{4.7}
$$

$$
A(t, \sigma) \equiv \|A^{*1/4+\delta/2}R[\Phi_1(t+\sigma, t)x - \Phi_1(t+\sigma, t_1)x]\|^2_{L_2(\Omega)}, \tag{4.8a}
$$

$$
B(t, \sigma) = \|A^{*1/4+\delta/2}R[\Phi_1(t+\sigma, t_1)x - \Phi_1(t_1+\sigma, t_1)x]\|^2_{L_2(\Omega)}. \tag{4.8b}
$$

Since $t > t_1$, as below (2.21) we have $\Phi_1(t+\sigma, t_1)x = \Phi_1(t+\sigma, t)\Phi(t, t_1)x$ and

$$
A(t, \sigma) \leqq \|A^{*1/4+\delta/2}R\Phi_1(t+\sigma, t)[x - \Phi(t, t_1)x]\|_{L_2(\Omega)}. \tag{4.9}
$$

Returning to the integral of $A(t, \sigma)$, we use there the Schwarz inequality, (4.9), and Theorem 3.16(iia), [(3.78)]. We obtain

$$
\begin{aligned}
\int_0^{T-t} A(t, \sigma)\, d\sigma &\leqq C_T \int_0^{T-t} \|A^{*1/4+\delta/2}R\Phi_1(t+\sigma, t)[x - \Phi(t, t_1)x]\|^2_{L_2(\Omega)}\, d\sigma \\
&\leqq C_T\|x - \Phi(t, t_1)x\|^2_{Y_r} \to 0 \quad \text{as } t \downarrow t_1
\end{aligned} \tag{4.10}
$$

as desired, with the right-hand side going to zero, by Theorem 3.16(ia). As to $B(t, \sigma)$, the same Theorem 3.16(ia) implies that for $|t - t_1| = |(t+\sigma) - (t_1+\sigma)|$ sufficiently small, and all $\sigma \in [0, T]$, the difference in square brackets in (4.8b) is arbitrarily small in the norm of $H^{1/2-\delta}(\Omega)$. But $A^{*1/4+\delta/2}R$: continuous $H^{1/2-\delta}(\Omega) \to L_2(\Omega)$. Hence,

$$
\int_0^{T-t} B(t, \sigma)\, d\sigma \to 0 \quad \text{as } t \downarrow t_1. \tag{4.11}
$$

Returning to (4.7) with (4.10) and (4.11), we conclude $\|I_1(t)\|_{L_2(\Gamma)} \to 0$ as well as $t \downarrow t_1$.

Left continuity. We interchange the role of t and t_1 in the above proof, i.e. we keep t fixed and let t_1 run $\uparrow t$. The above proves $\mathscr{B}^*\mathscr{P}(t)x \in C([0, T]; L_2(\Gamma))$, $x \in Y_r$ and by (4.4), Theorem 4.1 follows.

Remark 4.2. From the pointwise relation (2.28), we obtain again, via Theorem 4.1 and Theorem 3.11(i), ((3.51)), that for $y \in Y_r$, the corresponding optimal control $u^0(t) = u^0(t, 0; y) \in C([0, T]; L_2(\Gamma))$, a regularity result contained a fortiori in (3.54a).

4.2. Derivation of the Riccati differential equation for $\mathscr{P}(t)$. With the help of the fundamental Theorem 4.1, we can now proceed to derive the Riccati differential equation. In fact, we can now assert that the quadratic term $(\mathscr{B}^*\mathscr{P}(t)x,\ \mathscr{B}^*\mathscr{P}(t)y)_\Gamma$ in the R.D.E. is in $C[0, T]$ for $x, y \in Y_r$, [(3.76)].

LEMMA 4.2. *Let R satisfy* ((H.1) *and*) (H.2) = (3.43). *Let $x, y \in Y_r$*, [(3.76)]. *Then, both E-inner products $(\mathscr{P}(t)x, \mathscr{A}y)_E$ and $(\mathscr{P}(t)\mathscr{A}x, y)_E$ are well-defined at each t and in $C[0, T]$.*

Proof. Because of self-adjointedness of $\mathscr{P}(t)$ (Lemma 2.3), it is enough to consider the first inner product. For fixed t, we use (2.24) with $R\Phi_1(\tau, t)x \in C([t, T]; \mathscr{D}(A^{1/4+\delta/2}))$ for $x \in Y_r$ along with (1.5). Details are omitted (see also Lemma 2.3(i)). □

THEOREM 4.3. *Let R satisfy* ((H.1) *and*) (H.2) = (3.43). *Then the operator $\mathscr{P}(t)$ defined by* (2.22a) *or* (2.24) *satisfies the following Riccati differential equation*

$$\frac{d}{dt}(\mathscr{P}(t)x, y)_E = -(x_1, y_1)_\Omega - (\mathscr{P}(t)\mathscr{A}x, y)_E - (\mathscr{P}(t)x, \mathscr{A}y)_E + (\mathscr{B}^*\mathscr{P}(t)x, \mathscr{B}^*\mathscr{P}(t)y)_\Gamma$$

for all $x, y \in Y_r = H^{1/2-\delta}(\Omega) \times H^{-1/2-\delta}(\Omega)$, $t \in [0, T)$ with terminal condition $\mathscr{P}(T) = 0$.

Proof. Step 1. From (2.22a),

$$(\mathscr{P}(t)x, y)_E = \int_t^T \left(\begin{vmatrix} R\Phi_1(\tau, t)x \\ 0 \end{vmatrix}, e^{\mathscr{A}(\tau - t)}y \right)_E d\tau$$

we compute with $x, y \in Y_r$ and $\Phi_1(t, t)x = [\Phi(t, t)x]_1 = x_1$; using Lemma 4.2

$$\begin{aligned} \frac{d}{dt}(\mathscr{P}(t)x, y)_E &= -\left(\begin{vmatrix} x_1 \\ 0 \end{vmatrix}, y \right)_E \\ &\quad + \int_t^T \left(\left| \frac{\partial \Phi_1(\tau, t)x}{\partial t} \right|, R[e^{\mathscr{A}(\tau - t)}y]_1 \right)_\Omega d\tau - (\mathscr{P}(t)x, \mathscr{A}y)_E \end{aligned} \tag{4.12}$$

where we have to justify and evaluate the second term on the right-hand side of (4.12). This will be done below.

We first recall the operator $\mathscr{B}^*: E \supset \mathscr{D}(\mathscr{B}^*) \to L_2(\Gamma)$ from (2.25):

$$\mathscr{B}^*v = D^*A^{*1/2}A^{-1/2}v_2, \qquad v = [v_1, v_2] \in \mathscr{D}(\mathscr{B}^*)$$

where $\mathscr{D}(\mathscr{B}^*) \supset L_2(\Omega) \otimes H^{-1/2}(\Omega)$. For $v \in \mathscr{D}(\mathscr{B}^*)$ and $u \in L_2(\Gamma)$, the inner product $(u, \mathscr{B}^*v)_\Gamma$ is well-defined on $L_2(\Gamma)$ and we then introduce an operator $\mathscr{B}$, through the duality pairing defined on E:

$$(\mathscr{B}u, v)_E = (u, \mathscr{B}^*v)_\Gamma, \qquad u \in L_2(\Gamma), \quad v \in \mathscr{D}(\mathscr{B}^*). \tag{4.13}$$

Thus, for $x \in Y_r$ and $y = [y_1, y_2] \in L_2(\Omega) \otimes L_2(\Omega) \subset \mathscr{D}(\mathscr{B}^*)$ we compute the following well-defined expression (see Theorem 4.1) via (1.7)

$$\begin{aligned} (\mathscr{B}^*\mathscr{P}(t)x, \mathscr{B}^*y)_\Gamma &= (\mathscr{B}\mathscr{B}^*\mathscr{P}(t)x, y)_E \\ &= ([\mathscr{B}\mathscr{B}^*\mathscr{P}(t)x]_1, y_1)_\Omega + (A^{-1/2}[\mathscr{B}\mathscr{B}^*\mathscr{P}(t)x]_2, A^{-1/2}y_2)_\Omega \end{aligned} \tag{4.14a}$$

in $C[0, T]$ and on the other hand via (2.25) we have

$$\begin{aligned} (\mathscr{B}^*\mathscr{P}(t)x, \mathscr{B}^*y)_\Gamma &= (\mathscr{B}^*\mathscr{P}(t)x, D^*A^{*1/2}A^{-1/2}y_2)_\Gamma \\ &= (A^{1/2}D\mathscr{B}^*\mathscr{P}(t)x, A^{-1/2}y_2)_\Omega. \end{aligned} \tag{4.14b}$$

We compare (4.14a) and (4.14b), with $y_1 \in L_2(\Omega)$ and $A^{-1/2}y_2$ which exhausts all of $\mathscr{D}(A^{1/2}) = H_0^1(\Omega)$, as y_2 runs over $L_2(\Omega)$. We thus deduce

$$[\mathscr{B}\mathscr{B}^*\mathscr{P}(t)x]_1 = 0, \tag{4.15a}$$

$$A^{-1/2}[\mathscr{B}\mathscr{B}^*\mathscr{P}(t)x]_2 = A^{1/2}D\mathscr{B}^*\mathscr{P}(t)x \in H^{-1}(\Omega), \qquad x \in E. \tag{4.15b}$$

Step 2. We collect a few facts.

LEMMA 4.4. (1) *For $x \in Y_r \equiv H^{1/2-\delta}(\Omega) \times H^{-1/2-\delta}(\Omega) = \mathscr{D}(A^{1/4-\delta/2}) \times [\mathscr{D}(A^{1/4+\delta/2})]'$ and for each t fixed, we have*

$$\text{(a)} \quad S(\tau - t)AD\mathscr{B}^*\mathscr{P}(t)x \underset{(\text{in } \tau)}{\in} C([t, T]; H^{-1/2-\delta}(\Omega)) \tag{4.16}$$

where A can indifferently be moved in front of $S(\cdot)$

$$(4.17)\quad (b)\qquad [I_t + L_t L_t^* R]^{-1}\{S(\cdot - t)AD\mathscr{B}^*\mathscr{P}(t)x\}(\tau) = \Phi_1(\tau, t)\mathscr{B}\mathscr{B}^*\mathscr{P}(t)x \in L_2(t, T; H^{-1/2-\delta}(\Omega)).$$

(2) *In addition, let* $y \in Y_r$ *as well, and let R satisfy* (H.2) = (3.43). *Then*

(a) *the following duality pairing on* $H^{-1/2-\delta}(\Omega) \times H_0^{1/2+\delta}(\Omega)$ *is well-defined in the sense that*

$$(4.18)\qquad ([I_t + L_t L_t^* R]^{-1}\{S(\cdot - t)AD\mathscr{B}^*\mathscr{P}(t)x\},\ R[e^{\mathscr{A}(\tau - t)}y]_1)_\Omega \in L_2(t, T),$$

(b) *the following expression is well-defined in the sense that*

$$(4.19)\qquad (\mathscr{B}^*\mathscr{P}(t)x,\ \mathscr{B}^*\mathscr{P}(t)y)_\Gamma = \int_t^T \Big([I_t + L_t L_t^* R]^{-1} S(\cdot - t)AD\mathscr{B}^*\mathscr{P}(t)x,\ R[e^{\mathscr{A}(\tau - t)}y]_1\Big)_\Omega d\tau \in C[0, T].$$

Proof of Lemma 4.4. (1) Use Theorem 4.1 and (1.10) to obtain (4.16) (where $A^{1/4+\delta/2}$ is viewed as the isomorphic extension $L_2(\Omega) \to [\mathscr{D}(A^{1/4+\delta/2})]'$). Then, (4.17) follows from (4.16), via Theorem 3.14(ii). (2) In turn, (4.18) follows from (4.17), since $R[e^{\mathscr{A}\cdot}y]_1 \in C([t, T]; H_0^{1/2+\delta}(\Omega))$. (2b) By selfadjointness of $\mathscr{P}(t)$ (Lemma 2.3(ii.1)) and by (2.22a), we compute for $x, y \in Y_r$ the following expression, where, by Theorem 4.1, the left-hand side, and thus the right-hand side also, is well-defined and in $C([0, T]$:

$$(\mathscr{B}^*\mathscr{P}(t)x,\ \mathscr{B}^*\mathscr{P}(t)y)_\Gamma = (\mathscr{P}(t)\mathscr{B}\mathscr{B}^*\mathscr{P}(t)x, y)_E = \int_t^T (\Phi_1(\tau, t)\mathscr{B}\mathscr{B}^*\mathscr{P}(t)x,\ R[e^{\mathscr{A}(\tau - t)}y]_1)_\Omega\, d\tau$$

[recalling from (2.9c) and (2.13),

$$\begin{aligned} y^0(\tau, t; x) = \Phi_1(\tau, t)x &= \{[I_t + L_t L_t^* R]^{-1}[C(\cdot - t)x_1 + S(\cdot - t)x_2]\}(\tau),\ x \in E] \\ &= \int_t^T (\{[I_t + L_t L_t^* R]^{-1} \times \{C(\cdot - t)[\mathscr{B}\mathscr{B}^*\mathscr{P}(t)x]_1 \\ &\qquad + S(\cdot - t)[\mathscr{B}\mathscr{B}^*\mathscr{P}(t)x]_2\}\}(\tau),\ R[e^{\mathscr{A}(\tau - t)}y]_1)_\Omega\, d\tau \\ &= \int_t^T (\text{zero} + [I_t + L_t L_t^* R]^{-1} S(\cdot - t)AD\mathscr{B}^*\mathscr{P}(t)x,\ R[e^{\mathscr{A}(\tau - t)}y]_1)_\Omega\, d\tau \end{aligned} \qquad (4.20)$$

where in the last step we have used (4.15), and (4.19) follows. The correctness of these expressions is also asserted in (4.18). □

Step 3. If $x \in Y_r$, then $A^{1/2}S(\cdot - t)A^{1/2}x_1 - C(\cdot - t)x_2 \in C([t, T]; H^{-1/2-\delta}(\Omega))$ and via (1.5) and (4.20) we obtain

$$(4.21)\qquad \{I_t + L_t L_t^* R]^{-1}[A^{1/2}S(\cdot - t)A^{1/2}x_1 - C(\cdot - t)x_2]\}(\tau) = -\Phi_1(\tau, t)\mathscr{A}x \in L_2(t, T; H^{-1/2-\delta}(\Omega))$$

by Theorem 3.14(ii). Thus, for $x, y \in Y_r$, (4.19), (4.21) yield the well-defined expression

in $C[0, T]$ via Lemma 4.2

$$\int_t^T ([I_t + L_t L_t^* R]^{-1}[S(\cdot - t)AD\mathcal{B}^*\mathcal{P}(t)x + A^{1/2}S(\cdot - t)A^{1/2}x_1 - C(\cdot - t)x_2], R[e^{\mathcal{A}(\tau - t)}y]_1)_\Omega \, d\tau$$

$$= (\mathcal{B}^*\mathcal{P}(t)x, \mathcal{B}^*\mathcal{P}(t)y)_\Gamma - \int_t^T \left(\begin{vmatrix} R\Phi_1(\tau, t)\mathcal{A}x \\ 0 \end{vmatrix}, e^{\mathcal{A}(\tau - t)}y \right)_E d\tau \tag{4.22}$$

$$= (\mathcal{B}^*\mathcal{P}(t)x, \mathcal{B}^*\mathcal{P}(t)y)_\Gamma - (\mathcal{P}(t)\mathcal{A}x, y)_E$$

where, in the last step, we have used the definition (2.22a) of $\mathcal{P}(t)$.

Step 4. We now return to (4.20), with $L_t L_t^* = L_t L^*$ ((2.15)), and $x \in Y_r$, rewritten as

$$\Phi_1(\cdot, t)x + L_t L_t^* R\Phi_1(\cdot, t)x = C(\cdot - t)x_1 + S(\cdot - t)x_2 \tag{4.23}$$

and compute first the term

$$\frac{\partial}{\partial t}\{L_t L^* R\Phi_1(\cdot, t)x\} = \frac{\partial}{\partial t}\int_t^\tau AS(\tau - \sigma)D[L^* R\Phi_1(\cdot, t)x](\sigma)\, d\sigma$$

$$= -AS(\tau - t)D[L^* R\Phi_1(\cdot, t)x](t) + \left\{ L_t L_t^* R \frac{\partial \Phi_1}{\partial t}(\cdot, t)x \right\}(\tau)$$

$$= -AS(\tau - t)D\mathcal{B}^*\mathcal{P}(t)x + \left\{ L_t L_t^* R \frac{\partial \Phi_1}{\partial t}(\cdot, t)x \right\}(\tau) \quad \text{(by (2.26))}.$$

Hence by differentiating (4.23) in t with $x \in Y_r$

$$\left\{ [I_t + L_t L_t^* R] \frac{\partial \Phi_1}{\partial t}(\cdot, t)x \right\}(\tau) = AS(\tau - t)D\mathcal{B}^*\mathcal{P}(t)x + A^{1/2}S(\tau - t)A^{1/2}x_1 - C(\tau - t)x_2$$

$$\in L_2(t, T; H^{-1/2-\delta}(\Omega)) \tag{4.24}$$

which, in τ, belongs for each t to $L_2(t, T; H^{-1/2-\delta}(\Omega))$ by Lemma 4.4, ((4.16)) so that

$$\frac{\partial \Phi_1}{\partial t}(\cdot, t)x = [I_t + L_t L_t^* R]^{-1}\{AS(\cdot - t)D\mathcal{B}^*\mathcal{P}(t)x + A^{1/2}S(\cdot - t)A^{1/2}x_1 - C(\cdot - t)x_2\}$$

$$\in L_2(t, T; H^{-1/2-\delta}(\Omega)), \qquad x \in Y_r \tag{4.25}$$

is well-defined and in $L_2(t, T; H^{-1/2-\delta}(\Omega)$, as guaranteed by Lemma 4.4, [(4.17)].

Step 5. By (4.25) and (4.22), with $x, y \in Y_r$

$$\int_t^T \left(\left\{ \frac{\partial \Phi_1}{\partial t}(\cdot, t)x \right\}(\tau), R[e^{\mathcal{A}(\tau - t)}y]_1 \right)_\Omega d\tau$$

$$= (\mathcal{B}^*\mathcal{P}(t)x, \mathcal{B}^*\mathcal{P}(t)y)_\Gamma - (\mathcal{P}(t)\mathcal{A}x, y)_E \tag{4.26}$$

which is well-defined as an $C[0, T]$-function, Theorem 4.1 and Lemma 4.2. But the left-hand side of (4.26) is, in fact, the second term in (4.12) which we sought to compute. Hence, inserting (4.26) into (4.12) gives, finally, the claimed Riccati differential equation of the statement. □

We explicitly single out a result, which is essentially already contained in (4.25). Using (4.17) plus (4.21), we can rewrite (4.25) as

$$\frac{\partial \Phi_1(\cdot, t)x}{\partial t} = -\Phi_1(\cdot, t)[\mathcal{A} - \mathcal{B}\mathcal{B}^*\mathcal{P}(t)]x \in L_2(t, T; H^{-1/2-\delta}(\Omega)), \qquad x \in Y_r$$

as a well-defined expression, which is in $L_2(t, T; H^{-1/2-\delta}(\Omega))$ for each t. A similar result can be obtained for Φ_2. Thus, we obtain the following.

PROPOSITION 4.5. *For $x \in Y_r$, we have*

$$\frac{\partial \Phi(\cdot, t)x}{\partial t} = -\Phi(\cdot, t)[\mathscr{A} - \mathscr{B}\mathscr{B}^*\mathscr{P}(t)]x$$

as a well-defined expression in $L_2(t, T; [\mathscr{D}(A^{1/4+\delta/2})]' \times [\mathscr{D}(A^{3/4+\delta/2})]')$ for each t.

As a corollary to Theorem 4.3, we obtain a Riccati integral equation for $\mathscr{P}(t)$ in a standard way.

COROLLARY 4.6. *The Riccati operator $\mathscr{P}(t)$ of Theorem 4.3 satisfies the following integral equation*:

$$(\mathscr{P}(t)x, y)_E = \int_t^T \left(\begin{bmatrix} R & 0 \\ 0 & 0 \end{bmatrix} e^{\mathscr{A}(\tau - t)}x, e^{\mathscr{A}(\tau - t)}y \right)_E d\tau + \int_t^T (\mathscr{B}^*\mathscr{P}(\tau)\, e^{\mathscr{A}(\tau - t)}x, \mathscr{B}^*\mathscr{P}(\tau)\, e^{\mathscr{A}(\tau - t)}y)_\Gamma \, d\tau \tag{R.I.E.}$$

for all $x, y \in Y_r$ and all $t \in [0, T]$.

Remark 4.3. Plainly, the solution (2.1a, b), (2.2a, b) of problem (1.1) can also be written as

$$\begin{vmatrix} y(t) \\ \dot{y}(t) \end{vmatrix} = e^{\mathscr{A}t} \begin{vmatrix} y_0 \\ y_1 \end{vmatrix} + (\mathscr{L}u)(t). \tag{4.27}$$

See (1.5), (2.23), where we have introduced the operator

$$(\mathscr{L}u)(t) = \begin{vmatrix} (Lu)(t) \\ \left(\dfrac{dLu}{dt}\right)(t) \end{vmatrix} = \mathscr{A} \int_0^t e^{\mathscr{A}(t-\tau)}\mathscr{A}^{-1}\mathscr{B}u(\tau)\, d\tau \tag{4.28}$$

$$: \text{continuous } L_2(0, T; L_2(\Gamma)) \to C([0, T]; E)$$

$\mathscr{B}$ as in (4.13) so that

$$\mathscr{A}^{-1}\mathscr{B}u = \begin{vmatrix} -Du \\ 0 \end{vmatrix}, \qquad \mathscr{A}^{-1}\mathscr{B} \in \mathscr{L}(L_2(\Gamma), E).$$

We shall find it convenient to use (4.27)-(4.28) in § 5 (below (5.3)), where, by convention, we shall write (4.28) more expediently as

$$(\mathscr{L}u)(t) = \int_0^t e^{\mathscr{A}(t-\tau)}\mathscr{B}u(\tau)\, d\tau.$$

See also [D-L-T.1], for a *direct* study of the Riccati integral equation for hyperbolic dynamics written in the form (4.27)-(4.28).

5. The quadratic regulator problem for (1.1). The present section is devoted to a quick study of the optimal quadratic control problem over an infinite time interval (quadratic regulator) for the hyperbolic dynamics (1.1). Our treatment, in particular, has the important advantage that it does *not* require the usual limit process—present in most, if not all, available literature, see e.g. [B.1]—of deriving the algebraic Riccati equation from the differential (or integral) Riccati equation on $[0, T]$, as $T \uparrow \infty$. This advantage pays off particularly in cases like the hyperbolic dynamics, where regularity

properties of $\mathscr{B}^*\mathscr{P}(\mathscr{P}$ = Riccati operator of problem on $[0, \infty])$ can be deduced directly with minimal assumptions on the observation operators, in particular, without passing through the limit of $\mathscr{B}^*\mathscr{P}_T(t)(\mathscr{P}_T(t)$ = Riccati operator of problem on $[0, T])$, whose very definition may require, by contrast, stronger assumptions on the observation operators. In line with problem (1.2a) on $[0, T]$, $T<\infty$ of the preceding sections, the problem that we study now is: minimize the cost

$$J_\infty(u, y(u)) = \int_0^\infty \{(Ry(t), y(t))_\Omega + |u(t)|^2_\Gamma\}\, dt \tag{5.1}$$

over all $u \in L_2(0, \infty; L_2(\Gamma)) \equiv L_2(\Sigma)$, with $y(u)$ solution of (1.1) corresponding to u. R satisfies *only* assumption (H.1) of nonnegative definiteness on $L_2(\Omega)$; in particular, R may be the identity. A minimal assumption which we must impose for the above problem to make sense is, as usual:

(H) For each initial data $[y_0, y_1] \in E$, there exists a $u \in L_2(\Sigma_\infty)$, whose corresponding solution gives Ry in $L_2(0, \infty; E)$, whereby the corresponding cost J_∞ is finite.

Remark 5.1. It is proved in [L-T.5] that, under some assumptions on the bounded domain Ω, the boundary feedback $u(t) = -D^*y_t(t)$, once inserted in (1.1) with $-A(\xi, \partial) = \Delta$, gives rise to a closed loop hyperbolic system with the following properties:

(i) The corresponding *closed loop feedback* dynamics $[y(t), y_t(t)]$ is a s.c. contraction semigroup $S_F(t)$ on $L_2(\Omega) \times H^{-1}(\Omega)$; i.e. it can be represented here by

$$\begin{vmatrix} y(t) \\ y_t(t) \end{vmatrix} = S_F(t) \begin{vmatrix} y_0 \\ y_1 \end{vmatrix}.$$

(ii) $u = -D^*y_t \in L_2(0, \infty; L_2(\Gamma))$.

(iii) The semigroup $S_F(t)$ in (i) is exponentially stable in the corresponding uniform operator topology; i.e. there are constants C, δ both positive, such that

$$\|S_F(t)\| \leqq C e^{-\delta t}, \qquad t \geqq 0$$

where $\|\ \|$ is here the uniform operator norm on $L_2(\Omega) \times H^{-1}(\Omega)$.

As a consequence of this result, for each initial piece of data $[y_0, y_1] \in E$, there exists a $u \in L_2(\Sigma)$—indeed $u = -D^*y_t$—such that the cost

$$\int_0^\infty \{(Ry(t), y(t))_{L_2(\Omega)} + (R'y_t(t), y_t(t))_{H^{-1}(\Omega)} + |u(t)|^2_{L_2(\Gamma)}\}\, dt. \tag{5.2}$$

R' nonnegative definite on $H^{-1}(\Omega)$ is *finite* and the corresponding regulator problem could be studied, indeed with the techniques of the present section modulo only minor additions. For sake of simplicity, and in line with §§ 2-4, we shall, however, concentrate on problem (5.1), rather than (5.2) (see also footnote 2). Thus, a fortiori, assumption (H) above is satisfied, indeed in a boundary feedback, constructive manner. Reference [L-T.5] gives a version of the *uniform* feedback stabilization problem, with feedback acting in the Dirichlet B.C. Treatment of the same problem, with feedback acting, instead, in the Neumann B.C. on the nontrivial portion Γ_1 or Γ, with homogeneous Dirichlet B.C. in the possibly empty portion Γ_0 of Γ, $\Gamma_0 \cup \Gamma_1 = \Gamma$, with $\bar{\Gamma}_0 \cap \bar{\Gamma}_1 = \phi$, was previously given, in its most general and complete version, in [L-4], improving upon [C.1].

It is a standard result that, as a consequence of assumption (H) being satisfied, there is a *unique optimal pair* u^0_∞, y^0_∞ of the quadratic problem (5.1).

We shall now proceed to obtain certain quantities of the optimal control problem (5.1) over $[0, \infty]$ (e.g. the Riccati operator $\mathscr{P}$, the optimal control and optimal solution, etc.) via a limit process as $T \uparrow \infty$ of the corresponding quantities of the corresponding optimal control problem (1.2a) over $[0, T]$. However, at the level of deriving the algebraic Riccati equation, we shall *not* take the limit of the differential Riccati equation, as done in much if not all, of the literature e.g. [B.1]. Rather, we shall first derive for $\mathscr{P}$ a relation (see (5.30) below), and then derive both the pointwise synthesis for the optimal control and the algebraic Riccati equation.

Relation (5.30) for $\mathscr{P}$ may be taken as a *defining* relation in the special but important case where the original s.c. (semi-) group $e^{\mathscr{A}t}$ is *uniformly bounded* on E, a situation large enough to include most physical systems, in particular, the canonical wave equation $-A(\xi, \partial) = \Delta$, where $e^{\mathscr{A}t}$ is indeed a unitary group on E. When $e^{\mathscr{A}t}$ is uniformly bounded on E, our treatment becomes particularly simple and direct, see Remark 5.2 below.

With $0 < T < \infty$ fixed, we shall henceforth indicate with a subscript "T" the quantities related to the optimal control problem (1.3a) on $[0, T]$ of §§ 2-4. Thus, $u_T^0(t) \equiv u_T^0(t, 0; x)$, $y_T^0(t) \equiv y_T^0(t, 0, x)$ and $\dot{y}_T^0(t) \equiv \dot{y}_T^0(t, 0; x)$ are the optimal control and solution of (1.3a) on $[0, T]$, starting at x, while $\mathscr{P}_T(t)$, $J_T(\cdot, \cdot)$, $\Phi_T(t, s)x = [y_T^0(t, s; x), \dot{y}_T^0(t; s; x)]$ are the corresponding Riccati operator, cost, and evolution operator. Extension by zero beyond T of f_T will be indicated by $\tilde{f}_T$. Thus, $\tilde{y}_T^0(t) \equiv y_T^0(t)$, $0 \leqq t \leqq T$, and $\tilde{y}_T^0(t) \equiv 0$, $t > T$ and similarly for $\tilde{u}_T^0(t)$ and $\tilde{\Phi}_T$. The same quantities with $T = \infty$ refer to the problem (5.1) over $[0, \infty]$. We begin by considering $\Phi_T(t, s)$.

Recalling Remark 4.3, we introduce the operators

$$\mathscr{R} \equiv \begin{vmatrix} R & 0 \\ 0 & 0 \end{vmatrix} \text{ nonnegative definite on } E, \tag{5.3}$$

$$\mathscr{L} \equiv \begin{vmatrix} L \\ \dfrac{dL}{dt} \end{vmatrix} : \text{continuous } L_2(0, T; L_2(\Gamma)) \to C([0, T]; E) \text{ by (2.3a, b).} \tag{5.4}$$

If $\mathscr{L}^* = [L^*, dL^*/dt]$ is the dual operator $(\mathscr{L}u, v)_{L_2(0, T; E)} = (u, \mathscr{L}^* v)_{L_2(0, T; L_2(\Gamma))}$, then $\mathscr{L}^*\mathscr{R} = [L^*R, 0]$ and

$$[I + \mathscr{L}\mathscr{L}^*\mathscr{R}] = \begin{vmatrix} I + LL^*R \\ I + + \dfrac{dL}{dt} L^*R \end{vmatrix} = \text{boundedly invertible on } L_2(0, T; E) \tag{5.5}$$

(same argument as below (2.8d)). Moreover, the optimal dynamics (2.31a) can be rewritten as

$$\begin{aligned} \Phi_T(t, s)x &= e^{\mathscr{A}(t-s)}x + \{\mathscr{L}_s u_T^0(\cdot, s; x)\}(t) \\ &= e^{\mathscr{A}(t-s)}x - \{\mathscr{L}_s\mathscr{L}_s^*\mathscr{R}\Phi_T(\cdot, s; x)\}(t) \end{aligned} \tag{5.6}$$

where, with $\mathscr{B}$ as in (4.13)

$$(\mathscr{L}_s u)(t) \equiv \int_s^t e^{\mathscr{A}(t-\tau)}\mathscr{B}u(\tau)\, d\tau, \tag{5.7}$$

$$(\mathscr{L}_s^* v)(t) \equiv \begin{cases} \mathscr{B}^* \displaystyle\int_s^T e^{\mathscr{A}^*(\tau - t)} v(\tau)\, d\tau, & s \leqq t \leqq T, \\ 0, & 0 \leqq t \leqq s, \end{cases} \tag{5.8}$$

using the same convention as in Remark 4.3.

LEMMA 5.1. *In the notation introduced above* (5.3), *we have*

$$\Phi_{T-t}(\sigma, 0) \equiv \Phi_T(t+\sigma, t) \quad \text{on } E, \quad 0 \leqq t < T, \quad 0 \leqq \sigma \leqq T-t.$$

Proof. From (5.6) we have for $x \in E$

$$\Phi_{T-t}(\sigma, 0)x + \{\mathscr{L}\mathscr{L}^* \mathscr{R} \Phi_{T-t}(\cdot, 0)x\}(\sigma) = e^{\mathscr{A}\sigma}x, \qquad 0 \leqq \sigma \leqq T-t$$

or explicitly, from (5.7), (5.8)

$$(5.9) \quad e^{\mathscr{A}\sigma}x = \Phi_{T-t}(\sigma, 0)x + \int_0^\sigma e^{\mathscr{A}(\sigma-\tau)}\mathscr{B}\left(\mathscr{B}^* \int_\tau^{T-t} e^{\mathscr{A}^*(r-\tau)}\mathscr{R}\Phi_{T-t}(r, 0)x\, dr\right) d\tau.$$

Similarly

$$\begin{aligned} e^{\mathscr{A}(t+\sigma-t)}x &= \Phi_T(t+\sigma, t)x + \{\mathscr{L}_t\mathscr{L}_t^*\mathscr{R}\Phi_T(\cdot, t)x\}(t+\sigma), \qquad 0 \leqq \sigma \leqq T-t \\ &= \Phi_T(t+\sigma, t)x + \int_t^{t+\sigma} e^{\mathscr{A}(t+\sigma-\tau)}\mathscr{B}(\mathscr{B}^* \int_\tau^T e^{\mathscr{A}^*(\alpha-\tau)}\mathscr{R}\Phi_T(\alpha, t)x\, d\alpha\, d\tau. \end{aligned}$$

Setting $\tau - t = \beta$ in the first integral and then $\alpha - t = r$ in the second integral yields

$$(5.10) \quad e^{\mathscr{A}\sigma}x = \Phi_T(t+\sigma, t)x + \int_0^\sigma e^{\mathscr{A}(\sigma-\beta)}\mathscr{B}\left(\mathscr{B}^* \int_\beta^{T-t} e^{\mathscr{A}^*(r-\beta)}\mathscr{R}\Phi_T(t+r, t)x\, dr\right) d\tau.$$

Comparison between (5.9) and (5.10) shows that both $\Phi_{T-t}(\sigma, 0)$ and $\Phi_T(t+\sigma, t)$ satisfy the same equation, say (5.10). But then the difference

$$(5.11) \qquad z(\sigma, t) \equiv \Phi_T(t+\sigma, t)x - \Phi_{T-t}(\sigma, 0)x \underset{(\text{in } \sigma)}{\in} C([0, T-t]; E)$$

satisfies $[I + \mathscr{L}\mathscr{L}^*\mathscr{R}]z(\cdot, t) = 0$. By (5.5), we deduce that $z(\sigma, t)$ is the zero element in $L_2(0, T-t; E)$ and by (5.11) the conclusion follows. □

THEOREM 5.2. *In the notation introduced above* (5.3), *we have*:

(i) *The* (*self-adjoint*) *nonnegative definite operator* $\mathscr{P}_T(0)$ *converges strongly on* E *to a* (*self-adjoint*) *nonnegative definite operator* $\mathscr{P}$ *as* $T \uparrow \infty$; *i.e.*

$$(5.12) \qquad \mathscr{P}x = \lim_{T \uparrow \infty} \mathscr{P}_T(0)x = \lim_{T \uparrow \infty} \int_0^T e^{\mathscr{A}^*\tau} \left| \begin{matrix} R\Phi_{T,1}(\tau, 0)x \\ 0 \end{matrix} \right| d\tau, \qquad x \in E.$$

(ii) $$\mathscr{P}_{T-t}(0) = \mathscr{P}_T(t), \quad 0 \leqq t < T.$$

(5.13)

(iii) $\mathscr{P}$ *in* (i) *can likewise be defined by*

$$(5.14) \qquad \mathscr{P}x = \lim_{T \uparrow \infty} \mathscr{P}_T(t)x = \lim_{T \uparrow \infty} \int_t^T e^{\mathscr{A}^*(\tau-t)} \left| \begin{matrix} R\Phi_{T,1}(\tau, t)x \\ 0 \end{matrix} \right| d\tau, \qquad x \in E$$

independently on t, $0 \leqq t < T$.

(iv) *For* $x \in E$

$$(5.15) \quad J_\infty^0 \equiv J_\infty^0(u_\infty^0(\cdot, 0; x), y_\infty^0(\cdot, 0; x)) = \int_0^\infty |u_\infty^0(t)|_\Gamma^2 + (Ry_\infty^0(t), y^0(t))_\Omega\, dt = (\mathscr{P}x, x)_E.$$

(v) *The optimal pair on* $[0, T]$ *for problem* (1.3a) *converges to the optimal pair on* $[0, \infty]$ *for problem* (5.1) *strongly in* L_2; *more precisely*

$$(5.16) \qquad \begin{aligned} &\tilde{u}_T^0 \to u_\infty^0 \quad \text{in } L_2(\Sigma_\infty), \\ &R^{1/2}\tilde{y}_T^0 \to R^{1/2}y_\infty^0 \quad \text{in } L_2(Q_\infty), \end{aligned}$$

for a suitable subsequence in $T \uparrow \infty$.

(vi) *For each t fixed, we have*

$$\left.\begin{aligned} &y_T^0(t) \to y_\infty^0(t) \text{ in } L_2(\Omega) \quad &(5.17a)\\ &\dot{y}_T^0(t) \to \dot{y}_\infty^0(t) \text{ in } H^{-1}(\Omega) \quad &(5.17b)\end{aligned}\right\} \textit{uniformly on bounded intervals}$$

as $T \uparrow \infty$, $t < T$. *Moreover, for* $x \in E$

$$\left|\begin{matrix} y_\infty^0(t, 0; x) \\ \dot{y}_\infty^0(t, 0; x)\end{matrix}\right| \equiv \left|\begin{matrix} y_\infty^0(t) \\ \dot{y}_\infty^0(t)\end{matrix}\right| \in C([0, T_0]; E) \quad \textit{for any } T_0 < \infty. \tag{5.17c}$$

Proof. Property (i). From the nonnegative definite operator $\mathscr{P}_T(0)$ satisfying

$$(\mathscr{P}_T(0)x, x)_E = J_T(u_T^0(\cdot, 0; x), y_T^0(\cdot, 0; x)) \equiv J_T^0, \qquad x \in E \tag{5.18}$$

(Lemma 2.3(iii)) one obtains, in a standard way (e.g. [B.1, p. 270]), the operator $\mathscr{P}$ as in (5.12) left; then (5.12) right follows from (2.22a).

Property (ii). Identity (5.13), which is intuitive in view of Lemma 2.3(iii), is a direct consequence of definition (2.22a) of $\mathscr{P}_T(t)$, via Lemma 5.1.

Property (iii) then follows by taking the limit in (5.13) as $T \to \infty$ and using (5.12) left, and (2.22a).

Property (iv). *Step* 1. As remarked below Remark 5.1, the existence of a unique optimal pair $u_\infty^0 \in L_2(\Sigma_\infty)$ and $y_\infty^0 \in L_2(Q_\infty)$ for problem (5.1) is a standard consequence of the assumption on the finiteness of the cost for some pair. Let $u_{\infty,T}^0$ and $y_{\infty,T}^0$ be the "cut" functions of u_∞^0 and y_∞^0 at $t = T$; i.e. $u_{\infty,T}^0(t)$ (resp. $y_{\infty,T}^0(t)$) coincides with $u_\infty^0(t)$ (resp. $y_\infty^0(t)$) over $0 \leqq t \leqq T$, and vanishes for $t > T$. From (5.18) with $x \in E$

$$\begin{aligned} (\mathscr{P}_T(0)x, x)_E &= \int_0^T |u_T^0(t)|_\Gamma^2 + (Ry_T^0(t), y_T^0(t))_\Omega \, dt \\ &\leqq \int_0^T |u_{\infty,T}^0(t)|_\Gamma^2 + (Ry_{\infty,T}^0(t), y_{\infty,T}^0(t))_\Omega \, dt \\ &\leqq J_\infty(u_\infty^0(\cdot, 0; x), y_\infty^0(\cdot, 0; x)) \equiv J_\infty^0 < \infty. \end{aligned} \tag{5.19}$$

Thus, the extended functions $\{\tilde{u}_T^0\}$ and $\{R^{1/2}\tilde{y}_T^0\}$ are contained in a fixed ball of $L_2(\Sigma_\infty)$ and $L_2(Q_\infty)$ respectively, for all T. Hence, we can extract subsequences

$$\tilde{u}_T^0 \text{ weakly convergent to, say, some } \tilde{u} \text{ in } L_2(\Sigma_\infty), \tag{5.20a}$$

$$R^{1/2}\tilde{y}_T^0 \text{ weakly convergent to, say, some } R^{1/2}\tilde{y} \text{ in } L_2(Q_\infty). \tag{5.20b}$$

Step 2. With reference to (5.20) we have, in fact, that $\tilde{y}$ is the solution of (1.1) due to $\tilde{u}$; i.e. for any $0 < T_0 < \infty$

$$\tilde{y} = C(\cdot)x_1 + S(\cdot)x_2 + L\tilde{u} \in C([0, T_0]; L_2(\Omega)). \tag{5.21}$$

Indeed, with $T > T_0$, $Lu_T^0 = L\tilde{u}_T^0$ converges weakly to $L\tilde{u}$ in $L_2(0, T_0; L_2(\Omega))$, while the extension by zero of $R^{1/2}$ applied to $y_T^0 = C(\ )x_1 + S(x)x_2 + Lu_T^0$ converges weakly to $R^{1/2}\tilde{y}$ in $L_2(0, T_0; L_2(\Omega))$. By uniqueness of the weak limit, (5.21) follows (first in $L_2(0, T_0; L_2(\Omega))$).

Step 3. Passing to the limit in (5.19) yields

$$(\mathscr{P}x, x)_E \leqq J_\infty^0 \tag{5.22}$$

by (5.12) left. On the other hand, the well-known lower semicontinuity of the quadratic cost J_∞ resulting from the weak convergence (5.20) [E-T.1, p. 11] completed with (5.21) gives the inequality in

$$(\mathscr{P}_T(0)x, x)_E = J_T(u_T^0, y_T^0) = J_\infty(\tilde{u}_T^0, \tilde{y}_T^0) \geqq J_\infty(\tilde{u}, \tilde{y})$$

from which, taking the limit via (5.12) left,

$$(\mathscr{P}x, x)_E \geqq J_\infty(\tilde{u}, \tilde{y}) \geqq J_\infty^0. \tag{5.23}$$

Then (5.22)-(5.23) yield (5.15):

$$(\mathscr{P}x, x)_E = J_\infty(\tilde{u}, \tilde{y}) = J_\infty^0 = J_\infty(u_\infty^0, y_\infty^0). \tag{5.24}$$

Property (v). By uniqueness of the optimal pair (*Step* 1), we conclude from (5.24) that

$$\tilde{u} = u_\infty^0 \quad \text{in } L_2(\Sigma_\infty), \qquad \tilde{y} = y_\infty^0 \quad \text{in } L_2(Q_\infty), \tag{5.25}$$

and thus, (5.20) becomes

(5.26a) $\tilde{u}_T^0$ converges weakly to u_∞^0 in $L_2(\Sigma_\infty)$,

(5.26b) $R^{1/2}\tilde{y}_T^0$ converges weakly to $R^{1/2}y_\infty^0$ in $L_2(Q_\infty)$.

On the other hand, the established convergence $J_T^0 \to J_\infty^0$ provides norm convergence:

$$\|\tilde{u}_T^0\|_{L_2(\Sigma_\infty)}^2 + \|R^{1/2}\tilde{y}_T^0\|_{L_2(Q_\infty)}^2 \to \|u_\infty^0\|_{L_2(\Sigma_\infty)}^2 + \|R^{1/2}y_\infty^0\|_{L_2(Q_\infty)}^2.$$

This, combined with the weak convergence in (5.26), yields the strong convergence (5.16) as desired.

Property (vi). For each fixed t, (5.16) implies $((Lu_T^0)(t) \to ((Lu_\infty^0)(t)$ in $L_2(\Omega)$ and $((dL/dt)u_T^0)(t) \to ((dL/dt)u_\infty^0)(t)$, by the continuity (2.3a, b) uniformly on bounded intervals. Then (5.17a, b) follow via the limit on the optimal dynamics on $[0, T]$. (5.17c) is a restatement of (5.21) via (5.25), used also for the corresponding dynamics of the velocity $\dot{y}_\infty^0$.

We next define the operator $\Phi_\infty(t)$ on E by

$$\Phi_\infty(t)x = \begin{vmatrix} \Phi_{\infty,1}(t)x \\ \Phi_{\infty,2}(t)x \end{vmatrix} = \begin{vmatrix} y_\infty^0(t, 0; x) \\ \dot{y}_\infty^0(t, 0; x) \end{vmatrix}, \qquad x \in E. \tag{5.27}$$

COROLLARY 5.3. *In the notation introduced above* (5.3) *and in* (5.27), *we have*:
(i)

$$\tilde{\Phi}_{T,1}(\cdot, 0)x \to \Phi_{\infty,1}(\cdot)x \quad \textit{in } L_2(Q_\infty), \qquad x \in E. \tag{5.28}$$

(ii) *For each fixed* $t > 0$, *as* $T \uparrow \infty$, $T > t$:

$$\Phi_T(t, 0)x \to \Phi_\infty(t)x,\ x \in E; \textit{ uniformly on bounded intervals.} \tag{5.29}$$

(iii) $\Phi_\infty(t)$ *is a strongly continuous semigroup on* E.
(iv) *The operator* $\mathscr{P}$ *defined on* E *by* (5.12) *or* (5.14) *satisfies*

$$\mathscr{P}x = \int_0^{t_0} e^{\mathscr{A}^* \tau} \begin{vmatrix} R\Phi_{\infty,1}(\tau)x \\ 0 \end{vmatrix} d\tau + e^{\mathscr{A}^* t_0}\mathscr{P}\Phi_\infty(t_0)x, \qquad x \in E \tag{5.30}$$

where t_0 *is an arbitrary point* $0 \leqq t_0 < \infty$.

Proof. The convergence properties (i) and (ii) are restatements of properties (5.16b) and (5.17a) in Theorem 5.2.

(iii) Recalling (5.27) and (5.17c), we see that $\Phi_\infty(t)$ is strongly continuous on E. The semigroup property of $\Phi_\infty(t)$ then follows from (5.29) via the evolution properties

of $\Phi_T(\ ,\ )$ and Lemma 5.1: indeed, for $x \in E$

$$\begin{aligned}\Phi_T(t+\tau, 0)x &= \Phi_T(t+\tau, \tau)\Phi_T(\tau, 0)x \quad \text{(by Lemma 2.1(ii))}\\ &= \Phi_{T-\tau}(t, 0)\Phi_T(\tau, 0)x \quad \text{(by Lemma 5.1)}\\ &= \Phi_{T-\tau}(t, 0)[\Phi_T(\tau, 0)x - \Phi_\infty(\tau, 0)x] + \Phi_{T-\tau}(t, 0)\Phi_\infty(\tau, 0)x.\end{aligned}$$

Taking the limit as $T \uparrow \infty$ we obtain by virtue of (5.29)

$$\Phi_\infty(t+\tau)x = \Phi_\infty(t, 0)\Phi_\infty(\tau, 0)x$$

as desired, since for fixed t, $\Phi_{T-\tau}(t, 0)$ is uniformly bounded in T in the operator norm of E by the Principle of Uniform Boundedness.

As to (iv), we have from (5.22) for an arbitrary $0 < t_0 < T$ and $x \in E$

$$\begin{aligned}\mathscr{P}x &= \lim_{T \uparrow \infty} \left\{ \int_0^{t_0} e^{\mathscr{A}^* \tau} \left| \begin{matrix} R\Phi_{T,1}(\tau, 0)x \\ 0 \end{matrix} \right| d\tau + e^{\mathscr{A}^* t_0} \int_{t_0}^{T} e^{\mathscr{A}^*(\tau - t_0)} \left| \begin{matrix} R\Phi_{T,1}(\tau, t_0)\Phi_T(t_0, 0)x \\ 0 \end{matrix} \right| d\tau \right\}\\ &= \int_0^{t_0} e^{\mathscr{A}^* \tau} \left| \begin{matrix} R\Phi_{\infty,1}(\tau, 0)x \\ 0 \end{matrix} \right| d\tau + e^{\mathscr{A}^* t_0} \lim_{T \uparrow \infty} \mathscr{P}_T(t_0)\Phi_T(t_0, 0)x\end{aligned}$$

where we have used (5.28) and (5.14). But

$$\begin{aligned}\lim_{T \uparrow \infty} \mathscr{P}_T(t_0)\Phi_T(t_0, 0)x &= \lim \{\mathscr{P}_T(t_0)[\Phi_T(t_0, 0)x - \Phi_\infty(t_0, 0)x] + \mathscr{P}_T(t_0)\Phi_\infty(t_0, 0)x\}\\ &= \mathscr{P}\Phi_\infty(t_0, 0)x = \mathscr{P}\Phi_\infty(t_0)x \qquad (5.31)\end{aligned}$$

by (5.29), the uniform boundedness of $\mathscr{P}_T(t_0)$ in T for t_0 fixed, and (5.14).

Remark 5.2. With reference to Remark 5.1, suppose that we are studying the minimization problem, over all $u \in L_2(\Sigma_\infty)$, of the cost functional (5.2), which penalizes also the velocity component. Let $\mathscr{R} = \left| \begin{matrix} R & 0 \\ 0 & R' \end{matrix} \right|$ be positive definite on E. Then, the same procedure culminating in Corollary 5.3 leads to the further conclusions that:

(i) The corresponding semigroup $\Phi_\infty(t)$ on E satisfies

$$\Phi_\infty(t)x \in L_2(0, \infty; E), \qquad x \in E.$$

(ii) The corresponding operator $\mathscr{P}$ satisfies

$$\mathscr{P}x = \int_0^{t_0} e^{\mathscr{A}^* \tau} \mathscr{R}\Phi_\infty(\tau)x\, d\tau + e^{\mathscr{A}^* t_0}\mathscr{P}\Phi_\infty(t_0)x, \qquad x \in E, \tag{5.32a}$$

a counterpart of (5.30). Suppose in addition, that the original semigroup $e^{\mathscr{A}t}$ is *uniformly bounded* on E: $\|e^{\mathscr{A}t}\|_{\mathscr{L}(E)} \leqq C$, $t \geqq 0$, a situation which includes many physical hyperbolic systems, in particular the canonical wave equation $-A(\xi, \partial) = \Delta$, where $e^{\mathscr{A}t}$ is indeed a unitary s.c. group on E. Then, by a well-known result [D.1], property (i) above: $\Phi_\infty(t)x \in L_2(0, \infty; E)$ implies that $\Phi_\infty(t)$ is indeed exponentially stable on E:

$$\|\Phi_\infty(t)\|_{\mathscr{L}(E)} \leqq M_\omega e^{-\omega t}, \qquad t \geqq 0, \quad \text{some } \omega > 0.$$

Then, in (5.32) we can let $t_0 \to \infty$, thereby obtaining

$$\mathscr{P}x = \int_0^{\infty} e^{\mathscr{A}^* \tau} \mathscr{R}\Phi_\infty(\tau)x\, d\tau, \qquad x \in E, \tag{5.32b}$$

a *defining* formula for $\mathscr{P}$. Thus, under the assumptions of the present remark, the procedure which follows simplifies considerably with $\Phi_\infty(t_0) = \Phi_\infty(\infty) = 0$. □

THEOREM 5.4. *With $\mathscr{P}$ and $\Phi_\infty(\ )$ defined by* (5.12) *and below* (5.27), *and $\mathscr{B}^*$ as in* (2.25):

$$-u_\infty^0(t,0;x) = -u_\infty^0(t) = \mathscr{B}^*\mathscr{P}\Phi_\infty(t)x, \qquad x \in E \tag{5.33a}$$

where

$$\mathscr{B}^*\mathscr{P}\Phi_\infty(t)\text{: continuous } E \to L_2(\Sigma_\infty). \tag{5.33b}$$

Proof. Recall

$$e^{\mathscr{A}^* t}y = \begin{vmatrix} C^*(t)y_1 - S^*(t)A^{*1/2}A^{-1/2}y_2 \\ A^{1/2}A^{*1/2}S^*(t)y_1 + A^{1/2}C^*(t)A^{-1/2}y_2 \end{vmatrix}, \qquad y \in E \tag{5.34}$$

computed as in obtaining (2.24). By (2.25)

$$\mathscr{B}^* e^{\mathscr{A}^* t}y = D^*A^*S^*(t)y_1 + D^*A^{*1/2}C^*(t)A^{-1/2}y_2 \tag{5.35a}$$

$$\text{: continuous } E \to L_2(0, T; L_2(\Gamma)) \tag{5.35b}$$

which follows from Theorem 3.2. Next, recall (2.28) and (2.22a)

$$-u_T^0(t,0;x) = -u_T^0(t) = \mathscr{B}^*\mathscr{P}_T(t)\Phi_T(t,0)x = I_{1T}(t) + I_{2T}(t), \tag{5.36}$$

$$I_{1T}(t) = \mathscr{B}^* \int_t^{t_0} e^{\mathscr{A}^*(\tau - t)} \begin{vmatrix} R\Phi_{T,1}(\tau,0)x \\ 0 \end{vmatrix} d\tau, \tag{5.37a}$$

$$I_{2T}(t) = \mathscr{B}^* e^{\mathscr{A}^*(t_0 - t)} \int_{t_0}^{T} e^{\mathscr{A}^*(\tau - t_0)} \begin{vmatrix} R\Phi_{T,1}(\tau, t_0)\Phi_T(t_0,0)x \\ 0 \end{vmatrix} d\tau. \tag{5.37b}$$

But by (5.35a) and (3.6b)

$$\begin{aligned} I_{1,T}(t) &= -D^*A^*S^*(t) \int_t^{t_0} C^*(\tau)R\Phi_{T,1}(\tau,0)x\, d\tau \\ &\quad + D^*A^{*1/2}C^*(t)A^{*1/2} \int_t^{t_0} S^*(\tau)R\Phi_{T,1}(\tau,0)x\, d\tau \\ &\equiv \{\bar{F}(\cdot)\Phi_{T,1}(\cdot,0)x\}(t) \end{aligned}$$

where, by Lemma 3.1, $\bar{F}(\cdot)$ is continuous $L_1(0, T_1; L_2(\Omega)) \to L_2(0, T_1; L_2(\Gamma))$ for any finite T_1. Thus, by (5.28) in Corollary 5.3, we conclude from (5.37a)

$$\begin{aligned} \lim_{T \uparrow \infty} I_{1T}(t) &= \mathscr{B}^* \int_t^{t_0} e^{\mathscr{A}^*(\tau - t)} \begin{vmatrix} R\Phi_{\infty,1}(\tau)x \\ 0 \end{vmatrix} d\tau \\ &= \mathscr{B}^* \int_0^{t_0 - t} e^{\mathscr{A}^*\sigma} \begin{vmatrix} R\Phi_{\infty,1}(\sigma)\Phi_\infty(t)x \\ 0 \end{vmatrix} d\sigma \end{aligned} \tag{5.38}$$

the limit being taken in the $L_2(0, T_1; L_2(\Gamma))$-topology. As to I_{2T} in (5.37b), recall (2.22a) and (5.31) to get

$$\lim_{T \uparrow \infty} I_{2T}(t) = \lim_{T \uparrow \infty} \mathscr{B}^* e^{\mathscr{A}^*(t_0 - t)}\mathscr{P}_T(t_0)\Phi_T(t_0,0)x = \mathscr{B}^* e^{\mathscr{A}^*(t_0 - t)}\mathscr{P}\Phi_\infty(t_0)x \tag{5.39}$$

by (5.35b), the limit being taken as in (5.38). We return to (5.36) and use (5.16a) on the left, and (5.38), (5.39), (5.30) (with t_0 in (5.30) replaced by $t_0 - t$ now) on the right, thus obtaining (5.33).

DEFINITION 5.1. Henceforth, we let $\mathscr{A}_F$ (F for "feedback") be the closed, densely defined infinitesimal generator of the s.c. semigroup $\Phi_\infty(t)$ on E of Corollary

5.3(iii): $\Phi_\infty(t) = e^{\mathscr{A}_F t}$:

$$\frac{d\Phi_\infty(t)}{dt}x = \mathscr{A}_F\Phi_\infty(t)x = \Phi_\infty(t)\mathscr{A}_F x,\ x \in \mathscr{D}(\mathscr{A}_F). \tag{5.40}$$

LEMMA 5.5. *With $\mathscr{P}$ defined by* (5.12), *we have $\mathscr{D}(\mathscr{B}^*\mathscr{P}) \supset \mathscr{D}(\mathscr{A}_F)$ and for $x \in \mathscr{D}(\mathscr{A}_F)$*

$$\begin{aligned}\mathscr{B}^*\mathscr{P}x = D^*Rx_1 + D^*\int_0^{t_0} C^*(\tau)R\Phi_{\infty,1}(\tau)\mathscr{A}_F x\,d\tau \\ -D^*C^*(t_0)R\Phi_{\infty,1}(t_0)x + \mathscr{B}^*\,e^{\mathscr{A}^*t_0}\mathscr{P}\Phi_\infty(t_0)x \in L_2(\Gamma)\end{aligned} \tag{5.41}$$

where t_0 (depending on x) is chosen so that the last term in (5.41) *is well-defined on $L_2(\Gamma)$. See the proof below (the measure of the set of all such t_0's contained in a finite interval is the length of this interval).*

Proof. We use now (5.30) complemented by (5.35a). Integration by parts on the integral term $\int_0^{t_0}$ of (5.30) produces the first three terms at the right of (5.41), for an arbitrary $0 < t_0$. Next, note that

$$\mathscr{B}^*\,e^{\mathscr{A}^*t_0}\mathscr{P}\Phi_\infty(t_0)x = \mathscr{B}^*\,e^{\mathscr{A}^*t_0}\mathscr{P}\int_0^{t_0}\Phi_\infty(t)\mathscr{A}_F x\,dt + \mathscr{B}^*\,e^{\mathscr{A}^*t_0}\mathscr{P}x \tag{5.42}$$

since $d\Phi_\infty(t)x/dt = \Phi_\infty(t)\mathscr{A}_F x$. But each of the two terms on the right of (5.42) are in $L_2(0, T_1; L_2(\Gamma))$ in the variable t_0, for any finite T_1, by (5.35a, b) and Lemma 3.1. For each x, the measure of all t_0 (depending on x) in $[0, T_1]$ for which both terms are in $L_2(\Gamma)$ is equal to T_1.

We next provide information on $\mathscr{A}_F$.

LEMMA 5.6. *For $x \in E$ and $t \geqq 0$*

$$\frac{d\Phi_\infty(t)x}{dt} = [\mathscr{A} - \mathscr{B}\mathscr{B}^*\mathscr{P}]\Phi_\infty(t)x \in [\mathscr{D}(\mathscr{A}^*)]'. \tag{5.43}$$

Thus,

$$[\mathscr{A} - \mathscr{B}\mathscr{B}^*\mathscr{P}]\Phi_\infty(t)x = \mathscr{A}_F\Phi_\infty(t)x = \Phi_\infty(t)\mathscr{A}_F x,\ x \in \mathscr{D}(\mathscr{A}_F),\qquad t > 0, \tag{5.44a}$$

$$[\mathscr{A} - \mathscr{B}\mathscr{B}^*\mathscr{P}]x = \mathscr{A}_F x,\qquad x \in \mathscr{D}(\mathscr{A}_F). \tag{5.44b}$$

Proof. From the optimal dynamics and the optimal control in (5.33a),

$$(\Phi_\infty(t)x, z)_E = (e^{\mathscr{A}t}x, z)_E - \left(\int_0^t e^{\mathscr{A}(t-\tau)}\mathscr{B}\mathscr{B}^*\mathscr{P}\Phi_\infty(\sigma)x\,d\tau, z\right)_E. \tag{5.45}$$

We differentiate in t with $x \in E$ and $z \in \mathscr{D}(\mathscr{A})$

$$\begin{aligned}\left(\frac{d\Phi_\infty}{dt}(t)x, z\right)_E = (e^{\mathscr{A}t}x, \mathscr{A}^*z)_E - (\mathscr{B}\mathscr{B}^*\mathscr{P}\Phi_\infty(t)x, z)_E \\ -\left(\int_0^t e^{\mathscr{A}(t-\tau)}\mathscr{B}\mathscr{B}^*\mathscr{P}\Phi_\infty(\tau)x\,d\tau, \mathscr{A}^*z\right)_E\end{aligned} \tag{5.46}$$

where the second term on the right of (5.46) is well-defined as a duality pairing on $[\mathscr{D}(\mathscr{A}^*)]' \times \mathscr{D}(\mathscr{A}^*)$ by (5.33b) and, say $\mathscr{A}^{-1}\mathscr{B}$ being bounded on E. Re-using (5.45) with z replaced by $\mathscr{A}^*z$ inside (5.46) yields (5.43).

LEMMA 5.7. *With $\mathscr{P}$ defined by* (5.12), *we have*

(i) $\mathscr{A}^*\mathscr{P}: \mathscr{D}(\mathscr{A}_F) \to E$.

(ii) $\mathscr{A}_F^*\mathscr{P}: \mathscr{D}(\mathscr{A}) \to E$.

Proof. Again, we use (5.30).

(i) For $x \in \mathscr{D}(\mathscr{A}_F)$, we apply $\mathscr{A}^*$ to (5.30) and integrate both sides in t_0 (variable) on some $[0, T_0]$. Performing integration by parts on the right side, we easily see that $T_0\mathscr{A}^*\mathscr{P}x \in E$.

(ii) To show $\mathscr{P}\mathscr{A}_F: E \to [\mathscr{D}(\mathscr{A})]'$, we take $x \in E$, $z \in \mathscr{D}(\mathscr{A})$ and compute $(\mathscr{P}\mathscr{A}_F x, z)_E$ via (5.30). Integrating both sides of this expression in t_0 over some $[0, T_0]$, (the right side by parts) we likewise find that $T_0(\mathscr{P}\mathscr{A}_F x, z)_E$ is well-defined.

LEMMA 5.8. *With $\mathscr{P}$ defined by* (5.12), *we have*

$$\text{(i)} \qquad -\mathscr{A}^*\mathscr{P}x = \begin{vmatrix} Rx_1 \\ 0 \end{vmatrix} + \mathscr{P}\mathscr{A}_F x \in E, \qquad x \in \mathscr{D}(\mathscr{A}_F).$$

$$\text{(ii)} \qquad -\mathscr{A}_F^*\mathscr{P}z = \begin{vmatrix} Rz_1 \\ 0 \end{vmatrix} + \mathscr{P}\mathscr{A}z \in E, \qquad z \in \mathscr{D}(\mathscr{A}).$$

Proof. Differentiate in t_0 the expression $(\mathscr{P}x, z)_E$ with $\mathscr{P}$ given by (5.30) and then set $t_0 = 0$. We get

$$\text{(5.47)} \qquad \left(\begin{vmatrix} Rx_1 \\ 0 \end{vmatrix}, z \right)_E + (\mathscr{P}\mathscr{A}_F x, z)_E + (\mathscr{P}x, \mathscr{A}z)_E = 0, \quad x \in \mathscr{D}(\mathscr{A}_F), \quad z \in \mathscr{D}(\mathscr{A}).$$

From here, using the a priori regularity of $\mathscr{A}^*\mathscr{P}$ and $\mathscr{A}_F^*\mathscr{P}$ given by Lemma 5.7, we extend the above inner products by continuity to all $z \in E$ with $x \in \mathscr{D}(\mathscr{A}_F)$, and to all $x \in E$ with $z \in \mathscr{D}(\mathscr{A})$. Lemma 5.8 follows. □

COROLLARY 5.9. *For $\mathscr{P}$ defined by* (5.12)

$(\mathscr{B}^*\mathscr{P}x, \mathscr{B}^*\mathscr{P}y) =$ *well-defined for $x, y \in \mathscr{D}(\mathscr{A})$, or else $x, y \in \mathscr{D}(\mathscr{A}_F)$. Thus,*

$$\mathscr{B}^*\mathscr{P}: \begin{vmatrix} \mathscr{D}(\mathscr{A}) \\ \mathscr{D}(\mathscr{A}_F) \end{vmatrix} \to L_2(\Gamma).$$

Proof. From Lemma 5.6, (5.44b)

$$\text{(5.48)} \qquad -(\mathscr{B}^*\mathscr{P}x, \mathscr{B}^*\mathscr{P}y)_\Gamma = (\mathscr{P}\mathscr{A}_F x, y)_E - (\mathscr{P}\mathscr{A}x, y)_E$$

where the first term on the right is well-defined by Lemma 5.8(ii) for $x \in E$ and $y \in \mathscr{D}(\mathscr{A})$ (or $x \in \mathscr{D}(\mathscr{A}_F)$ and $y \in E$) and the second term is well-defined for $x \in \mathscr{D}(\mathscr{A})$ and $y \in E$ (or, by Lemma 5.8(i), for $x \in E$ and $y \in \mathscr{D}(\mathscr{A}_F)$). □

THEOREM 5.10. (i) *The operator $\mathscr{P}$ defined by* (5.12) *satisfies the following algebraic Riccati equation*

$$\text{(5.49)} \qquad (Rx_1, y_1)_\Omega + (\mathscr{P}x, \mathscr{A}y)_E + (\mathscr{P}\mathscr{A}x, y)_E = (\mathscr{B}^*\mathscr{P}x, \mathscr{B}^*\mathscr{P}y)_\Gamma$$

for $x, y \in \mathscr{D}(\mathscr{A})$, or else for $x, y \in \mathscr{D}(\mathscr{A}_F)$.

Proof. (i) Combine Lemma 5.8 and Corollary 5.9. □

Remark 5.3. In the study of the minimization of the functional cost (5.2) which penalizes also the velocity component, we have already noted in Remark 5.2 that, for $\mathscr{R} = \begin{vmatrix} R & 0 \\ 0 & R' \end{vmatrix}$ positive definite on E, the corresponding semigroup $\Phi_\infty(t)$ is *exponentially stable here*: $(\#)$ $\|\Phi_\infty(t)\|_{\mathscr{L}(E)} \leqq M_\omega e^{-\omega t}$, $t \geqq 0$, $\omega < 0$. From here, it then follows, via a standard argument along the lines of [B.1, Corollary 5.3.1, pp. 272–273] or the notes [DaP.1], that the corresponding algebraic Riccati equation ((5.48) with the first term replaced by $(\mathscr{R}x, z)_E$) admits a unique self-adjoint nonnegative definite solution within

the class of linear operators $\mathscr{P} \in \mathscr{L}(E)$ such that $\mathscr{B}^*\mathscr{P} \in \mathscr{L}(\mathscr{D}(\mathscr{A}_F), L_2(\Gamma))$. Indeed, a generalization of this uniqueness result is available, as stated by:

THEOREM 5.11. *With reference to the optimal problem* (5.2), *let* $\mathscr{R} = \left|\begin{smallmatrix} R & 0 \\ 0 & R' \end{smallmatrix}\right|$ *be nonnegative definite on E, so that the existence statement of the corresponding algebraic Riccati equation* ((5.48) *with the first term replaced by* $(\mathscr{R}x, z)_E$) *holds true, as in Theorem* 5.10. *Assume further that*[11] *there exists an operator* $\mathscr{K} = \left|\begin{smallmatrix} K & 0 \\ 0 & K' \end{smallmatrix}\right|$ *with* $K, K' \in \mathscr{L}(L_2(\Omega)) \cap \mathscr{L}(\mathscr{D}(A))$ *such that the s.c. group generator* $\mathscr{A}_K = \mathscr{A} + \mathscr{K}\mathscr{R}^{1/2}$ *is exponentially stable as* $t \to +\infty$: $\|e^{\mathscr{A}_K t}\|_{\mathscr{L}(E)} \leqq M_\delta e^{-\delta t}$, $t \geqq 0$ *for some* $\delta > 0$. *Then, said algebraic Riccati equation admits a unique self-adjoint nonnegative definite solution within the class of linear operators* $\mathscr{P} \in \mathscr{L}(E)$ *such that* $\mathscr{B}^*\mathscr{P}^* \in \mathscr{L}(\mathscr{D}(\mathscr{A}_F), L_2(\Gamma))$.

Proof. As noted in the first paragraph, it suffices to show inequality (#), i.e. the exponential stability of $\Phi_\infty(t)$ on E for $t \geqq 0$. To this end, we write as usual

$$\frac{d\Phi_\infty(t)}{dt}x = (\mathscr{A} - \mathscr{B}\mathscr{B}^*\mathscr{P})\Phi_\infty(t)x + \mathscr{K}\mathscr{R}^{1/2}\Phi_\infty(t)x - \mathscr{K}\mathscr{R}^{1/2}\Phi_\infty(t)x, \qquad x \in \mathscr{D}(\mathscr{A}_F)$$

on E, or else

$$\frac{d\Phi_\infty(t)x}{dt} = \mathscr{A}_K\Phi_\infty(t)x - \mathscr{K}\mathscr{R}^{1/2}\Phi_\infty(t)x - \mathscr{B}\mathscr{B}^*\mathscr{P}\Phi_\infty(t)x$$

on $[\mathscr{D}(\mathscr{A})]'$. Thus, for $x \in E$

$$\Phi_\infty(t)x = e^{\mathscr{A}_K t}x + \int_0^t e^{\mathscr{A}_K(t-\tau)}\mathscr{K}\mathscr{R}^{1/2}\Phi_\infty(\tau)x\,d\tau + \int_0^t e^{\mathscr{A}_K(t-\tau)}\mathscr{B}\mathscr{B}^*\mathscr{P}\Phi_\infty(\tau)x\,d\tau$$

where

$$\int_0^\infty \|e^{\mathscr{A}_K t}x\|_E^2\,dt + \int_0^\infty \left\| \int_0^t e^{\mathscr{A}_K(t-\tau)}\mathscr{K}\mathscr{R}^{1/2}\Phi_\infty(\tau)x\,d\tau \right\|^2 dt \leqq C\|x\|_E^2, \qquad x \in E$$

since $\mathscr{R}^{1/2}\Phi_\infty(t)x \in L_2(0, \infty; E)$, by the finiteness of the optimal cost, as usual. Thus, in order to apply [D.1] and conclude with exponential stability of $\Phi_\infty(t)$ on E, what is left is to show the following result, the *crux* of our proof. This is the counterpart of the regularity Theorem 1.1(i), via Remark 4.3, extended to the infinite time interval for exponentially stable $e^{\mathscr{A}_K t}$.

LEMMA 5.12. *The map*

$$(\mathscr{L}_K u)(t) \equiv \int_0^t e^{\mathscr{A}_K(t-\tau)}\mathscr{B}u(\tau)\,d\tau$$

is continuous $L_2(0, \infty; L_2(\Gamma)) \to L_2(0, \infty; E)$; *equivalently, the map* $\mathscr{L}_K^*$: $(\mathscr{L}_K u, v)_{L_2(0,\infty;E)} = (u, \mathscr{L}_K^* v)_{L_2(0,\infty;L_2(\Gamma))}$, *given by*

$$(\mathscr{L}_K^* v)(t) \equiv \mathscr{B}^* \int_t^\infty e^{\mathscr{A}_K^*(\tau - t)}v(\tau)\,d\tau$$

is continuous $L_2(0, \infty; E) \to L_2(0, \infty; L_2(\Gamma))$.

Proof. The proof for $\mathscr{L}_K^*$ is carried out by applying to an hyperbolic dynamics the same multiplier technique used in [L-L-T.1, see in particular Remark 3.1] for a

[11] This assumption is plainly satisfied with $\mathscr{K} = -k\mathscr{R}^{-1/2}$, k sufficiently large, when $\mathscr{R}$ is positive definite.

finite time interval or in [L-T.5, Prop. 3.1] on an infinite time interval with a decaying factor $e^{-\beta t}$, $\beta > 0$. To this end, use is made of the definition $\mathscr{B}^*[v_1, v_2] = D^* v_2$ ((2.25) in the self-adjoint case) along with $D^* v_2 = D^* A A^{-1} v_2 = (\partial/\partial \nu) A^{-1} v_2$ on Γ, see (3.8). We refer to these last two references for details. □

Appendix 1. Proof of Proposition 2.2. By identity (3.6b), we re-write (2.21) (as done a few times in § 3, e.g. in (3.21)) as

$$\begin{aligned}(2.21) = D^* A^* S^*(t) \int_t^T C^*(\tau)\Phi_1(\tau, 0) y \, d\tau \\ -D^* A^{*1/2} C^*(t) \int_t^T A^{*1/2} S^*(\tau)\Phi_1(\tau, 0) y \, d\tau, \qquad y \in E.\end{aligned} \tag{A1.1}$$

The abstract Lemma 3.1 can then be applied with $X = L_2(\Omega)$, $Y = L_2(\Gamma)$, the integrands of (A1.1) in $C([0, T]; X)$, by Lemma 2.1(iii) and (1.10)] and the operators outside the integrals continuous $X \to L_2(0, T; X)$, see Theorem 3.2. As a result, (A1.1) $\in$ $L_2(0, T; L_2(\Gamma))$. □

Appendix 2. Change of order of integration below (2.32) Lemma 2.3(ii). In order to justify the change of order of integration, one first considers the regularized problem; i.e. the same problem with J replaced by J_ε

$$J_\varepsilon = \int_0^T (R_\varepsilon y(t), y(t))_\Omega \, dt + \int_0^T |u(t)|_\Gamma^2 \, dt$$

where $R_\varepsilon \to R$ strongly, $\varepsilon \downarrow 0$, R_ε positive self-adjoint and $A^* R_\varepsilon$ bounded for each $\varepsilon > 0$ (for example, one can use the resolvent $R(\lambda, A)$ of A and take $R_\varepsilon = (1/\varepsilon) R(1/\varepsilon, A^*) R$). For the problem with R_ε, it can be readily shown that

$$\mathscr{B}^* \mathscr{P}_\varepsilon \text{: continuous } L_2(\Omega) \to C([0, T]; L_2(\Gamma)).$$

Then, absolute integrability holds

$$\int_t^T \int_t^\tau |(D^* A^* S^*(\tau - \sigma)\Phi_1(\tau, t)x, \mathscr{B}^* \mathscr{P}_\varepsilon(\sigma)\Phi(\sigma, t)y)_\Gamma| \, d\sigma \, d\tau$$

the function on the left of the inner product being in L_2 in the variable σ, the one on the right being continuous in σ. Hence, change of the order of integration is legal for the problem with ε, and one obtains

$$\begin{aligned}(\mathscr{P}_\varepsilon(t)x, y)_E = \int_t^T (R_\varepsilon \Phi_{1,\varepsilon}(\tau, t)x, \Phi_{1,\varepsilon}(\tau, t)y)_\Omega \, d\tau \\ + \int_t^T (\mathscr{B}^* \mathscr{P}_\varepsilon(\tau)\Phi_\varepsilon(\tau, t)x, \mathscr{B}^* \mathscr{P}_\varepsilon(\tau)\Phi_\varepsilon(\tau, t)y)_\Gamma \, d\tau.\end{aligned}$$

It is then straightforward to show that $\mathscr{P}_\varepsilon \to \mathscr{P}$, $\Phi_\varepsilon \to \Phi$; $u_\varepsilon \to u$ in $L_2(0, T; \cdot)$. Passing to the limit yields (2.29). □

REFERENCES

[A.1] J. P. AUBIN, *Un théorème de compacité*, C.R. Acad. Sci., 256 (1963), pp. 5042-5044.

[B.1] A. V. BALAKRISHNAN, *Applied Functional Analysis*, Second Edition 1981, Springer-Verlag, New York.

[C.1] G. CHEN, *A note on the boundary stabilization of the wave equation*, this Journal, 19 (1981), pp. 106-113.

[C-P.1] R. CURTAIN AND A. PRITCHARD, *An abstract theory for unbounded control action for distributed parameter systems*, this Journal, 15 (1977), pp. 566-611.

[C-P.2] ———, *Infinite Dimensional Linear Systems Theory*, Lecture Notes in Control 8, Springer-Verlag, New York, 1978.

[DaP.1] G. DAPRATO, *Notes on Riccati's equations* (in Italian), 1983.

[DaP-L-T.1] G. DAPRATO, I. LASIECKA AND R. TRIGGIANI, *A direct study of Riccati equations arising in hyperbolic boundary control problems*, to appear.

[D.1] R. DATKO, *Extending a theorem of Liapunov to Hilbert space*, J. Math. Appl., 32 (1970), pp. 610-616.

[E-T.1] I. EKELAND AND R. TEMAN, *Convex Analysis and Variational Problems*, North-Holland, Amsterdam, 1976.

[F.1] H. O. FATTORINI, *Ordinary differential equations in linear topological spaces*, I *and* II, J. Differential Equations, 5 (1968), pp. 72-105, and 6 (1969), pp. 537-565.

[F.2] D. FUJIWARA, *Concrete characterizations of domains of fractional powers of some elliptic differential operators of the second order*, Proc. Acad. Japan, 43 (1967), pp. 82-86.

[F.3] F. FLANDOLI, *Infinite dimensional algebraic Riccati equation arising in a boundary control problem*, preprint.

[F.4] H. O. FATTORINI, *Un teorema de perturbacion para generadores de functiones coseno Revista Unione Matem*, Argentina, 25 (1971).

[H-P.1] E. HILLE AND R. S. PHILLIPS, *Functional Analysis and Semigroups*, AMS, Colloquium Publ. Vol. XXXI, American Mathematical Society, Providence, RI, 1957.

[K.1] T. KATO, *Perturbation Theory of Linear Operators*, Springer-Verlag, New York, 1966.

[K.N.1] J. L. KOHN AND L. NIRENBERG, *An algebra of pseudo-differential operators*, Comm. Pure Appl. Math., XVIII (1965), pp. 269-305.

[L.1] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, 1971.

[L.2] I. LASIECKA, *Unified theory for abstract parabolic boundary problems—a semigroup approach*, Appl. Math. Optim., 6 (1980), pp. 31-62.

[L.3] J. L. LIONS, *Espaces d'interpolation et domain de puissances fractionnaires d'opérateurs*, J. Math. Soc. Japan, 14 (1962), pp. 233-241.

[L.4] J. LAGNESE, *Decay of solutions of wave equations in a bounded region with boundary dissipation*, J. Differential Equations, 50 (1983), pp. 163-182.

[L.5] D. G. LUENBERGER, *Optimization by Vector Space Methods*, John Wiley, New York, 1969.

[L-T.1] I. LASIECKA AND R. TRIGGIANI, *A cosine operator approach to modelling $L_2(0, T; L_2(\Gamma))$-boundary input hyperbolic equations*, Appl. Math. Optim., 7 (1981), pp. 35-93.

[L-T.2] ———, *Dirichlet boundary control problems for parabolic equations with quadratic cost: analyticity and Riccati's feedback synthesis*, this Journal, 21 (1983), pp. 41-67.

[L-T.3] ———, *Regularity of hyperbolic equations under $L_2(0, T; L_2(\Gamma))$—Dirichlet boundary terms*, Appl. Math. Optim., 10 (1983), pp. 275-286.

[L-T.4] ———, *The quadratic cost problem for $L_2(0, T; L_2(\Gamma))$-boundary input hyperbolic equations*, presented at Workshop on Control Theory for Distributed Parameter Systems, University of Graz, Vorau, Austria, July 1982.

[L-T.5] ———, *Uniform exponential energy decay of the wave equation in a bounded domain with $L_2(0, \infty; L_2(\Gamma))$-boundary feedback in the Dirichlet B.C.*, to appear

[L-L-T.1] I. LASIECKA, J. L. LIONS AND R. TRIGGIANI, *Non homogeneous boundary value problems for second order hyperbolic operators*. J. Math. Pures, Appl., to appear.

[L-M.1] J. L. LIONS AND E. MAGENES, *Non-homogeneous Boundary Value Problems and Applications*, Vols. I, II, Springer-Verlag, Berlin-Heidelberg-New York, 1972.

[N.1] J. NECAS, *Les methodes directes en théories des équations elliptiques*, Masson et cie, 1967, Paris.

[N.2] B. NAGY, *On cosine operator functions in Banach spaces*, Acta Scientiarum Matematicarum, 36 (1974), pp. 281-289.

[S.1] M. SOVA, *Cosine operator functions*, Rozprawy Mat., 49 (1966), pp. 3-46.

[T-W.1] C. C. TRAVIS AND G. F. WEBB, *Second order differential equations in Banach spaces*, in Nonlinear Equations in Abstracts Spaces, V. Lakshmikantham, ed., Academic Press, New York, 1978, pp. 331-361.

[V-J.1] R. B. VINTER AND T. JOHNSON, *Optimal control of nonsymmetric systems in N variables on the half space*, this Journal, 15 (1977), pp. 129-143.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 5, September 1986

ESTIMATION OF COEFFICIENTS AND BOUNDARY PARAMETERS IN HYPERBOLIC SYSTEMS*

H. T. BANKS† AND K. A. MURPHY‡

Abstract. We consider semi-discrete Galerkin approximation schemes in connection with inverse problems for the estimation of spatially varying coefficients and boundary condition parameters in second order hyperbolic systems typical of those arising in 1-D surface seismic problems. Spline based algorithms are proposed for which theoretical convergence results along with a representative sample of numerical findings are given.

Key words. hyperbolic systems, parameter estimation, spline approximations

AMS(MOS) subject classifications. 35R30, 41A15, 63N30

1. Introduction. In this paper we consider computational techniques for the following class of inverse problems: For the system

$$\rho(x)\frac{\partial^2 v}{\partial t^2}=\frac{\partial}{\partial x}\left(E(x)\frac{\partial v}{\partial x}\right), \quad t>0, \quad 0\leqq x\leqq 1, \tag{1.1}$$

$$\frac{\partial v}{\partial x}(t,0)+k_1 v(t,0)=s(t;\tilde{k}), \tag{1.2}$$

$$\frac{\partial v}{\partial t}(t,1)+k_2\frac{\partial v}{\partial x}(t,1)=0, \tag{1.3}$$

$$v(0,x)=\phi(x), \qquad v_t(0,x)=\psi(x), \tag{1.4}$$

given observations $\{\hat{y}_{ij}\}$ for $\{v(t_i, x_j)\}$, choose, from some admissible set, "best" estimates for the parameters ρ, E, k_1, k_2, $\tilde{k}$. These problems are motivated by certain versions of the so-called "1-D Seismic Inversion Problem" (see, e.g. [1], [9]). Roughly speaking, one has an elastic medium (e.g., the earth) with density ρ and elastic modulus E. A perturbation of the system (explosions, or vibrating loads from specially designed trucks) near the surface ($x=0$) produces a source s for particle disturbances v that travel as elastic waves, being partially reflected due to the inhomogeneous nature of the medium. An important but difficult problem involves using the observed disturbances at the surface or at points along a "bore hole" to determine properties (represented by parameters in the system) of the medium. In the highly idealized 1-D "surface seismic" problem, one assumes that data are collected at the same point ($x=0$) where the original disturbance or "source" is located. In addition to this hypothesis which cannot be true, other unrealistic special assumptions are made about the nature of the traveling and reflected waves. Although the standard 1-D formulations are far from reality, exploration seismologists have developed techniques for processing actual field data (performing a series of experiments and "stacking" the data) so that the 1-D

* Received by the editors March 9, 1984, and in revised form June 10, 1985. This research was supported in part by the National Science Foundation under grant MCS-8205355, the Air Force Office of Scientific Research under contract 81-0198, and the Army Research Office under contract ARO-DAAG-29-83-K0029.

† Lefschetz Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. Parts of this research were carried out while this author was a visitor at the Institute for Computer Applications in Science and Engineering (ICASE), NASA Langley Research Center, Hampton, VA, which is operated under NASA contracts NAS1-16394 and NAS1-17130.

‡ Department of Mathematics, Southern Methodist University, Dallas, Texas 75275.

problems are generally accepted as useful and worthy subjects of investigation. Consequently, numerous papers (for some interesting references, see the bibliographies of [1], [9]) on the 1-D problems can be found in the research literature.

In many formulations of the seismic inverse problem, the medium is assumed to be the half-line $x>0$ (with $x=0$ the surface) while in others (especially some of those dealing with computational schemes) one finds the assumption of an artificial finite boundary (say at $x=1$) at which no downgoing waves are reflected (an "absorbing" boundary). While there are several ways to approximate such a condition in 2- or 3-dimensional problems (see [13], [23]), for the 1-D formulation this condition is embodied in a simple boundary condition of the form (1.3); here $k_2 \approx \sqrt{E(1)/\rho(1)}$ and one can view this boundary condition as resulting from factoring the wave equation (1.1) at $x=1$ and imposing the condition of "no upgoing waves" at $x=1$.

Equation (1.1) is a 1-D version of the equations for an isotropic elastic medium while (1.2) represents an "elastic" boundary condition at the surface $x=0$ (k_1 represents an elastic modulus for the restoring force produced by the medium).

As is the case in many inverse or "identification" problems, the problems described above tend to be ill-posed (including a computationally undesirable instability) unless careful restrictions are imposed on the admissible parameter class (for some discussions of these aspects, see [1], [11]). We shall not focus on this aspect here. Rather, the purpose of our presentation in this paper is to demonstrate the feasibility of a certain theoretical approach and certain approximations in developing computational schemes for problems in which there are (i) unknown boundary parameters and (ii) unknown spatially varying coefficients in the system equations. We are, to our knowledge, the first to develop a sound theoretical framework for problems such as those considered here. Certain technical difficulties arise when one includes unknown boundary parameters in the estimation problems and we demonstrate one means of successfully treating (both theoretically and computationally) these difficulties. Ideas for estimation of variable coefficients in parabolic equations were carefully developed in [6] and we show here that these same techniques can be readily employed to give a rather complete theory for efficient schemes for problems involving hyperbolic systems.

We choose the "1-D seismic inverse problem" involving (1.1)–(1.4) as a test example to exhibit the efficacy of our ideas. However the technical features and notions we present are of importance in a number of other applications. There are rather easily motivated and fundamental problems in dealing with large elastic structures (large space structures—e.g. beamlike structures with tip bodies) that involve estimation of boundary condition parameters. In these cases the models are often hybrid models with distributed system (Euler–Bernoulli, Timoshenko) state equations and ordinary differential equation boundary conditions (see, for example, [2], [10], [20], [22]). A second class of problems for which the techniques introduced in this paper have immediate use are related to bioturbation [8], [14]. This is the mixing of lake and deep-sea sediments by burrowing activities of organisms. Understanding of this phenomenon is fundamental to geologists in interpreting geologic records contained in sediment core samples. The best models to date involve parabolic state equations (for a nonuniform "mixing chamber") with unknown parameters in the boundary conditions describing the flux into and out of the chamber.

In our approach here we employ the Trotter–Kato theorem to obtain theoretical convergence results (assuming regularity of parameter sets to guarantee existence of solutions to the inverse problems) for spline approximation schemes for the states. Boundary parameter estimation is treated directly via mappings that iteratively change the parameter-dependent spline basis elements into "conforming" elements (i.e., ele-

ments which satisfy the appropriate boundary conditions). We deal only with estimation of regular spatially-varying coefficients in (1.1), where again splines are used for parameters in a secondary approximation. Estimation of discontinuous coefficients (including location of the discontinuities) in problems such as those that are the focus of our attention in this paper can be effectively treated theoretically and numerically in a framework similar to that here using, for example, tau-Legendre state approximation schemes [4].

We turn then to the estimation problem for (1.1)-(1.4). It is theoretically and numerically advantageous to deal with homogeneous boundary conditions by transforming the problem so that the source term s in (1.2) appears in the initial data and in a term in the state equation. We make the transformation $u = v + G$ where (here "$\cdot$" represents differentiation with respect to t)

$$G(t, x; q) = -\left(\frac{1}{k_1}\right)s(t; \tilde{k}) + \left(\frac{1}{k_1 k_2}\right)x^2(x-1)\dot{s}(t; \tilde{k})$$

and obtain the system

$$\begin{aligned} &q_1(x)\frac{\partial^2 u}{\partial t^2} = \frac{\partial}{\partial x}\left(q_2(x)\frac{\partial u}{\partial x}\right) + F(t, x; q), \\ &u_x(t, 0) + q_3 u(t, 0) = 0, \\ &u_t(t, 1) + q_4 u_x(t, 1) = 0, \\ &u(0, x) = \tilde{\phi}(x; q), \qquad u_t(0, x) = \tilde{\psi}(x; q). \end{aligned} \tag{1.5}$$

Here the forcing function F is given by

$$\begin{aligned} F(t, x; q) \equiv q_1(x)&\left\{-\left(\frac{1}{q_3}\right)\ddot{s}(t; \tilde{k}) + \left(\frac{1}{q_3 q_4}\right)x^2(x-1)\ddot{s}(t; \tilde{k})\right\} \\ &-\frac{\partial}{\partial x}\left\{q_2(x)\left(\frac{1}{q_3 q_4}\right)(3x^2 - 2x)\dot{s}(t; \tilde{k})\right\}, \end{aligned}$$

where here and throughout we adopt the notation $q = (q_1, q_2, q_3, q_4, \tilde{k})$ with $q_1 \equiv \rho$, $q_2 \equiv E$, $q_3 = k_1$, and $q_4 = k_2$. The transformed initial conditions have the form

$$\begin{aligned} \tilde{\phi}(x; q) &= \phi(x) - \left(\frac{1}{q_3}\right)s(0; \tilde{k}) + \left(\frac{1}{q_3 q_4}\right)x^2(x-1)\dot{s}(0; \tilde{k}), \\ \tilde{\psi}(x; q) &= \psi(x) - \left(\frac{1}{q_3}\right)\dot{s}(0; \tilde{k}) + \left(\frac{1}{q_3 q_4}\right)x^2(x-1)\ddot{s}(0; \tilde{k}). \end{aligned}$$

We assume henceforth that we have observations $\hat{y}_i = (\hat{y}_{i1}, \cdots, \hat{y}_{im})$, $i = 1, 2, \cdots, n$, corresponding to $w(t_i; q) = (u(t_i, x_1), \cdots, u(t_i, x_m))$ where u is the solution of (1.5). For a criterion in determining a best estimate $\hat{q}$ of the parameters we use a least-squares function

$$J(q) = \sum_{i=1}^{n} |\hat{y}_i - w(t_i; q)|^2 \tag{1.6}$$

which we seek to minimize as q ranges over some admissible parameter set Q. We remark that in the event our observations $\hat{\eta}_i = (\hat{\eta}_{i1}, \cdots, \hat{\eta}_{im})$ are for the original system (1.1)-(1.4), we may apply directly the theory and techniques of this paper by considering

in place of (1.6) the criterion

$$\tilde{J}(q)=\sum_{i=1}^{n}|\hat{\eta}_i+\tilde{G}(t_i;\,q)-w(t_i;\,q)|^2 \tag{1.7}$$

where $\tilde{G}(t_i;\,q)\equiv(G(t_i,x_1;\,q),\cdots,G(t_i,x_m;\,q))$.

We make some standing assumptions to facilitate consideration of our problem in subsequent discussions. We shall search for q in a set $Q\subset C(0,1)\times H^1(0,1)\times R\times R\times R^k$ (we shall sometimes write Q as $Q_1\times Q_2\times Q_3\times Q_4\times Q_5$). We further assume that Q is compact in the $C\times H^1\times R^{2+k}$ topology, and that there exist positive constants $\underline{q}_i$, $\bar{q}_i$, $i=1,2,3,4$ such that

$$\begin{aligned}
&\underline{q}_i\leqq q_i(x)\leqq\bar{q}_i \quad \text{for } q_i\in Q_i, \quad i=1,2,\\
&\underline{q}_3\leqq -q_3\leqq\bar{q}_3 \quad \text{for } q_3\in Q_3, \quad \text{and}\\
&\underline{q}_4\leqq q_4\leqq\bar{q}_4 \quad \text{for } q_4\in Q_4.
\end{aligned}$$

Finally, we assume $\phi\in H^1(0,1)$, $\psi\in H^0(0,1)$, and $s(\cdot\,;\tilde{k})\in H^3(0,T)$ for each $\tilde{k}\in Q_5$, where $t_i\in[0,T]$, $T<\infty$, and that $\tilde{k}\to s(\cdot\,;\tilde{k})$ is a continuous mapping from Q_5 to $H^3(0,T)$.

We turn next to the theoretical foundations of the approximation schemes we propose to use in solving our inverse problem of minimizing J over Q, subject to (1.5).

2. Abstract formulation. The object in this section is to lay the theoretical foundation for the problem. First, we shall write our partial differential equation as an abstract ordinary differential equation in a Hilbert space, then determine a set of approximating ordinary differential equations. Each of these abstract equations will have an associated identification problem; the original will be referred to as (ID), the Nth approximating problem will be referred to as (ID^N). We shall use the theory of semigroups to obtain existence and uniqueness of solutions to the differential equations. We can then fit our problem into the theoretical framework developed in [5], and deduce that, under conditions stated there (reiterated below for clarity), one can solve (ID^N) for each N, and these parameter estimates thus obtained will "lead to" a solution of (ID).

The equation (1.5) can be rewritten as a first order system, motivating the use of a product $(V(q)\times L^2(q))$ of two spaces to be our Hilbert space $X(q)$.

Define $V(q)$ to be $H^1(0,1)$ with inner product defined by $\langle v,w\rangle_{V(q)}=\int_0^1 q_2DvDw\,dx-q_2(0)q_3v(0)w(0)$. ($D$ denotes the spatial differentiation operator $\partial/\partial x$.) It can be readily shown that for any $q\in Q$, $V(q)$ is a Hilbert space, and moreover, the assumptions made about Q imply that the $V(q)$ norm is uniformly equivalent to the H^1 norm as q ranges over Q. Let $V_B(q)$ contain those elements of $V(q)$ which satisfy the elastic boundary condition, i.e., $V_B(q)=\{v\in V(q)\cap H^2(0,1)\,|\,Dv(0)+q_3v(0)=0\}$.

We define $L^2(q)$ to be $H^0(0,1)$ with inner product given by $\langle v,w\rangle_{0,q}=\int_0^1 q_1vw\,dx$, and note that for each $q\in Q$, $L^2(q)$ is a Hilbert space and its norm is uniformly equivalent to the standard H^0 norm as q ranges over Q.

As described earlier, we take $X(q)=V(q)\times L^2(q)$ with inner product given by $\langle x,y\rangle_q=\langle x_1,y_1\rangle_{V(q)}+\langle x_2,y_2\rangle_{0,q}$ (where $x=(x_1,x_2)^T$ and $y=(y_1,y_2)^T$). It is clear from our remarks above that for $q\in Q$, $X(q)$ is a Hilbert space, and the X norm is uniformly equivalent to the $H^1\times H^0$ norm as q ranges over Q. We can formally write (1.5) as an abstract equation in $X(q)$:

$$\begin{aligned}
&\dot{z}(t)=A(q)z(t)+G(t;\,q),\\
&z(0)=z_0(q),
\end{aligned} \tag{2.1}$$

where we have identified $z(t) \in X(q)$ with $\binom{u(t,\cdot)}{u_t(t,\cdot)}$. The boundary conditions are incorporated into the domain of $A(q)$ by defining $\operatorname{dom} A(q) = \{\binom{u}{v} \in V_B(q) \times H^1(0,1) \mid v(1) + q_4 Du(1) = 0\}$, and A is the unbounded linear operator given by

$$A(q) = \begin{pmatrix} 0 & I \\ (1/q_1)D(q_2 D) & 0 \end{pmatrix}.$$

The function G and the initial condition are given by

$$G(t; q) = \begin{pmatrix} 0 \\ F(t, \cdot; q) \end{pmatrix} \quad \text{and} \quad z_0(q) = \begin{pmatrix} \tilde{\phi}(\cdot; q) \\ \tilde{\psi}(\cdot; q) \end{pmatrix}.$$

It can be shown that for each $q \in Q$, $A(q)$ is the infinitesimal generator of a C_0-semigroup, $T(t; q)$ on $X(q)$, so that we have the existence of mild solutions to (2.1), given by

$$z(t; q) = T(t; q) z_0(q) + \int_0^t T(t-s; q) G(s; q)\, ds \tag{2.2}$$

with $z(\cdot; q) \in C(0, T; X(q))$. In this context, the inverse problem can be stated as:

(ID) Given observations $\hat{y} = \{\hat{y}_i\}_{i=1}^n$, minimize $J(z(\cdot; q), \hat{y})$ over $q \in Q$ subject to $z(\cdot; q)$ satisfying (2.2).

Here, $J(q) \equiv J(z(\cdot; q), \hat{y}) = \sum_{i=1}^n |\hat{y}_i - \xi(t_i, q)|^2$ where $\xi(t_i, q) = (z_1(t_i, x_1; q), \cdots, z_1(t_i, x_m; q))$ and z_1 denotes the first component of z.

To prove that for each q, $A(q)$ generates a C_0-semigroup, one can use the Lumer-Phillips theorem [17, p. 16]. To employ this theorem, one must show the operator is dissipative, densely defined, and satisfies a certain range statement. To demonstrate the dissipativity of $A(q)$, we take $f \in \operatorname{dom} A(q)$, $q \in Q$, and compute (with an integration by parts)

$$\begin{aligned}
\langle A(q)f, f\rangle_q &= \left\langle \begin{pmatrix} f_2 \\ (1/q_1)D(q_2 Df_1) \end{pmatrix}, \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} \right\rangle_q \\
&= \langle f_2, f_1\rangle_{V(q)} + \langle (1/q_1)D(q_2 Df_1), f_2\rangle_{0,q} \\
&= \int_0^1 q_2 Df_1 Df_2\, dx - q_2(0)q_3 f_1(0)f_2(0) + \int_0^1 D(q_2 Df_1)f_2\, dx \\
&= -q_2(0)q_3 f_1(0)f_2(0) - q_2(0)Df_1(0)f_2(0) + q_2(1)Df_1(1)f_2(1) \\
&= -q_2(1)q_4(Df_1(1))^2 \leqq 0.
\end{aligned}$$

By relating $\operatorname{dom} A(q)$ to other subsets (see [16] for details) which are known to be dense in $H^1 \times H^0$, one can easily argue that for each $q \in Q$, $\operatorname{dom} A(q)$ is dense in $X(q)$. One can also argue that $\mathcal{R}(\lambda - A(q)) = X(q)$ for some $\lambda > 0$, by demonstrating that given $\binom{f_1}{f_2} \in X(q)$, there exists $\binom{u}{v} \in \operatorname{dom} A(q)$ such that

$$\begin{pmatrix} \lambda u - v \\ -(1/q_1)D(q_2 Du) + \lambda v \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}.$$

This is equivalent to solving the following two point boundary value problem:

$$-(1/q_1)D(q_2Du)+\lambda^2 u=\lambda f_1+f_2,$$
$$Du(0)+q_3u(0)=0,$$
$$\lambda u(1)+q_4Du(1)=f_1(1),$$

for $u\in H^2(0,1)$, and setting $v(x)=\lambda u(x)-f_1(x)$.

If we let $y=u-(1/q_4)x^2(x-1)f_1(1)$ the above problem is transformed to an equivalent one with homogeneous boundary conditions:

$$(-1/q_1)D(q_2Dy)+\lambda^2 y=F,$$
$$Dy(0)+q_3y(0)=0,$$
$$q_4Dy(1)+\lambda y(1)=0,$$

where $F\in L^2(q)$. One can then use the theory of self-adjoint operators (again see [16]) to argue that a solution exists for any $F\in L^2(q)$.

We now turn to the approximation of our equation (2.1). We shall obtain a solution z^N to an approximating equation (to be discussed in detail below) in a finite dimensional subspace of $X(q)$, denoted $X^N(q)$. Specifically, let $S^3(\Delta^N)$ represent the standard subspace of C^2 cubic splines corresponding to the partition $\Delta^N=\{x_i\}_{i=0}^N$, $x_i=i/N$ (see [18, pp. 78-81]); then, given $q\in Q$, we take $X^N(q)$ to be that subspace of $S^3(\Delta^N)\times S^3(\Delta^N)$ whose elements satisfy the boundary conditions corresponding to q (i.e., $X^N(q)\subset \operatorname{dom} A(q)$). Let B_j^N, $j=-1,\cdots,N+1$, be the B-spline basis elements for $S^3(\Delta^N)$. Then $X^N(q)$ is the $(2N+3)$-dimensional subspace spanned by the following set of basis functions:

$$\beta_1^N=\begin{pmatrix}\frac{4q_3}{N}B_{-1}^N+\left(3-\frac{q_3}{N}\right)B_0^N\\ 0\end{pmatrix},\qquad \beta_2^N=\begin{pmatrix}-\frac{4q_3}{N}B_1^N+\left(3+\frac{q_3}{N}\right)B_0^N\\ 0\end{pmatrix},$$

$$\beta_3^N=\begin{pmatrix}B_2^N\\ 0\end{pmatrix},\quad\cdots,\quad \beta_{N-1}^N=\begin{pmatrix}B_{N-2}^N\\ 0\end{pmatrix},$$

$$\beta_N^N=\begin{pmatrix}B_{N-1}^N\\ \frac{3Nq_4}{4}B_N^N\end{pmatrix},\qquad \beta_{N+1}^N=\begin{pmatrix}B_N^N\\ 0\end{pmatrix},\qquad \beta_{N+2}^N=\begin{pmatrix}B_{N+1}^N\\ -\frac{3Nq_4}{4}B_N^N\end{pmatrix}$$

$$\beta_{N+3}^N=\begin{pmatrix}-1/(3Nq_4)B_{N+1}^N\\ B_{N+1}^N\end{pmatrix},\qquad \beta_{N+4}^N=\begin{pmatrix}-1/(3Nq_4)B_{N+1}^N\\ B_{N-1}^N\end{pmatrix},$$

$$\beta_{N+5}^N=\begin{pmatrix}0\\ B_{N-2}^N\end{pmatrix},\quad\cdots,\quad \beta_{2N+1}^N=\begin{pmatrix}0\\ B_2^N\end{pmatrix},$$

$$\beta_{2N+2}^N=\begin{pmatrix}0\\ \frac{-4q_3}{N}B_1^N+\left(3+\frac{q_3}{N}\right)B_0^N\end{pmatrix},\qquad \beta_{2N+3}^N=\begin{pmatrix}0\\ \frac{4q_3}{N}B_{-1}^N+\left(3-\frac{q_3}{N}\right)B_0^N\end{pmatrix}.$$

Let $P^N(q):X(q)\to X^N(q)$ denote the orthogonal projection of $X(q)$ onto $X^N(q)$, i.e., given $f\in X(q)$, $P^N(q)f$ is that element in $X^N(q)$ which satisfies $|P^N(q)f-f|_q\leqq |g-f|_q$ for all $g\in X^N(q)$. For each $q\in Q$, we define an operator $A^N(q)$ on $X(q)$ given

by $A^N(q) = P^N(q)A(q)P^N(q)$, and then the approximating equation to (2.1) is written as:

$$\begin{aligned} \dot{z}^N(t) &= A^N(q)z^N(t) + P^N(q)G(t; q), \\ z^N(0) &= P^N(q)z_0(q), \end{aligned} \tag{2.3}$$

where $z^N(t) \in X^N(q)$. Using the fact that $A(q)$ is closed, $P^N(q)$ is bounded, and the Closed Graph Theorem, one finds that $A^N(q)$ is bounded. The operator $A^N(q)$ inherits the dissipativity of $A(q)$, and therefore it follows that for each $q \in Q$, $A^N(q)$ is the infinitesimal generator of a C_0-semigroup of contractions $T^N(t; q)$ on $X(q)$. It is readily seen that $T^N(t; q)$ leaves $X^N(q)$ invariant. Thus, for each $q \in Q$ and each $N = 1, 2, \cdots$, there exists a unique mild solution $z^N(\cdot; q) \in C(0, T; X^N(q))$ of (2.3), which can be expressed as

$$z^N(t; q) = T^N(t; q)P^N(q)z_0(q) + \int_0^t T^N(t-s; q)P^N(q)G(s; q)\, ds. \tag{2.4}$$

The associated approximate identification problem is given by

(ID^N) Given observations $\hat{y} = \{\hat{y}_i\}_{i=1}^n$, minimize $J(z^N(\cdot; q), \hat{y})$ over $q \in Q$ subject to $z^N(\cdot; q)$ satisfying (2.4).

Here, $J^N(q) \equiv J(z^N(\cdot; q), \hat{y}) = \sum_{i=1}^n |\hat{y}_i - \xi^N(t_i, q)|^2$ where $\xi^N(t_i, q) = (z_1^N(t_i, x_1; q), \cdots, z_1^N(t_i, x_m; q))$ and z_1^N denotes the first component of z^N.

Since $X^N(q)$ is finite-dimensional, (2.3) is in fact a system of $2N+3$ ordinary differential equations, which can be solved using standard numerical packages. Similarly, there are numerical packages available to solve (ID^N), provided solutions exist and we have some computationally feasible representation for q_1 and q_2. A detailed description of our numerical implementation, including a discussion of possible representations of q_1 and q_2, will be deferred to subsequent sections. First, our concern is to determine under what conditions solutions of (ID^N) exist and how they relate to a solution of (ID). This is the subject of the next theorem, a slight modification of that given in [5, p. 820].

THEOREM 2.1. *Assume Q is compact in the $C \times H^1 \times R^{2+k}$ topology. If $q \to z_0(q)$, $q \to P^N(q)f$, $q \to T^N(t; q)f$, $f \in X = X(q)$ are continuous in this same Q-topology, with the latter uniformly in $t \in [0, T]$, then:*

(i) *There exists for each N a solution $\hat{q}^N$ of* (ID^N) *and the sequence $\{\hat{q}^N\}$ possesses a convergent subsequence $\hat{q}^{N_k} \to \hat{q}$.*

(ii) *If we further assume that, for any sequence $\{q^j\}$ in Q with $q^j \to \tilde{q}$, we have $|z^j(t; q^j) - z(t; \tilde{q})|_{q^j} \to 0$ as $j \to \infty$, uniformly in $t \in [0, T]$, then $\hat{q}$ is a solution of* (ID).

The reader may, at first glance, find the convergence statement of (ii) suspect in that $z^j(t; q^j) \in X^j(q^j)$ and $z(t; \tilde{q}) \in X(\tilde{q})$, but this statement is meaningful in view of the following observation. In defining the spaces $V(q)$, $L^2(q)$, and $X(q)$, it was noted that $V(q)$, $L^2(q)$, and $X(q)$ are uniformly equivalent to H^1, H^0, and $H^1 \times H^0$, respectively, as q ranges over Q. This implies that the $X(q)$ are setwise equal as q ranges over Q. To be technically precise, we should use the canonical isomorphism when relating an element of $X(q^j)$ to its counterpart in $X(\tilde{q})$, but to simplify our presentation, we shall throughout abuse notation and omit the isomorphism.

It is easily seen from the form of $z_0(q)$ that $q \to z_0(q)$ is continuous. It is also true that for our $P^N(q)$, $T^N(t; q)$ we have $q \to P^N(q)f$ and $q \to T^N(t; q)f$ continuous; this will be readily seen from the matrix representations for our approximating scheme, and so further discussion is postponed until § 5.

The next theorem gives sufficient conditions for the hypothesis of (ii) from Theorem 2.1 to hold.

THEOREM 2.2. *Let q^N, $\tilde{q}$ be arbitrary in Q such that $q^N \to \tilde{q}$ as $N \to \infty$ (recall convergence is in the $C \times H^1 \times R^{2+k}$ topology). Suppose that the projections $P^N(q)$ are such that $|(P^N(q^N) - I)f|_{q^N} \to 0$ as $N \to \infty$ for all $f \in X(\tilde{q})$, that $f \in X(\tilde{q})$ implies $|T^N(t; q^N)f - T(t; \tilde{q})f|_{q^N} \to 0$ as $N \to \infty$, uniformly in $t \in [0, T]$, and that $|z_0(q^N) - z_0(\tilde{q})|_{q^N} \to 0$ as $N \to \infty$. Then the mild solutions $z^N(t; q^N)$ of (2.3) converge to the mild solution $z(t; \tilde{q})$ of (2.1) uniformly in $t \in [0, T]$.*

The proof of this theorem, which is based on a standard "variation-of-constants" representation for solutions z and z^N in terms of the semigroups T and T^N, essentially follows immediately from [5, Thm. 3.1, p. 823]. One only needs to verify that our spaces, operators, etc. satisfy the conditions required in [5].

It is clear from the continuity of $q \to z_0(q)$ that $|z_0(q^N) - z_0(\tilde{q})|_{q^N} \to 0$ as $q^N \to \tilde{q}$. It remains only to show the convergence of the projections and the semigroups. The main result of the next section is the convergence of the semigroups; the convergence of the projections is obtained as an intermediate proposition. In summary then, at the end of the next section, we will be able to deduce from Theorem 2.2 that $z^N(t; q^N)$ converges to $z(t; \tilde{q})$ whenever $q^N \to \tilde{q}$, and hence by Theorem 2.1 we are assured that the sequence of iterates $\{\hat{q}^N\}$ we obtain by solving (ID^N), has a subsequence which converges to a solution, $\hat{q}$, of (ID).

3. Convergence arguments. This section will be devoted to establishing the result: For each convergent sequence $q^N \to \tilde{q}$ in Q, and for any $f \in X(\tilde{q})$, $|T^N(t; q^N)f - T(t; \tilde{q})f|_{q^N} \to 0$ as $N \to \infty$, uniformly in $t \in [0, T]$. As explained in the previous section, this convergence result is crucial in arguing that $z^N(t; q^N) \to z(t; \tilde{q})$ whenever $q^N \to \tilde{q}$, which in turn is necessary to ensure that our candidate (the limit of our approximating subsequence) is indeed a solution to our inverse problem.

We shall first prove a slightly different form of convergence of the semigroups using the following version of the Trotter–Kato Theorem [3].

THEOREM 3.1. *Let $(\mathcal{B}, |\cdot|)$ and $(\mathcal{B}^N, |\cdot|_N)$, $N = 1, 2, \cdots$, be Banach spaces and let $\Pi^N : \mathcal{B} \to \mathcal{B}^N$ be bounded linear operators. Further assume that $T(t)$ and $T^N(t)$ are C_0-semigroups on $\mathcal{B}$ and $\mathcal{B}^N$ with infinitesimal generators $\tilde{A}$ and $\tilde{A}^N$, respectively. If*

(i) $\lim_{N\to\infty} |\Pi^N f|_N = |f|$ *for all* $f \in \mathcal{B}$,

(ii) *there exist constants M, ω independent of N such that $|T^N(t)|_N \leqq M e^{\omega t}$, for $t \geqq 0$,*

(iii) *there exists a set $\mathcal{D} \subset \mathcal{B}$, $\mathcal{D} \subset \mathrm{dom}(\tilde{A})$, with $\overline{(\lambda_0 - \tilde{A})\mathcal{D}} = \mathcal{B}$ for some $\lambda_0 > 0$, such that for all $f \in \mathcal{D}$ we have*

$$|\tilde{A}^N \Pi^N f - \Pi^N \tilde{A} f|_N \to 0 \quad \text{as } N \to \infty,$$

then $|T^N(t)\Pi^N f - \Pi^N T(t) f|_N \to 0$ as $N \to \infty$, for all $f \in \mathcal{B}$, uniformly in t on compact intervals in $[0, \infty)$.

It will be a standing assumption throughout this section that $q^N \to \tilde{q}$ in Q with this convergence in the $C \times H^1 \times R^{2+k}$ topology. Let $\mathcal{B} = X(\tilde{q})$ with norm denoted by $|\cdot|_{\tilde{q}}$, $\mathcal{B}^N = X(q^N)$ with norm $|\cdot|_{q^N}$ for $N = 1, 2, \cdots$, $\tilde{A} = A(\tilde{q})$ with corresponding semigroup $T(t) = T(t; \tilde{q})$, and $\tilde{A}^N = A^N(q^N) = P^N(q^N)A(q^N)P^N(q^N)$ with corresponding semigroup $T^N(t) = T^N(t; q^N)$ (as described in § 2). For each N, $\Pi^N : X(\tilde{q}) \to X(q^N)$ will be a bounded linear operator which will map elements of $\mathrm{dom}\, A(\tilde{q})$ into elements of $\mathrm{dom}\, A(q^N)$. Define

$$g^N(x) = \exp([\tilde{q}_3 - q_3^N]x) - (x^2/2)[\tilde{q}_3 - q_3^N] \exp[\tilde{q}_3 - q_3^N];$$

given $f=\binom{f_1}{f_2}$, let Π^N be defined by

$$\Pi^N f=\begin{pmatrix} g^N f_1 \\ (q_4^N/\tilde{q}_4)g^N f_2 \end{pmatrix}.$$

The functions g^N are defined so that as $N\to\infty$, $g^N(x)\to 1$, and $D^j(g^N(x))\to 0$ for any positive integer j, where in each case the convergence is uniform in $x\in[0,1]$.

A simple computation demonstrates that if $f\in \operatorname{dom} A(\tilde{q})$, then $\Pi^N f\in \operatorname{dom} A(q^N)$. For each N, Π^N is a bounded linear operator from $X(\tilde{q})$ to $X(q^N)$, but moreover, the set of operators $\{\Pi^N\}$ is uniformly bounded. This statement can be proved using the assumptions on Q and the properties of g^N mentioned above. Similar comments apply to the proof of our first proposition.

PROPOSITION 3.1. *For any* $f\in X(\tilde{q})$, $|\Pi^N f-f|_{q^N}\to 0$ *as* $N\to\infty$.

In order to argue the convergence of the infinitesimal generators, we shall need error estimates for the spline approximations and their derivatives. These will be variations of estimates such as those found in [21], modified to take into account our q-dependent norm, and the presence of the operator Π^N.

The following notation will be used throughout this section. Given a vector function f, we shall use f_i or $(f)_i$ to denote the ith component of f. Given the scalar function h, $I^N h$ will denote the standard cubic spline interpolant of h (thus $I^N h\in S^3(\Delta^N)$). For a vector function $f=\binom{f_1}{f_2}$, $I^N f$ will be the vector whose components are the spline interpolants of the components of f, i.e.,

$$I^N f=\begin{pmatrix} I^N f_1 \\ I^N f_2 \end{pmatrix} \quad \text{and} \quad I^N f\in S^3(\Delta^N)\times S^3(\Delta^N).$$

The interpolant of f which satisfies the boundary conditions corresponding to q will be written as $I_B^N(q)f$. While $I^N f$ interpolates f_1 and f_2 at the values $\{i/N\}_{i=0}^N$ and the derivatives of f_1 and f_2 at 0 and 1, $I_B^N(q)f$ will interpolate f_1 and f_2 at the values $\{i/N\}_{i=0}^N$, and will additionally satisfy

$$[D(I_B^N(q)f)_i](0)+q_3[(I_B^N(q)f)_i](0)=0, \quad \text{or equivalently,}$$
$$[D(I_B^N(q)f)_i](0)=-q_3 f_i(0) \quad \text{for } i=1,2,$$

and

$$[(I_B^N(q)f)_2](1)+q_4[D(I_B^N(q)f)_1](1)=0, \quad \text{or equivalently,}$$
$$[D(I_B^N(q)f)_1](1)=-(1/q_4)f_2(1).$$

We note that if f satisfies the boundary conditions involving q, then $I_B^N(q)f=I^N f$.

The first estimates involve cubic interpolants for scalar functions.

LEMMA 3.1. *If* $h\in H^2$, *then*

$$|D^2(h-I^N h)|_0\to 0 \quad \text{as } N\to\infty,$$
$$|D(h-I^N h)|_0\leqq N^{-1}|D^2(h-I^N h)|_0\leqq N^{-1}|D^2 h|_0,$$
$$|h-I^N h|_0\leqq N^{-2}|D^2(h-I^N h)|_0\leqq N^{-2}|D^2 h|_0.$$

The convergence statement of this lemma follows immediately from the density of H^3 in H^2, the estimates of [21, Thm. 6.9], and the first integral relation (4.15) of [21]. The estimates follow from (4.24) and (4.25), respectively, of [21] and the first integral relation.

One can use the results of Lemma 3.1 and the equivalence of the $X(q)$ and $H^1\times H^0$ norms to derive similar statements for the interpolants in the $X(q)$ norm.

LEMMA 3.2. *If* $f \in H^2 \times H^2$ *and* $q \in Q \subset C \times H^1 \times R^{2+k}$, *then*

$$|I^N f - f|_q \leqq K_1 N^{-1}(|D^2(f_1 - I^N f_1)|_0^2 + |D^2(f_2 - I^N f_2)|_0^2)^{1/2}$$
$$\leqq K_1 N^{-1}(|D^2 f_1|_0^2 + |D^2 f_2|_0^2)^{1/2},$$
$$|D(I^N f - f)|_q \leqq K_2(|D^2(f_1 - I^N f_1)|_0^2 + |D^2(f_2 - I^N f_2)|_0^2)^{1/2}$$

where K_1, K_2 *are constants which are independent of* f, q, *and* N.

Again, due to the equivalence of norms, the Schmidt inequality of [21, Thm. 1.5] can be modified and used component-wise to give a Schmidt type inequality in the $X(q)$ norm.

LEMMA 3.3. *If* $f \in S^3(\Delta^N) \times S^3(\Delta^N)$ *and* $q \in Q$, *then* $|Df|_q \leqq K_3 N |f|_q$, *where* K_3 *is a constant independent of* f, N, *and* q.

The preceding estimates can be used to establish convergence properties for the canonical projections $P^N(q^N)$ where $q^N \to \tilde{q}$ in Q.

PROPOSITION 3.2. *If* $f \in X(\tilde{q})$, *then*

$$|P^N(q^N) f - f|_{q^N} \to 0 \quad \text{as } N \to \infty.$$

Proof. First consider $f \in \operatorname{dom} A(\tilde{q}) \cap (H^2 \times H^2)$. For such f, $\Pi^N f \in \operatorname{dom} A(q^N) \cap (H^2 \times H^2)$ and $I_B^N(q^N)\Pi^N f = I^N \Pi^N f$. We use Lemma 3.2 in the triangle inequalities below to derive

$$|P^N(q^N) f - f|_{q^N} \leqq |P^N(q^N)[f - \Pi^N f]|_{q^N} + |P^N(q^N)\Pi^N f - \Pi^N f|_{q^N} + |\Pi^N f - f|_{q^N}$$
$$\leqq 2|\Pi^N f - f|_{q^N} + |I_B^N(q^N)\Pi^N f - \Pi^N f|_{q^N}$$
$$= 2|\Pi^N f - f|_{q^N} + |I^N \Pi^N f - \Pi^N f|_{q^N}$$
$$\leqq 2|\Pi^N f - f|_{q^N} + K_1 N^{-1}(|D^2(\Pi^N f)_1|_0^2 + |D^2(\Pi^N f)_2|_0^2)^{1/2}.$$

Thus we have $|P^N(q^N) f - f|_{q^N}$ bounded by terms which we can show converge to zero using Proposition 3.1 and the properties of g^N.

The $P^N(q^N)$ are uniformly bounded, and the set $\operatorname{dom} A(\tilde{q}) \cap (H^2 \times H^2)$ is dense in $X(\tilde{q})$, hence one can use standard arguments to conclude that the statement of the proposition holds for all $f \in X(\tilde{q})$.

PROPOSITION 3.3. *For each* $f \in X(\tilde{q})$, $|(P^N(q^N) - I)\Pi^N f|_{q^N} \to 0$ *as* $N \to \infty$, *and for each* $f \in \operatorname{dom} A(\tilde{q}) \cap (H^2 \times H^2)$, $|D[(P^N(q^N) - I)\Pi^N f]|_{q^N} \to 0$ *as* $N \to \infty$.

Proof. The first statement is proved within the proof of Proposition 3.2; specifically, it was shown that $|P^N(q^N)\Pi^N f - \Pi^N f|_{q^N} \leqq K_1 N^{-1}(|D^2(\Pi^N f)_1|_0^2 + |D^2(\Pi^N f)_2|_0^2)^{1/2}$.

The proof of the second statement is obtained from the following triangle inequality (here we also use Lemmas 3.3, 3.2):

$$|D(P^N(q^N)\Pi^N f - \Pi^N f)|_{q^N} \leqq |D(P^N(q^N)\Pi^N f - I^N(\Pi^N f))|_{q^N} + |D(I^N(\Pi^N f) - \Pi^N f)|_{q^N}$$
$$\leqq K_3 N |P^N(q^N)\Pi^N f - I^N(\Pi^N f)|_{q^N}$$
$$+ |D(I^N(\Pi^N f) - \Pi^N f)|_{q^N}$$
$$\leqq K_3 N |(P^N(q^N) - I)\Pi^N f|_{q^N} + K_3 N |\Pi^N f - I^N \Pi^N f|_{q^N}$$
$$+ |D[I^N(\Pi^N f) - \Pi^N f]|_{q^N}$$
$$\leqq 2K_3 N |I^N \Pi^N f - \Pi^N f|_{q^N} + |D[I^N \Pi^N f - \Pi^N f]|_{q^N}$$
$$\leqq (2K_1 K_3 + K_2)(|D^2[(\Pi^N f)_1 - I^N(\Pi^N f)_1]|_0^2$$
$$+ |D^2[(\Pi^N f)_2 - I^N(\Pi^N f)_2]|_0^2)^{1/2}.$$

Thus the conclusion $|D[(P^N(q^N)-I)\Pi^N f]|_{q^N}\to 0$ as $N\to\infty$ follows from the observation that for $i=1,2$

$$\begin{aligned}|D^2[I^N(\Pi^N f)_i-(\Pi^N f)_i]|_0 &\leqq |D^2[I^N((\Pi^N f)_i-f_i)]|_0\\ &\quad +|D^2[I^N f_i-f_i]|_0+|D^2[f_i-(\Pi^N f)_i]|_0\\ &\leqq 2|D^2[(\Pi^N f)_i-f_i]|_0+|D^2[I^N f_i-f_i]|_0,\end{aligned}$$

with the latter terms approaching zero because of the properties of g^N and Lemma 3.1, respectively.

In later arguments, it will be helpful to have bounds (in the H^1 and H^0 norms) on one component of an element of X in terms of a bound (in the $X(q)$ norm) on the entire element. Thus, we consider for $f\in X(q)$, $|f|_q^2=|f_1|_{V(q)}^2+|f_2|_{0,q}^2$ which is equivalent to $|Df_1|_0^2+|f_1|_0^2+|f_2|_0^2$, so that there exist constants k_1 and k_2 such that $|Df_1|_0^2\leqq k_1|f|_q^2$ and $|f_2|_0^2\leqq k_2|f|_q^2$. Similarly, $|Df|_q^2=|Df_1|_{V(q)}^2+|Df_2|_{0,q}^2$ which is equivalent to $|D^2f_1|_0^2+|Df_1|_0^2+|Df_2|_0^2$ so we infer the existence of constants k_3 and k_4 such that $|D^2f_1|_0^2\leqq k_3|Df|_q^2$ and $|Df_2|_0^2\leqq k_4|Df|_q^2$. For future reference, we combine and label these observations as

$$\begin{aligned}&|Df_1|_0^2\leqq k_1|f|_q^2,\\ &|D^2f_1|_0^2\leqq k_3|Df|_q^2,\\ &|f_2|_1^2\leqq k_2|f|_q^2+k_4|Df|_q^2.\end{aligned}\tag{3.1}$$

It is now possible to state and prove the following convergence theorem.

THEOREM 3.2. *Suppose $q^N\to\tilde{q}$ in Q (convergence is in the $C\times H^1\times R^{2+k}$ topology). Then*

$$|T^N(t;q^N)\Pi^N f-\Pi^N T(t;\tilde{q})f|_{q^N}\to 0\quad \text{as } N\to\infty,$$

for all $f\in X(\tilde{q})$, uniformly in t on compact intervals in $[0,\infty)$.

Proof. The result is an immediate consequence of Theorem 3.1, once the hypotheses of that theorem have been shown to hold. Part (i) follows from Proposition 3.1, while part (ii) holds since $T^N(t;q)$ and $T(t;q)$ are contraction semigroups for each N and $q\in Q$. It remains only to verify (iii), for which we take $\mathcal{D}$ to be the set $\operatorname{dom} A(\tilde{q})\cap(H^2\times H^2)$. Let $f\in\mathcal{D}$. Then

$$\begin{aligned}|A^N(q^N)\Pi^N f-\Pi^N A(\tilde{q})f|_{q^N} &= |P^N(q^N)A(q^N)P^N(q^N)\Pi^N f-\Pi^N A(\tilde{q})f|_{q^N}\\ &\leqq |P^N(q^N)[A(q^N)P^N(q^N)\Pi^N f-\Pi^N A(\tilde{q})f]|_{q^N}\\ &\quad +|P^N(q^N)\Pi^N A(\tilde{q})f-\Pi^N A(\tilde{q})f|_{q^N}\\ &\leqq |A(q^N)P^N(q^N)\Pi^N f-\Pi^N A(\tilde{q})f|_{q^N}\\ &\quad +|(P^N(q^N)-I)\Pi^N A(\tilde{q})f|_{q^N}\\ &\equiv \varepsilon_1(N)+\varepsilon_2(N).\end{aligned}$$

It follows directly from Proposition 3.3 that $\varepsilon_2(N)\to 0$ as $N\to\infty$. We must work harder to establish that $\varepsilon_1(N)\to 0$. We begin by breaking the norm into its two components

and treat each separately. Thus

$$
\begin{aligned}
[\varepsilon_1(N)]^2 &= \left| \begin{pmatrix} 0 & 1 \\ (1/q_1^N)D(q_2^N D) & 0 \end{pmatrix} (P^N(q^N)\Pi^N f) - \Pi^N \begin{pmatrix} 0 & 1 \\ (1/\tilde{q}_1)D(\tilde{q}_2 D) & 0 \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} \right|^2_{q^N} \\
&= |(P^N(q^N)\Pi^N f)_2 - g^N f_2|^2_{V(q^N)} \\
&\quad + |(1/q_1^N)D[q_2^N D(P^N(q^N)\Pi^N f)_1] - (q_4^N/\tilde{q}_4) g^N (1/\tilde{q}_1) D[\tilde{q}_2 D f_1]|^2_{0,q^N} \\
&\equiv [\delta_1(N)]^2 + [\delta_2(N)]^2.
\end{aligned}
$$

We first observe that

$$
\begin{aligned}
\delta_1(N) &\leqq |(P^N(q^N)\Pi^N f)_2 - (q_4^N/\tilde{q}_4) g^N f_2|_{V(q^N)} + |[(q_4^N/\tilde{q}_4) - 1] g^N f_2|_{V(q^N)} \\
&= |(P^N(q^N)\Pi^N f)_2 - (\Pi^N f)_2|_{V(q^N)} + |((q_4^N/\tilde{q}_4) - 1) g^N f_2|_{V(q^N)}.
\end{aligned}
$$

It is more convenient, and due to the equivalence of the norms, it is sufficient, to establish the convergence in the H^1 norm. This can easily be done for the first term by invoking Proposition 3.3 and the inequalities (3.1). An argument can be made for the second term based on the properties of the g^N and the convergence $q^N \to \tilde{q}$.

We turn now to the estimation of $\delta_2(N)$. Using the equivalence of the $L^2(q^N)$ and H^0 norms, and the inequalities (3.1), we establish the following chain of inequalities:

$$
\delta_2(N) = \left| \frac{1}{q_1^N} D[q_2^N D(P^N(q^N)\Pi^N f)_1] - \left(\frac{q_4^N}{\tilde{q}_4} g^N \right) \frac{1}{\tilde{q}_1} D(\tilde{q}_2 D f_1) \right|_{0,q^N}
$$

which is equivalent to

$$
\begin{aligned}
&\left| \frac{1}{q_1^N} q_2^N D^2(P^N(q^N)\Pi^N f)_1 + \frac{1}{q_1^N} D q_2^N D(P^N(q^N)\Pi^N f)_1 \right. \\
&\qquad \left. - \frac{q_4^N}{\tilde{q}_4} g^N \frac{1}{\tilde{q}_1} \tilde{q}_2 D^2 f_1 - \frac{q_4^N}{\tilde{q}_4} g^N \frac{1}{\tilde{q}_1} D\tilde{q}_2 D f_1 \right|_0 \\
&\leqq \left| \frac{q_2^N}{q_1^N} D^2(P^N(q^N)\Pi^N f)_1 - \frac{\tilde{q}_2}{\tilde{q}_1} \left(\frac{q_4^N}{\tilde{q}_4} g^N \right) D^2 f_1 \right|_0 \\
&\quad + \left| \frac{D q_2^N}{q_1^N} D(P^N(q^N)\Pi^N f)_1 - \frac{D\tilde{q}_2}{\tilde{q}_1} \left(\frac{q_4^N}{\tilde{q}_4} g^N \right) D f_1 \right|_0 \\
&\leqq \left| \frac{q_2^N}{q_1^N} D^2(P^N(q^N)\Pi^N f)_1 - \frac{q_2^N}{q_1^N} D^2(\Pi^N f)_1 \right|_0 + \left| \frac{q_2^N}{q_1^N} D^2(\Pi^N f)_1 - \frac{\tilde{q}_2}{\tilde{q}_1} \left(\frac{q_4^N}{\tilde{q}_4} \right) g^N D^2 f_1 \right|_0 \\
&\quad + \left| \frac{D q_2^N}{q_1^N} D(P^N(q^N)\Pi^N f)_1 - \frac{D q_2^N}{q_1^N} D(\Pi^N f)_1 \right|_0 \\
&\quad + \left| \frac{D q_2^N}{q_1^N} D(\Pi^N f)_1 - \frac{D\tilde{q}_2}{\tilde{q}_1} \left(\frac{q_4^N}{\tilde{q}_4} g^N \right) D f_1 \right|_0 \\
&\leqq \sqrt{k_3} \left| \frac{q_2^N}{q_1^N} \right|_\infty |D[(P^N(q^N) - I)\Pi^N f]|_{q^N} + \left| \frac{D q_2^N}{q_1^N} \right|_0 |D((P^N(q^N) - I)\Pi^N f)_1|_\infty \\
&\quad + \left| \frac{q_2^N}{q_1^N} D^2(\Pi^N f)_1 - \frac{\tilde{q}_2}{\tilde{q}_1} \left(\frac{q_4^N}{\tilde{q}_4} \right) g^N D^2 f_1 \right|_0 + \left| \frac{D q_2^N}{q_1^N} D(\Pi^N f)_1 - \frac{D\tilde{q}_2}{\tilde{q}_1} \left(\frac{q_4^N}{\tilde{q}_4} g^N \right) D f_1 \right|_0 .
\end{aligned}
$$

We thus see that $\delta_2(N)$ can be bounded by four terms which go to zero as $N \to \infty$; the convergence of the first two terms is the result of Proposition 3.3 and the convergence of q^N to $\tilde{q}$, while the convergence of the second two can be argued using the properties of g^N and $q^N \to \tilde{q}$.

We can use this theorem, the convergence properties of the operators Π^N (Proposition 3.1), and the semigroup properties of T^N and T, to establish the final result we need, as a corollary.

COROLLARY 3.1. *Suppose* $q^N \to \tilde{q}$. *Then*

$$|T^N(t; q^N)f - T(t; \tilde{q})f|_{q^N} \to 0 \quad \text{as } N \to \infty$$

for all $f \in X(\tilde{q})$, *uniformly in* t *on compact intervals in* $[0, \infty)$.

We can now invoke the results (see Theorems 2.1 and 2.2) stated in § 2 to conclude that $\hat{q}$ (obtained there as the limit of an approximating subsequence, $\{\hat{q}^{N_k}\}$) is a solution to the identification problem.

4. Parameter approximation. In § 2, we pose the problem of minimizing $J^N(q)$ over Q. The arguments underlying Theorem 2.1 yield that (under certain assumptions) each Nth (approximate) problem has a solution $\hat{q}^N$, and for any convergent subsequence $\{\hat{q}^{N_k}\}$, with $\hat{q}^{N_k} \to \hat{q}$, we have $\hat{q}$ is a solution of the original identification problem. Recall, however, that q_1 and q_2 are functional coefficients, and hence each of the approximate optimization problems is in fact infinite-dimensional in nature. In this section, we discuss some methods for approximating these infinite-dimensional optimization problems by finite-dimensional ones, thus providing numerically tractable problems. This, of course, results in a second, or parameter, approximation that must be considered.

In § 5, we shall present the results of several numerical test examples. To facilitate our presentation, we set $q_1 = \rho \equiv 1$ and search for $q_2 \equiv E$, q_3, q_4, $\tilde{k}$, with q_2 the only functional unknown. We therefore restrict our theoretical discussions here to this case. (We note however that in principle, our methods and ideas can be applied to the estimation of both ρ and E.)

An approach that one might take would be to assume an a priori parameterization for q_2. Thus the estimation of the unknown function becomes the estimation of a set of unknown constants appearing in the parameterization. The convergence theory developed thus far is directly applicable to this method. However, it would only yield results for best approximates (through the criterion on state observations) to q_2 *within the fixed a priori parameterization class.* Little can be said about convergence to a "best fit parameter" $\hat{q}_2$ from the original parameter set Q.

An alternate approach, which does not require qualitative (e.g., shape) assumptions about the parameter class, is to search for the unknown parameter in a sequence of sets Q^M which are finite dimensional approximations to the set Q. For example, one might search for the unknown parameter in sequences of classes of linear combinations of spline (or members of any other suitably chosen approximation family) basis elements. Such an approach was given a careful theoretical development for problems involving parabolic equations in [6] and elliptic equations in [15]. Since the basic ideas do not depend on the particular type (e.g. parabolic, hyperbolic, elliptic) of system, they are easily adapted to the problems under investigation here. We therefore shall only sketch the pertinent results, referring the reader to [6], [15] for further details and to [21] for the necessary technical estimates.

We shall consider here two cases: Q^M as a set of linear spline interpolants, and Q^M as a set of cubic spline interpolants. For both cases we need to generalize the

theory developed in § 2, since we now have a "double index" (reflecting approximations for both the parameter and the state space) sequence of iterates, which we would like to argue converges to a solution of the original identification problem.

To be specific, let $Q = Q_2 \times Q_3 \times Q_4 \times Q_5 \subset H^1 \times R^{2+k}$, and assume we have a mapping $i^M: Q_2 \to H^1$. For I the identity map, define $\mathcal{I}^M = i^M \times (I)^{2+k}$, i.e., for $q \in Q$, we have $\mathcal{I}^M(q) = (i^M(q_2), q_3, q_4, q_5)$.

Let $Q^M = \mathcal{I}^M(Q)$. We assume

(4.1) The set $(Q^M)_2 \equiv i^M(Q_2)$ is compact in H^1.

(4.2) For $q_2 \in Q_2$, $i^M(q_2) \to q_2$ in H^1 as $M \to \infty$, and this convergence is uniform in $q_2 \in Q_2$.

The original set Q is assumed to be compact in $H^1 \times R^{2+k}$, so it follows from (4.1), the definition of $\mathcal{I}^M$, and Theorem 2.1 that for each N and M, a solution $\hat{q}_M^N$ exists to the problem of minimizing J^N over Q^M. From the definition $Q^M = \mathcal{I}^M(Q)$, we see that there exists $\bar{q}_M^N \in Q$ such that $\mathcal{I}^M(\bar{q}_M^N) = \hat{q}_M^N$ for each N and M. But the compactness of the original set Q then implies the existence of some subsequence $\{\bar{q}_{M_k}^{N_j}\}$ and an element $\hat{q} \in Q$ such that $\bar{q}_{M_k}^{N_j} \to \hat{q}$ in Q; moreover, this subsequence may be chosen so that both $N_j \to \infty$ and $M_k \to \infty$. The limit $\hat{q}$ is in fact a solution to the problem of minimizing J over Q; this claim is verified as follows: From the definition $\hat{q}_{M_k}^{N_j}$ we have

$$J^{N_j}(\hat{q}_{M_k}^{N_j}) \leqq J^{N_j}(q) \quad \text{for } q \in Q^{M_k}.$$

This implies

$$J^{N_j}(\hat{q}_{M_k}^{N_j}) \leqq J^{N_j}(\mathcal{I}^{M_k}(q)) \quad \text{for } q \in Q. \tag{4.3}$$

But $|\hat{q}_{M_k}^{N_j} - \hat{q}| \leqq |\mathcal{I}^{M_k}(\bar{q}_{M_k}^{N_j}) - \bar{q}_{M_k}^{N_j}| + |\bar{q}_{M_k}^{N_j} - \hat{q}|$, and thus $\hat{q}_{M_k}^{N_j} \to \hat{q}$ in Q as $N_j \to \infty$, $M_k \to \infty$ follows from (4.2), the definition of $\mathcal{I}^{M_k}$, and $\bar{q}_{M_k}^{N_j} \to \hat{q}$. If we take the limit in (4.3) as N_j, $M_k \to \infty$, we see that $J(\hat{q}) \leqq J(q)$ for $q \in Q$. Here we have used Theorem 2.2 with the observation that the convergence statement $z^N(t; q^N) \to z(t; \tilde{q})$ for any $q^N \to \tilde{q}$ is still valid if replaced by $z^N(t; q^j) \to z(t; \tilde{q})$ as $j, N \to \infty$, for any $q^j \to \tilde{q}$; this can be seen using a reindexing argument. These remarks are summarized in the following theorem.

THEOREM 4.1. *Let $Q^M = \mathcal{I}^M(Q)$ where (4.1) and (4.2) are satisfied. Let $\hat{q}_M^N$ be a solution to the problem of minimizing J^N over Q^M. Then for any convergent subsequence $\{\hat{q}_{M_k}^{N_j}\}$ with $N_j, M_k \to \infty$ and $\hat{q}_{M_k}^{N_j} \to \hat{q}$, the limit $\hat{q}$ is a solution to the problem of minimizing J over Q.*

We first consider the above results applied to the case where the Q^M are sets of linear spline interpolants. Let $S^1(\Delta^M)$ represent the subspace of piecewise linear splines corresponding to the partition $\Delta^M = \{x_i\}_{i=0}^M$, $x_i = i/M$, and let $i^M: H^1 \to S^1(\Delta^M)$ denote the standard linear spline interpolating operator. If, in addition to assuming Q_2 is compact in H^1, we assume Q_2 satisfies $Q_2 \subset \{q_2 \in H^2 \,|\, |D^2 q_2|_0 \leqq K\}$, then it is not difficult to show that (4.1) and (4.2) are true for Q^M and i^M as defined above. From a standard representation result for linear interpolating splines [21, p. 12], we infer the continuity of the operator i^M as a mapping from H^1 to H^1, and the compactness of $(Q^M)_2 = i^M(Q_2)$ in H^1 follows immediately. To establish (4.2) we appeal to standard estimates such as [21, (2.17) and (2.18)]. Having verified (4.1) and (4.2), we now state

THEOREM 4.2. *Suppose $Q = Q_2 \times Q_3 \times Q_4 \times Q_5$ is a compact subset of $H^1 \times R^{2+k}$ with Q_2 additionally satisfying $Q_2 \subset \{q_2 \in H^2 \,|\, |D^2 q_2|_0 \leqq K\}$. Let $Q^M \in \mathcal{I}^M(Q)$ where $\mathcal{I}^M \equiv i^M \times (I)^{2+k}$ and i^M is the linear spline interpolating operator. If $\hat{q}_M^N$ represents a solution obtained from minimizing J^N over Q^M, then for any subsequence $\{\hat{q}_{M_k}^{N_j}\}$ of $\{\hat{q}_M^N\}$ such that as $N_j, M_k \to \infty$, $\hat{q}_{M_k}^{N_j} \to \hat{q}$ in Q, we have that $\hat{q}$ is a minimizer for J over Q.*

Under slightly stronger assumptions on the set Q, we can develop a similar convergence result using cubic spline approximations to q_2. Let $S^3(\Delta^M)$ be the subspace of C^2 cubic splines corresponding to the partition Δ^M, and let $i^M: C^1 \to S^3(\Delta^M)$ denote the standard cubic spline interpolating operator (see §§ 2 and 3 for details). We assume Q_2 is a compact subset of C^1 satisfying also $Q_2 \subset \{q_2 \in H^2 \,||D^2 q_2|_0 \leqq K\}$. We again may use standard interpolating spline representations (see [21, p. 45]) to conclude that i^M is a continuous operator from C^1 to H^1, from whence it follows that $(Q^M)_2$ is compact in H^1. To verify (4.2), we again refer to (4.19) and (4.20) in [21]. Thus we have the following theorem.

THEOREM 4.3. *Suppose $Q = Q_2 \times Q_3 \times Q_4 \times Q_5$ is a compact subset of $C^1 \times R^{2+k}$ with $Q_2 \subset \{q_2 \in H^2 \,||D^2 q_2|_0 \leqq K\}$. Let $Q^M = \mathcal{I}^M(Q)$ where $\mathcal{I}^M \equiv i^M \times (I)^{2+k}$, and i^M is the cubic spline interpolating operator. If $\hat{q}_M^N$ represents a solution obtained from minimizing J^N over Q^M, then there exists $\hat{q} \in Q$ which minimizes J over Q, and a subsequence $\{\hat{q}_{M_k}^{N_j}\}$ of $\{\hat{q}_M^N\}$ such that as $N_j, M_k \to \infty$, $\hat{q}_{M_k}^{N_j} \to \hat{q}$.*

In the next section we present numerical findings for double (state and parameter) approximation schemes such as those described here.

5. Numerical implementation and examples. Recall from § 2 that the approximating identification problem is:

Given $\hat{y}$, minimize $J^N(q) = \sum_{i=1}^n |\hat{y}_i - \xi^N(t_i, q)|^2$ over $q \in Q$ (where ξ^N involves point evaluations, in space, of the first component of z^N) subject to $z^N(\cdot\,; q)$ satisfying the following ordinary differential equation:

$$\dot{z}^N(t) = A^N(q) z^N(t) + P^N(q) G(t; q),$$
$$z^N(0) = P^N(q) z_0(q).$$

(We continue our discussions in terms of the transformed system (1.5) and criterion (1.6) even though the numerical examples summarized in this section involve "data" for the original system (1.1)-(1.4) used in conjunction with the criterion (1.7).) Since $z^N \in X^N(q)$, z^N has a representation in terms of the basis elements of $X^N(q)$, $z^N(t; q) = \sum_{i=1}^{2N+3} w_i^N(t; q) \beta_i^N(x; q)$. If we let $[A^N(q)]$ and $[f^N]$ be the matrix and vector representations, respectively of $A^N(q)$ and $P_N(q)f$ (where f is an arbitrary function in $X(q)$) with respect to the basis elements of $X^N(q)$, and let $w^N(t; q) \equiv \text{col}\,(w_1^N(t; q), \cdots, w_{2N+3}^N(t; q))$, then $w^N(t; q)$ solves the following system of ordinary differential equations:

$$\dot{w}^N(t; q) = [A^N(q)] w^N(t; q) + [G^N(t; q)],$$
$$w^N(0; q) = [z_0^N(q)].$$

As in [5], this can be written more explicitly as:

$$Q^N \dot{w}^N(t; q) = K^N w^N(t; q) + R^N G(t; q), \tag{5.1}$$
$$Q^N w^N(0; q) = R^N z_0(q),$$

where Q^N and K^N are matrices, with elements described by $(Q^N)_{i,j} = \langle \beta_i^N, \beta_j^N \rangle_q$, $(K^N)_{i,j} = \langle \beta_i^N, A(q)\beta_j^N \rangle_q$, and $(R^N f)_i = \langle \beta_i^N, f \rangle_q$ for $f \in X(q)$. Due to the form of the B-spline basis elements we have chosen (see § 2), Q^N can be stored as a banded symmetric matrix; this banded, symmetric structure permits more efficient computations and requires less storage space. The matrix K^N has a similar sparse (although not symmetric) structure.

Each element of the matrices Q^N and K^N, and of the vector $R^N f$ depends continuously on q, therefore the representations $[A^N(q)]$ and $[f^N]$ are continuous in q. The basis elements for $X^N(q)$ depend linearly on q, and hence are continuous in q, which implies $q \to P^N(q)f$ and $q \to T^N(t; q)f$ (we note $T^N(t; q) = \exp(A^N(q)t)$ since $A^N(q)$ is a bounded operator) are continuous mappings (recall this was a necessary condition in Theorem 2.1).

We note that in the case where q_1 and q_2 are assumed to be constant, or to have a representation as, for example, a linear combination of spline elements, then the computations can be done more efficiently; in such cases, the numerical quadratures required to compute the inner products which form Q^N and K^N need be performed only once for each N. Then, to construct Q^N and K^N the appropriate multiples or linear combinations of these stored values are computed.

Many of the computations in the software package used to generate the following examples were done with IMSL subroutines (for example, the optimization, and the solution of the differential equation in (5.1)). Although much modification was necessary for the present application, the core of the package was developed by James Crowley [12]. The examples were computed either on an IBM VM/370, or a CDC 6600.

The optimization is done using a Levenberg-Marquardt algorithm. For fixed N, each iteration in the optimization is performed as follows. Given q, beginning at time zero ($t_1 = 0$), a Cholesky decomposition method is used to solve (5.1) for $\dot{w}^N(t; q)$ and $w^N(t_1; q)$; this is then integrated using Gear's method to obtain $w^N(t_2; q)$. We use the components of the vector $w^N(t_2; q)$ to recover $z_1^N(t_2; q)$ as the linear combination of the first components of the basis elements. The vector $\xi^N(t_2, q)$ is $z_1^N(t_2; q)$ evaluated at each of the spatial observation points. Using $w^N(t_2; q)$ as the initial value, (5.1) is solved again for $t \in [t_2, t_3]$, $\xi^N(t_3, q)$ is obtained, and this procedure is repeated until $\xi^N(t_i, q)$ has been evaluated at all times t_i; then $J^N(q)$ can be computed as the sum of the residuals, $|\hat{y}_i - \xi^N(t_i, q)|^2$. The data $\{\hat{y}_i\}$ is read in and stored at the beginning.

In the selection of examples to follow, the "data" has been generated with an independent finite difference scheme (an implicit method [19] was modified for our boundary conditions and the variable coefficient, $q_2(x)$) applied to the model with a priori chosen "true" values q^* of the parameters. In all examples, $q_1(x)$ is taken to be identically one (this is done to reduce ill-posedness, as mentioned in § 1). We begin each example with an initial guess, and a value of N; we solve (ID^N), to get converged values, $\bar{q}^N$ (these are numerical approximations (to $\hat{q}^N$) that result from the Levenberg-Marquardt algorithm), which we then use as starting values for the next value of N. So, in Example 5.1 (below) we begin with $N = 4$ and a guess q^0, and generate $\bar{q}^4$. We then start with $\bar{q}^4$ at $N = 8$, and generate $\bar{q}^8$.

The examples we present here are simple and are chosen mainly to illustrate how the scheme we have discussed theoretically performs when one is attempting to estimate boundary parameters or spatially varying elastic modulus. We have carried out numerous other numerical experiments to test the efficacy of the ideas. For example, we have found the algorithms to be rather robust with regard to the initial parameter guess q^0. That is, the integrity of the method is preserved in examples similar to Examples 5.1, 5.2 even though one starts with initial guesses which are as much as 200% in error from the "true" values. The optimization algorithm requires more iterations, CPU time is increased, but we still obtain convergence to accurate estimates of the parameters. Moreover, our theoretical and computational considerations can be modified to treat more realistic problems in which one has to estimate discontinuities in the elastic modulus E. Some initial numerical findings on such problems are reported in [4].

We remind the reader that the computations reported on below were carried out using "data" for the system (1.1)-(1.4) with criterion (1.7) and an appropriate approximate criterion for the Nth problem. (We have also successfully tested the methods on similar examples with the transformed system (1.5) and criterion (1.6), although, of course, this is not the typical formulation of the inverse problem for which data will be available.)

Example 5.1. For our first example we used "data" consisting of observations at $x=0$ and times $t=.25, .5, .75, \cdots, 2.0$. This is meant to simulate the situation in "surface seismic" experiments where only data at the surface are available. The source term was chosen as $s(t;\tilde{k})=q_5(1-e^{-5t})\,e^{q_6 t}$, a function which rises to a peak quickly and then gradually diminishes to zero; again this attempts to mimic the situation in seismic experiments. We assume vanishing initial conditions and seek to estimate a constant elastic modulus q_2 as well as the boundary parameters q_3, q_4 and the source parameters $\tilde{k}=(q_5, q_6)$. True values along with our estimates are given in the results summarized in Table 5.1. Graphs comparing the true solution at the surface $u(t, 0; q^*)$ with the approximate solution $u^N(t, 0; \bar{q}^N)$ are shown in Fig. 5.1. We also tested the method on this example using "data" for more spatial observations (data at $x=0, .5, 1.0$ and $t=.5, 1.0, 1.5$) with our findings given in Table 5.2. Based on these computations and a number of other tests, we suggest that there appears to be little difficulty with our method in the case where only one spatial observation is available as long as a sufficient number of time observations are available.

TABLE 5.1

	Converged values		
Initial guess	$N=4$	$N=8$	True values
$q_2^0=2.0$	$\bar{q}_2^4=2.96001$	$\bar{q}_2^8=3.0001$	$q_2^*=3.0$
$q_3^0=-1.0$	$\bar{q}_3^4=-1.98861$	$\bar{q}_3^8=-1.99012$	$q_3^*=-2.0$
$q_4^0=2.0$	$\bar{q}_4^4=0.97428$	$\bar{q}_4^8=1.00683$	$q_4^*=1.0$
$q_5^0=1.5$	$\bar{q}_5^4=1.97135$	$\bar{q}_5^8=1.99809$	$q_5^*=2.0$
$q_6^0=-0.5$	$\bar{q}_6^4=-0.98500$	$\bar{q}_6^8=-1.00506$	$q_6^*=-1.0$
No. of iterations[1]	11	2	
R.S.S.[2]	0.659×10^{-5}	0.119×10^{-5}	
CPU[3]	125.363	84.688	

[1] Number of iterations in the optimization algorithm.
[2] Residual sum of squares $=J^N(\bar{q}^N)$.
[3] The CPU time given in seconds.

Example 5.2. In this example we compared the performance of our method on problems with "noisy data" with that on those without noise in the data. We used the same source term as that in Example 5.1, zero initial conditions, but a "true" parameterized elastic modulus $E(x)=\frac{3}{2}+1/\pi \arctan[q_{21}(x-q_{22})]$. Data for observations at $x=$ 0.0, 0.5, 1.0 and $t=.416, .832, 1.248, 1.664, 2.08, 2.496$ were used. Results for the case of data without noise are summarized in Table 5.3, while findings employing data with a noise level of approximately 3% are given in Table 5.4. In both cases, the method converges nicely but as one might expect, the converged values of the parameters do

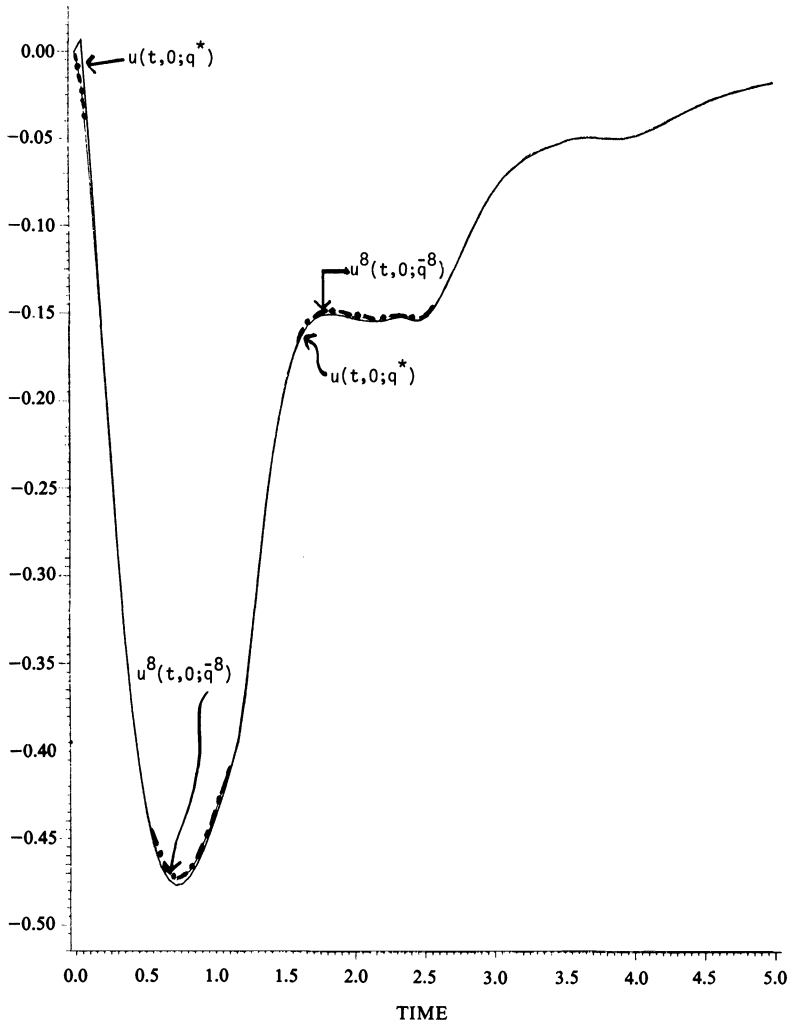


FIG. 5.1

TABLE 5.2

Initial guess	Converged values $N=4$	Converged values $N=8$	True values
$q_2^0=2.0$	$\bar{q}_2^4=2.98515$	$\bar{q}_2^8=2.99378$	$q_2^*=3.0$
$q_3^0=-1.0$	$\bar{q}_3^4=-1.92304$	$\bar{q}_3^8=-2.01999$	$q_3^*=-2.0$
$q_4^0=2.0$	$\bar{q}_4^4=1.01302$	$\bar{q}_4^8=1.00285$	$q_4^*=1.0$
$q_5^0=1.5$	$\bar{q}_5^4=1.97120$	$\bar{q}_5^8=2.00578$	$q_5^*=2.0$
$q_6^0=-0.5$	$\bar{q}_6^4=-1.03296$	$\bar{q}_6^8=-0.99172$	$q_6^*=-1.0$
No. of iterations	12	5	
R.S.S.	0.235×10^{-5}	0.441×10^{-5}	
CPU	117.597	147.323	

TABLE 5.3

	Converged values		
Initial guess	$N=4$	$N=8$	True values
$q_{21}^0=1.0$	$\bar{q}_{21}^4=2.97352$	$\bar{q}_{21}^8=2.99994$	$q_{21}^*=3.0$
$q_{22}^0=1.0$	$\bar{q}_{22}^4=0.51115$	$\bar{q}_{22}^8=0.50053$	$q_{22}^*=0.5$
$q_3^0=-2.0$	$\bar{q}_3^4=-0.99892$	$\bar{q}_3^8=-1.00026$	$q_3^*=-1.0$
$q_4^0=2.0$	$\bar{q}_4^4=3.05138$	$\bar{q}_4^8=3.01070$	$q_4^*=3.0$
$q_5^0=1.0$	$\bar{q}_5^4=2.00322$	$\bar{q}_5^8=2.00056$	$q_5^*=2.0$
$q_6^0=-2.0$	$\bar{q}_6^4=-1.01163$	$\bar{q}_6^8=-1.00217$	$q_6^*=-1.0$
No. of iterations	13	3	
R.S.S.	0.1025×10^{-3}	0.82859×10^{-5}	
CPU	269.696	196.335	

TABLE 5.4. (*Noisy data*).

	Converged values		
Initial guess	$N=4$	$N=8$	True values
$q_{21}^0=1.0$	$\bar{q}_{21}^4=3.30536$	$\bar{q}_{21}^8=3.29222$	$q_{21}^*=3.0$
$q_{22}^0=1.0$	$\bar{q}_{22}^4=0.53802$	$\bar{q}_{22}^8=0.53115$	$q_{22}^*=0.5$
$q_3^0=-2.0$	$\bar{q}_3^4=-0.86648$	$\bar{q}_3^8=-0.86017$	$q_3^*=-1.0$
$q_4^0=2.0$	$\bar{q}_4^4=2.99610$	$\bar{q}_4^8=2.96002$	$q_4^*=3.0$
$q_5^0=1.0$	$\bar{q}_5^4=2.09207$	$\bar{q}_5^8=2.09295$	$q_5^*=2.0$
$q_6^0=-2.0$	$\bar{q}_6^4=-1.15602$	$\bar{q}_6^8=-1.15571$	$q_6^*=-1.0$
No. of iterations	13	2	
R.S.S.	0.6509×10^{-3}	0.476×10^{-3}	
CPU	270.11	136.87	

not agree with the true parameters in the case of noisy data. In Figs. 5.2, 5.3, 5.4 and 5.5, we graphically depicted the curves for $\bar{E}^N$ and E^* in several cases.

Example 5.3. In this example we illustrate the ideas discussed in § 4 regarding parameter approximation in the set of linear and cubic splines. We do not assume an a priori shape for the elastic modulus $E(x)$, the "true" value of which is given by $E^*(x)=\frac{3}{2}+\tanh[6(x-.5)]$. Rather we first search for E in the class of linear spline approximations to E^*. We then carry out the search using cubic splines. Initial conditions are $u(0,x)=e^x$, $u_t(0,x)=-3\,e^x$ and no source term was assumed (i.e., $s\equiv 0$). Data for observations at 3 spatial points ($x=0.0$, 0.5, 1.0) and 6 time points ($t=.16, .32, \cdots, 1.0$) were used. Figure 5.6 depicts graphs of the true modulus E^*, the initial guess E^0, and the converged estimate $\bar{E}^4$ where we used linear splines (with 4 basis elements—$M=3$ in the notation of § 4) to approximate E and cubic splines ($N=4$) to approximate the state. At the same time we searched for the boundary parameters q_3, q_4 (true values $q_3^*=-1.0$, $q_4^*=3.0$) and obtained converged estimates $\bar{q}_3^4=-1.05425$, $\bar{q}_4^4=3.3576$ with a CPU time of 38 seconds and R.S.S. $=0.255\times10^{-2}$.

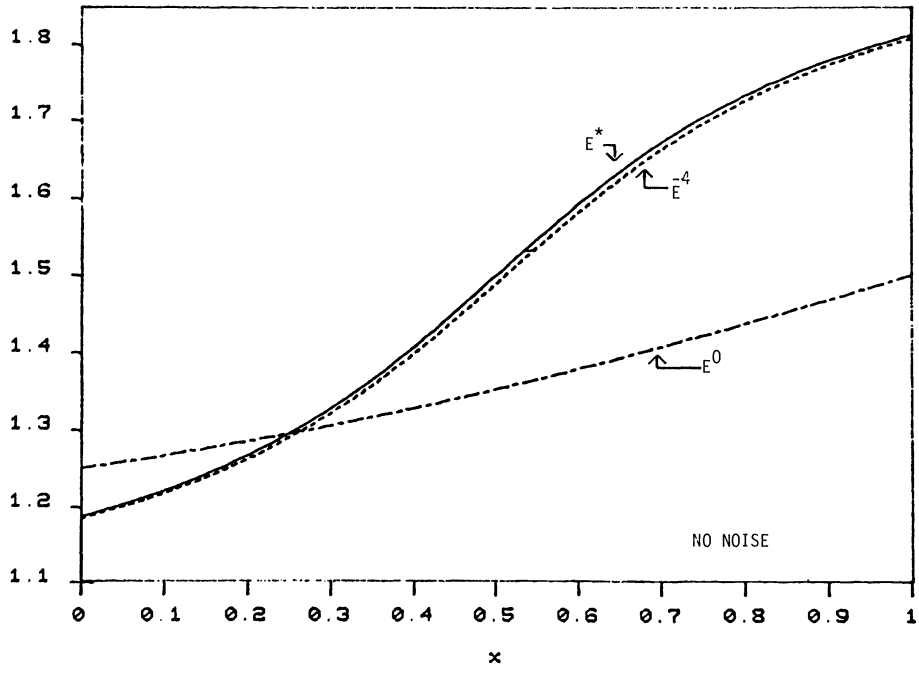

1.8
1.7
1.6
1.5
1.4
1.3
1.2
1.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
x
E*
E-4
E0
NO NOISE

FIG. 5.2

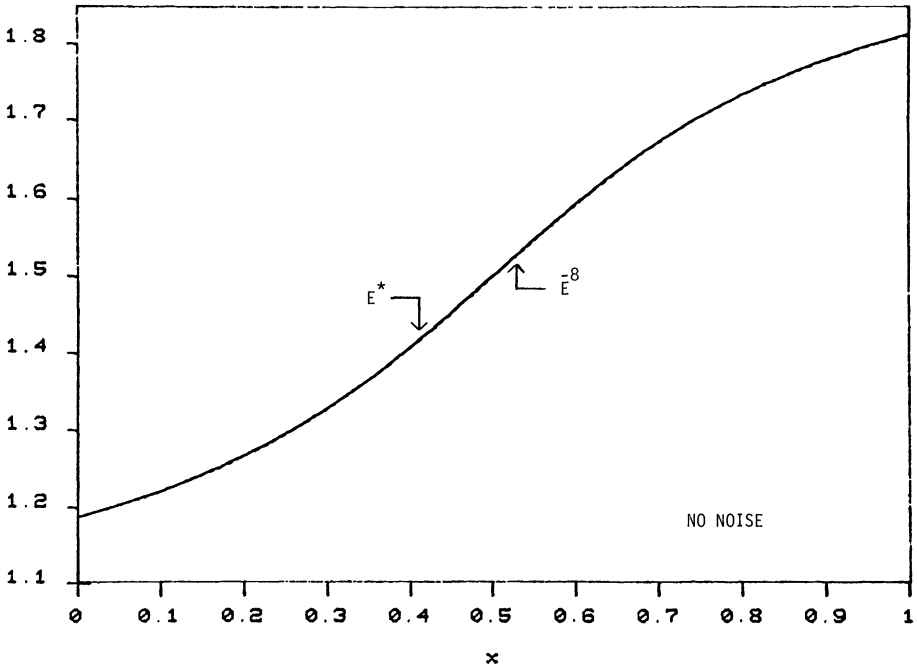

1.8
1.7
1.6
1.5
1.4
1.3
1.2
1.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
x
E*
E-8
NO NOISE

FIG. 5.3

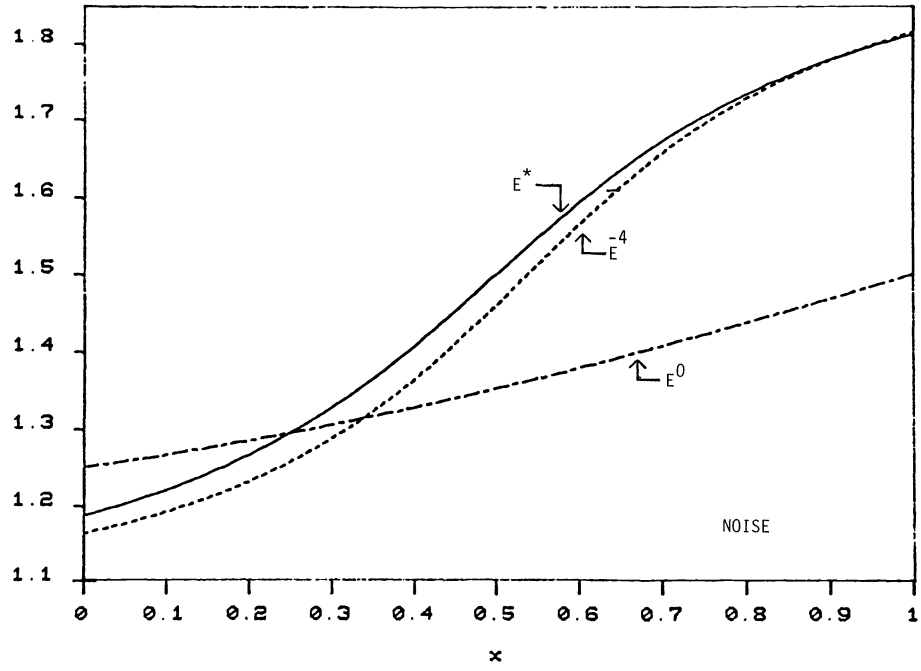

1.8
1.7
1.6
1.5
1.4
1.3
1.2
1.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
x
E*
E^-4
E^0
NOISE

FIG. 5.4

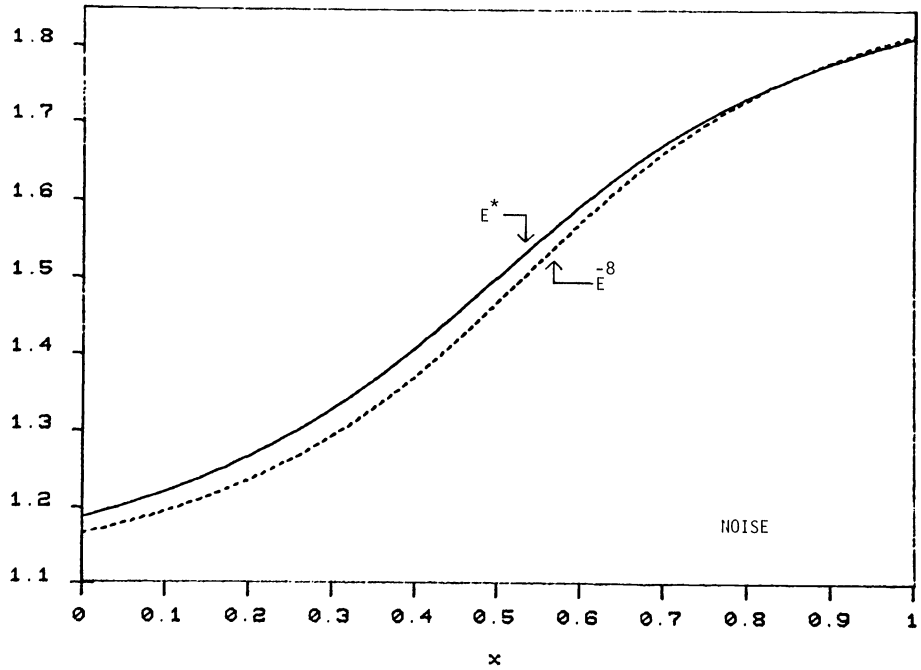

1.8
1.7
1.6
1.5
1.4
1.3
1.2
1.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
x
E*
E^-8
NOISE

FIG. 5.5

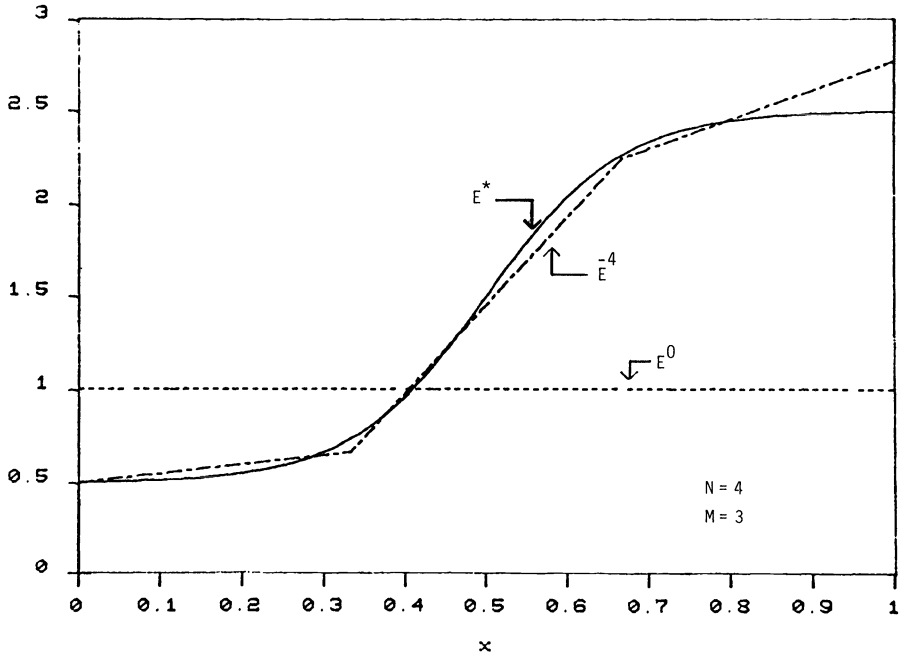

3
2.5
2
1.5
1
0.5
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
x
E*
E^-4
E^0
N = 4
M = 3

FIG. 5.6

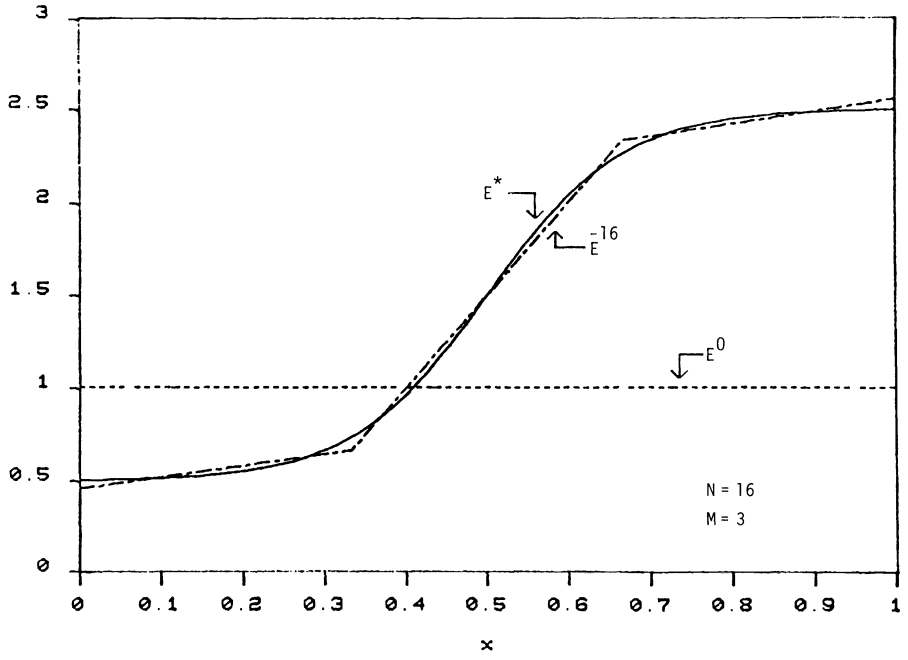

3
2.5
2
1.5
1
0.5
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
x
E*
E^-16
E^0
N = 16
M = 3

FIG. 5.7

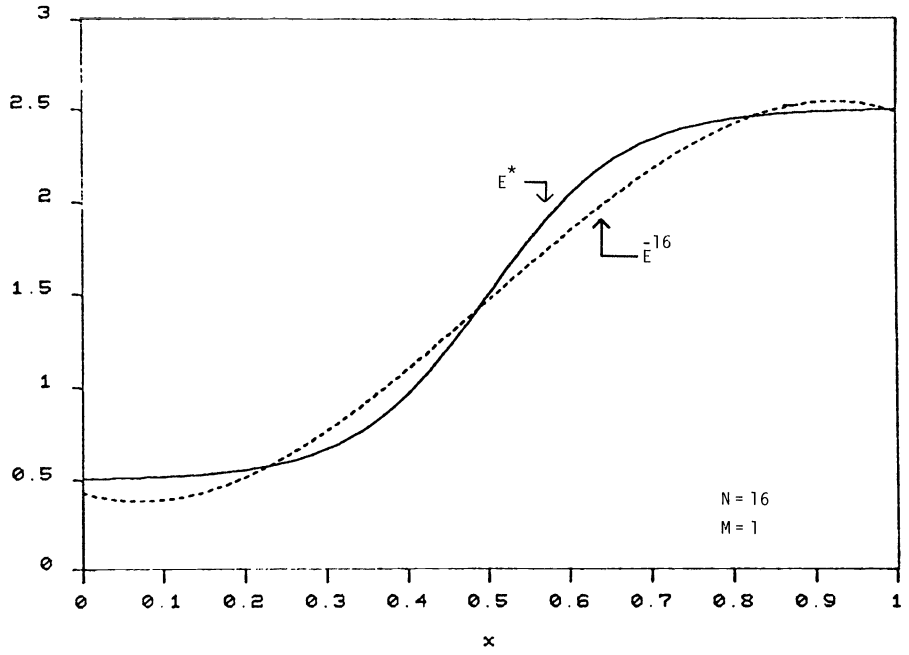


FIG. 5.8

Figure 5.7 contains graphs similar to those in Fig. 5.6 except $N = 16$ was used in the state approximations. Boundary parameter estimates corresponding to $\bar{E}^{16}$ were $\bar{q}_3^{16} = -1.10063$, $\bar{q}_4^{16} = 3.07049$ with CPU time of 118 seconds and R.S.S. $= 0.472 \times 10^{-4}$. The error (in the H^0 norm) in estimating E^* in each case was calculated to be $|E^* - \bar{E}^4| = .081$ and $|E^* - \bar{E}^{16}| = .030$.

We carried out similar calculations for the same example in which we employed cubic splines ($M = 1$ in the notation of § 4, i.e. 4 basis elements) for the parameter approximations. The graphs of E^*, E^0 and $\bar{E}^{16}$ are compared in Fig. 5.8. In this second test we did not search on the boundary parameters q_3, q_4 but rather held them fixed at their "true" values. The error at the converged parameter was $|E^* - \bar{E}^{16}| = .109$, with R.S.S. $= 0.293 \times 10^{-2}$ and a CPU time of 178 seconds.

6. Concluding remarks. We have presented in this paper both theoretical and numerical results using some of our ideas involving spline approximations for inverse or parameter estimation problems for hyperbolic systems. Among the novel features is the capability of estimating variable coefficients and boundary parameters with methods that are both theoretically sound and readily implementable. Our techniques (reported on earlier, [7]) involve the use of parameter dependent basis elements for the approximation subspaces in a Galerkin type semidiscrete scheme.

While we have focused on 1-dimensional space domain problems here, our ideas are in principle applicable to problems in 2- and 3-dimensional domains. We have devoted some thought to such problems in connection with use of basis elements that are tensor products of 1-D elements. These ideas offer some promise, given the parallelism that would be inherent in the resulting algorithms and given the emerging technology related to supercomputers and array processors. However, there are other ideas that also offer great promise; in particular, there are those involving spectral

methods such as the tau-Legendre for which we have reported preliminary findings in [4]. A fundamental difference between these techniques and those proposed in this paper is that in the tau-Legendre one does not require the approximation subspace basis elements to satisfy the boundary conditions. Instead the boundary conditions are essentially imposed as side constraints adjoined to the Galerkin type differential equations. This can offer significant computational advantages, especially in higher dimensional domain problems. We are currently pursuing investigations of these ideas.

The assumption of compactness on Q in our presentation was used to guarantee convergence results within our theoretical framework. It can also be used to establish stability results (i.e., a type of continuous dependence of parameter estimates $\hat{q}$ on the observations $\{\hat{y}_i\}$). When specifying a typical parameter set Q, this compactness requirement will often be manifested in terms of functional parameter constraints (e.g. bounds on certain derivatives) which can be important with regard to implementation. Thus, while we did not find it necessary to impose any such constraints in the algorithms used in producing the numerical results discussed in § 5 (this, we conjecture, was because the examples we considered here were relatively simple and rather well behaved—e.g. no other local extremals near the "true" values q^*), it is sometimes necessary to take them into consideration to obtain acceptable numerical performance. We have carried out numerical experiments with other problems for which this is a critical point, and satisfactory numerical performance will be guaranteed only if a constrained optimization procedure is employed. The compactness-related constraints can and should be implemented in certain cases. We are currently investigating further these questions and will report on our findings elsewhere. Here we just observe that the "compactness of Q" is an important computational as well as theoretical feature of our approach.

In closing we remark that the theoretical results presented above only guarantee convergence of subsequences $\{\hat{q}^{N_k}\}$ to a minimizer $\hat{q}$ for J. But for the class of problems investigated here and for a number of other types of inverse problems we have studied, we have in practice observed (numerically) convergence of the original sequence $\{\hat{q}^N\}$. This has been our experience even in examples with noisy data and may be due in many cases to the fact that the original problem of minimizing J over Q has a unique solution $\hat{q}$. In this situation, elementary and quite standard arguments can be employed to actually establish convergence of $\{\hat{q}^N\}$ itself to $\hat{q}$.

Acknowledgments. The authors would like to express their sincere appreciation to G. Moeckel (Mobil Oil Co.), R. Ewing (U. Wyoming), and K. Kunisch (U. Graz) for stimulating discussions during the course of some of the work reported above. They are also grateful for the support and hospitality received during their visit at Southern Methodist University where a substantial portion of the investigations reported on here were carried out.

REFERENCES

[1] A. BAMBERGER, G. CHAVENT AND P. LAILLY, *About the stability of the inverse problem in* 1-D *wave equations—application to the interpretation of seismic profiles*, Appl. Math. Optim., 5 (1979), pp. 1-47.

[2] H. T. BANKS AND J. M. CROWLEY, *Parameter estimation for distributed systems arising in elasticity*, Proc. Symposium on Engineering Sciences and Mechanics, National Cheng Kung University, Tainan, Taiwan, Dec. 28-31, 1981, pp. 158-177; LCDS Tech. Rep. 81-24, November, 1981, Brown University.

[3] H. T. BANKS, J. M. CROWLEY AND K. KUNISCH, *Cubic spline approximation techniques for parameter estimation in distributed systems*, LCDS Tech. Rep. 81-25, November, 1981, Brown University; IEEE Trans. Automat. Control, AC-28 (1983), pp. 773-786.

[4] H. T. BANKS, K. ITO AND K. A. MURPHY, *Computational methods for estimation of parameters in hyperbolic systems*, in Conference on Inverse Scattering: Theory and Application, J. B. Bednar, et al., eds., Society for Industrial and Applied Mathematics, Philadelphia, 1983, pp. 181-193.

[5] 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.

[6] H. T. BANKS AND P. DANIEL LAMM, *Estimation of variable coefficients in parabolic distributed systems*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 386-398.

[7] H. T. BANKS AND K. A. MURPHY, *Inverse problems for hyperbolic systems with unknown boundary parameters*, in Control Theory for Distributed Parameter Systems and Applications, F. Kappel, et al., eds., Springer-Verlag, Berlin, 1983, pp. 35-44.

[8] H. T. BANKS AND I. G. ROSEN, *Fully discrete approximation methods for the estimation of parabolic systems and boundary parameters*, LCDS Tech. Rep. 84-19, Brown University, 1984; Acta Applic. Math., to appear.

[9] K. P. BUBE AND R. BURRIDGE, *The one-dimensional inverse problem of reflection seismology*, SIAM Rev., 25 (1983), pp. 497-559.

[10] J. A. BURNS AND E. M. CLIFF, *An approximation technique for the control and identification of hybrid systems*, in Dynamics and Control of Large Flexible Spacecraft, 3rd VPISU/AIAA Symposium, 1981, pp. 269-284.

[11] 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. 45-58.

[12] J. M. CROWLEY, *Numerical methods of parameter identification for problems arising in elasticity*, Ph.D. Thesis, Brown University, May, 1982.

[13] B. ENGQUIST AND A. MAJDA, *Absorbing boundary conditions for the numerical simulation of waves*, Math. Comp., 31 (1977), pp. 629-651.

[14] N. L. GUINASSO AND D. R. SCHINK, *Quantitative estimates of biological mixing rates in abyssal sediments*, J. Geophys. Res., 80 (1975), pp. 3032-3043.

[15] K. KUNISCH AND L. WHITE, *Parameter estimation for elliptic equations in multidimensional domains with point and flux observations*, Nonlinear Anal.: TMA, to appear.

[16] K. A. MURPHY, *A spline-based approximation method for inverse problems for a hyperbolic system including unknown boundary parameters*, Ph.D. Thesis, Brown University, May, 1983.

[17] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Applied Mathematical Sciences 44, Springer, New York, 1983.

[18] P. M. PRENTER, *Splines and Variational Methods*, Wiley-Interscience, New York, 1975.

[19] R. D. RICHTMYER AND K. W. MORTON, *Difference Methods for Initial-Value Problems*, Wiley-Interscience, New York, 1967.

[20] I. G. ROSEN, *A numerical scheme for the identification of hybrid systems describing the vibration of flexible beams with tip bodies*, Report CSDL-P-1983, Draper Lab, 1984; J. Math. Anal. Appl., to appear.

[21] M. H. SCHULTZ, *Spline Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1973.

[22] J. STORCH AND S. GATES, *Planar dynamics of a uniform beam with rigid bodies affixed to the ends*, CSDL-R-1629, Draper Labs, Cambridge, MA, May, 1983.

[23] E. TURKEL, *Numerical methods for large-scale time-dependent partial differential equations*, in Computational Fluid Dynamics, W. Kollman, ed., Hemisphere Publ., Washington, 1980, pp. 127-262.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 5, September 1986

004

GENERALIZED SOLUTIONS OF CONSTRAINED OPTIMIZATION PROBLEMS*

TOMÁŠ ROUBÍČEK†

Abstract. A constrained optimization problem on a uniform space X is considered. Under certain assumptions, the problem may be extended onto the completion of X with respect to the precompact modification of the original uniformity on X, which yields the generalized problem. It is shown that the solution of the perturbed classical problem (or of the penalized classical problem) converges to the solution of the generalized problem, which makes the generalized problem more convenient than the classical one. Finally, the general framework is demonstrated in special cases in normed linear spaces.

Key words. optimization, generalized solution, stability

1. Introduction and notation. In [6] the unconstrained optimization problem on a noncompact domain X was investigated with the aim to construct a "natural" extension of the domain, denoted $\bar{X}$, which guarantees the existence of the generalized solution (which is an element of $\bar{X}$) together with the stability of the set of the generalized solutions with respect to certain perturbations of the function to be minimized. The domain X was considered as a uniform space and the extended domain $\bar{X}$ was constructed as a completion of X with respect to another (admissible) uniformity. The existence and stability of the generalized solution is then guaranteed if the later uniformity is sufficiently coarse. Moreover, there exists such a uniformity which is optimal from the stability point of view.

In this paper the approach mentioned above is applied to a constrained optimization problem where the situation is more complicated than that in the unconstrained case. In this section the basic notation and some results from [6], required in what follows, are introduced. The "classical" constrained optimization problem, its perturbed version, the generalized problem, and basic relations between them are given in § 2. In the further sections, a characterization of the set of the generalized solutions using the level sets of the classical problem, some convergence results for the perturbed classical problem, and for the unconstrained problem arising from the use of the well-known penalty-function method are stated. The general framework exploiting the uniform-space theory, used in §§ 2–5, is somewhat unusual in optimization; in § 6 it is applied to two special cases, both employing the more usual normed-linear-space notation.

Now, we start with some definitions (see e.g. [1]). The filter $\mathcal{F}$ on a set M is a nonempty collection of the nonempty subsets of M such that $A \in \mathcal{F}$, $B \supset A \Rightarrow B \in \mathcal{F}$ and $A, B \in \mathcal{F} \Rightarrow A \cap B \in \mathcal{F}$. Let X be a completely regular Hausdorff topological space endowed with an admissible uniformity $\mathcal{U}_X$. It should be recalled that the uniformity $\mathcal{U}_X$ on X is a filter on $X \times X$ with the properties: $\forall U \in \mathcal{U}_X: \Delta \subset U$, $U^{-1} \in \mathcal{U}_X$ and $\exists V \in \mathcal{U}_X$, $V^2 \subset U$, where $\Delta = \{(x_1, x_2) \in X^2;\ x_1 = x_2\}$, $U^{-1} = \{(x_1, x_2) \in X^2;\ (x_2, x_1) \in U\}$ and $V^2 = \{(x_1, x_2) \in X^2;\ \exists x_3,\ (x_1, x_3) \in V,\ (x_3, x_2) \in V\}$. The elements of the uniformity are called entourages. The uniformity $\mathcal{U}_X$ is admissible iff it generates the topology of X. $\mathcal{U}_X$ is precompact iff $\forall U \in \mathcal{U}_X \exists$ a finite set $S \subset X$: $U(S) = X$, where $U(S) = \{x \in X;\ \exists x_1 \in S,\ (x, x_1) \in U\}$. Also, $\mathcal{U}_X$ is precompact iff the completion of X with

* Received by the editors October 3, 1984, and in revised form July 26, 1985.

† Institute of Information Theory and Automation, Czechoslovak Academy of Sciences, Pod vodárenskou věží 4, 182 08 Praha 8, Czechoslovakia. This paper was prepared while the author was at the General Computing Centre of the Czechoslavak Academy of Sciences.

respect to $\mathcal{U}_X$ is compact. The finest from the precompact admissible uniformities on X which are coarser than $\mathcal{U}_X$, denoted $\mathcal{U}_X^*$, is called the precompact modification of $\mathcal{U}_X$. $\mathcal{U}_X^*$ is also the coarsest admissible uniformity on X which makes uniformly continuous all the functions $X \to [0, 1]$ which are uniformly continuous with respect to $\mathcal{U}_X$. A typical situation in most applications is the case when X is a normed linear space (with the norm denoted $\|\cdot\|_X$) and $\mathcal{U}_X$ is generated by means of the base $\{V_\varepsilon, \varepsilon > 0\}$, where $V_\varepsilon = \{(x_1, x_2) \in X^2; \|x_1 - x_2\|_X < \varepsilon\}$.

In [6] the minimization problem for a l.s.c. (lower semicontinuous) function $f: X \to \bar{R}$ is investigated; $\bar{R}$ denotes the usual two-point compactification of the real line, i.e. $\bar{R} = R \cup \{-\infty, +\infty\}$. As the extended domain $\bar{X}$ the completion of X with respect to $\mathcal{U}_X^*$ is taken. Obviously, since $\bar{X}$ is compact, there is a unique admissible uniformity on $\bar{X}$, denoted $\mathcal{U}_{\bar{X}}$, the entourages of which will be denoted with the bar, e.g. $\bar{U}$. Also elements of $\bar{X}$ will be denoted with the bar, e.g. $\bar{x}$. Note that the trace of $\mathcal{U}_{\bar{X}}$ on X is just $\mathcal{U}_X^*$, especially, $\bar{U} \cap X^2 \in \mathcal{U}_X^*$ for $\bar{U} \in \mathcal{U}_{\bar{X}}$. We can naturally extend f on $\bar{X}$ (the extension will be denoted again by f without any confusion) by the formula $f(\bar{x}) = \liminf_{x \to \bar{x}, x \in X} f(x)$; $x \to \bar{x}$ means, of course, the convergence in $\bar{X}$. Clearly, we have $\inf f(\bar{X}) = \inf f(X)$. An element $\bar{x} \in \bar{X}$ is said to be the generalized solution of the considered minimization problem iff $f(\bar{x}) = \inf f(X)$. The (nonempty) set of the generalized solutions is stable with respect to the perturbations of f in certain topology on the space of the functions $X \to \bar{R}$. This topology is induced by the uniformity on the space of the epigraphs which can be constructed by means of the uniformity $\mathcal{U}_X \times \mathcal{U}_{\bar{R}}$; for details see [6]. Moreover, the uniformity $\mathcal{U}_X^*$ is the finest uniformity by means of which the stable set of the generalized solutions can be constructed for arbitrary function f, see [6, Thm. 3]. The uniformity $\mathcal{U}_X^*$ has also the important property: $\forall A \subset X$, $\forall U \in \mathcal{U}_X$, $\exists V \in \mathcal{U}_X^*$: $V(A) \subset U(A)$; in other words, for every $A \subset X$ and $U \in \mathcal{U}_X$ there exists a $\mathcal{U}_X$-uniformly continuous function $\psi: X \to [0, 1]$ such that $\psi(A) = 0$ and $\psi(X \setminus U(A)) = 1$. This property of $\mathcal{U}_X^*$ enables us to describe certain sets in $\bar{X}$ by means of the original uniformity $\mathcal{U}_X$, which is remarkable because the uniformity $\mathcal{U}_X^*$ and the extended domain $\bar{X}$ have been defined in a somewhat nonconstructive manner, see also Remark in § 3.

For $\bar{x} \in \bar{X}$ we denote $\mathcal{N}(\bar{x})$ the filter on X which is the trace of the neighbourhood filter of $\bar{x}$ in $\bar{X}$. $\mathcal{N}(\bar{x})$ is a maximal $\mathcal{U}_X$-round filter on X, and the mapping $\bar{x} \mapsto \mathcal{N}(\bar{x})$ defines a one-to-one correspondence between the points in $\bar{X}$ and the filters of the mentioned property. Recall that the filter $\mathcal{F}$ on X is said to be $\mathcal{U}_X$-round (see [2]) iff $\forall A \in \mathcal{F} \exists B \in \mathcal{F} \exists V \in \mathcal{U}_X$: $V(B) \subset A$. Of course, $\mathcal{N}(\bar{x})$, being maximal in the class of the $\mathcal{U}_X$-round filters on X, can be constructed generally by means of the axiom of choice only. Still other characterizations of $\mathcal{N}(\bar{x})$ can be stated (see [6, Thm. 4]) but they use the uniformity $\mathcal{U}_X^*$ instead of the axiom of choice.

For $M \subset \bar{X}$ denote $\mathcal{N}(M) = \cap \{\mathcal{N}(\bar{x}); \bar{x} \in M\}$. Clearly, $\mathcal{N}(M)$ is again a filter on X. Moreover, let M_1, M_2 be closed subsets of $\bar{X}$, then $M_1 \neq M_2$ implies $\mathcal{N}(M_1) \neq \mathcal{N}(M_2)$. Thus the closed subsets of $\bar{X}$ are characterized by their filters $\mathcal{N}(\cdot)$, which will be employed in Theorem 3.

2. The classical and the generalized optimization problems. In the case of the constrained optimization problem, very little may be said under the condition that the mapping to the constraints is only continuous as usual in the classical theory of optimization. Here, the uniform continuity will be exploited. First, we will treat a general framework, and afterwards, in § 6, we will apply the obtained results to more usual and detailed structures.

Let X, Y be the uniform Hausdorff spaces endowed with the uniformities $\mathcal{U}_X$, $\mathcal{U}_Y$, respectively; $f: X \to \bar{R}$ be a l.s.c. function; $F: X \to Y$ be a uniformly continuous mapping; and C be a subset of Y. We shall deal with the classical optimization problem, denoted P_C:

$$P_C \quad \begin{array}{l} \text{minimize } f(x), \\ \text{subject to } x \in X, \qquad F(x) \in C. \end{array}$$

As usual, we define the admissible set $S_{ad}P_C = \{x \in X;\ F(x) \in C\}$, the value $\inf P_C = \inf f(S_{ad}P_C)$, and the set of the classical solutions $\operatorname{Arginf} P_C = \{x \in S_{ad}P_C;\ f(x) = \inf P_C\}$. Also we use the set of the approximate classical solutions $\operatorname{Arginf}_\varepsilon P_C = \{x \in S_{ad}P_C;\ f(x) \leqq \inf P_C + \varepsilon\}$, with $\varepsilon > 0$. From the "practical" or "engineering" point of view, some perturbed problems are very interesting. We employ the perturbations of the set C, namely in the form $\tilde{C} = V(C)$, $V \in \mathcal{U}_Y$. For the perturbed problem $P_{\tilde{C}}$ thus obtained, we can again define $S_{ad}P_{\tilde{C}}$, $\inf P_{\tilde{C}}$, and $\operatorname{Arginf}_\varepsilon P_{\tilde{C}}$ by replacing C by $\tilde{C}$ in the corresponding definitions for the problem P_C.

In most problems of "technical" origin, it suffices to find a solution from $\operatorname{Arginf}_\varepsilon P_{\tilde{C}}$, because such solution is "almost" optimal and, at the same time, the constraints are fulfilled with accuracy prescribed in advance, which is realistic in the technical problems. Similar approach has been already used in the book of J. Warga [7], in which, however, a specific structure of the optimal control problems on certain function spaces has been employed (our classical problems correspond to those which are named "original" in [7], and the generalized problems introduced below correspond roughly to the "extended" problems). There is, however, a fundamental difference between these two mentioned approaches because the extended problem in [7] is obtained by exploiting a topology which is weaker than the initial one, while our generalized problem will be obtained using the initial topology only.

Since the sets $\operatorname{Arginf}_\varepsilon P_{\tilde{C}}$ are of "technical" importance, there is naturally a question whether there exists a limit (in certain sense) when $\varepsilon \to 0$ and $\tilde{C} \to C$. Note that $\operatorname{Arginf}_\varepsilon P_{\tilde{C}}$ generally does not converge to $\operatorname{Arginf} P_C$. Thus we are motivated to introduce a generalized problem which is, in fact, a more natural setting of the classical problem P_C. To construct this generalized problem, denoted GP_C, we extend f and F on the space $\bar{X}$ (= the completion of X with respect to $\mathcal{U}_X^*$). The extension of f has been already defined in § 1. It is easy to see that F, being uniformly continuous with respect to $\mathcal{U}_X$ and $\mathcal{U}_Y$, is uniformly continuous with respect to $\mathcal{U}_X^*$ and $\mathcal{U}_Y^*$ as well. Denoting $\bar{Y}$ the completion of Y with respect to $\mathcal{U}_Y^*$, we may extend $F: \bar{X} \to \bar{Y}$ (the extended mapping is denoted again by F) in such a manner that F is continuous on $\bar{X}$. As X is dense in $\bar{X}$, this extension is unique. It should be emphasized that $\bar{X}$, $\bar{Y}$, $\bar{R}$ are compact, which together with the continuity of F and l.s.c. of f yields the base for most of the following results.

We introduce the generalized problem:

$$GP_C \quad \begin{array}{l} \text{minimize } f(\bar{x}) \\ \text{subject to } \bar{x} \in \bar{X}, \qquad F(\bar{x}) \in \bar{C} = \operatorname{cl}_{\bar{Y}} C, \end{array}$$

$\operatorname{cl}_{\bar{Y}} C$ denotes the closure of C in $\bar{Y}$. For this problem we define the admissible set, the value and the set of the generalized solutions by a straightforward manner, i.e. $S_{ad}GP_C = \{\bar{x} \in \bar{X};\ F(\bar{x}) \in \bar{C}\}$, $\inf GP_C = \inf f(S_{ad}GP_C)$ and $\operatorname{Arginf} GP_C = \{\bar{x} \in S_{ad}GP_C;\ f(\bar{x}) = \inf GP_C\}$.

The optimization problem is said to be nontrivial iff its value is less than $+\infty$. Especially, the nontrivial problem has a nonempty admissible set. Basic relationship between the optimization problems defined above is given in the following theorem.

THEOREM 1. *Let P_C be nontrivial, $\tilde{C} = V(C)$, $V \in \mathcal{U}_Y$, $\varepsilon > 0$. Then $P_{\tilde{C}}$ and GP_C are nontrivial, too*; $\operatorname{Arginf}_\varepsilon P_{\tilde{C}}$ *and* $\operatorname{Arginf} GP_C$ *are nonempty*; $\inf P_C \geqq \inf GP_C \geqq \inf P_{\tilde{C}}$; *and*

$$X \cap \operatorname{Arginf} GP_C = \begin{cases} \varnothing & \text{if } \inf P_C > \inf GP_C, \\ \operatorname{Arginf} P_{\mathrm{cl}_Y C} & \text{if } \inf P_C = \inf GP_C. \end{cases}$$

Proof. Clearly, $P_{\tilde{C}}$ and GP_C are nontrivial and $\operatorname{Arginf}_\varepsilon P_{\tilde{C}}$ is nonempty. $\operatorname{Arginf} GP_C$ is nonempty, as well, because f is l.s.c. on a compact set $S_{ad}GP_C$. Also the inequality $\inf P_C \geqq \inf GP_C$ is obvious. To prove $\inf GP_C \geqq \inf P_{\tilde{C}}$, we consider $\bar{x} \in \operatorname{Arginf} GP_C$ and $\bar{V} \in \mathcal{U}_{\bar{Y}}$ such that $V^2(C) \subset \tilde{C}$ with $V = \bar{V} \cap Y^2$. Since F is continuous on $\bar{X}$, there is $\bar{U} \in \mathcal{U}_{\bar{X}}$ such that $F(\bar{U}(\bar{x})) \subset \bar{V}(\bar{C})$. At the same time, $\inf f(X \cap \bar{U}(\bar{x})) \leqq f(\bar{x}) = \inf GP_C$. Thus for any $\varepsilon > 0$ there is $x \in X: f(x) \leqq \inf GP_C + \varepsilon$ and $F(x) \in \bar{V}(\bar{C}) \cap Y \subset V^2(C) \subset \tilde{C}$. As ε can be considered arbitrarily small, we have $\inf P_{\tilde{C}} \leqq \inf GP_C$. The last assertion of the theorem is straightforward. □

The situation in the constrained optimization problems differs essentially from that in the unconstrained problems, where the case $\inf P_C > \inf GP_C$ cannot appear (see [6]) and therefore the generalized solutions cannot be "better" than the classical ones. It will be shown in § 6.1 that certain "classical" conditions can guarantee that $\inf P_C = \inf GP_C$. Now, an example for the case $\inf P_C > \inf GP_C$ is given. An example of a similar nature has been given in [7, p. 247]. In our example, the problem P_C is even convex, but noncoercive. Problems which are coercive but nonconvex can be constructed, as well. To construct the example, we consider $X = l_2$, i.e. the Hilbert space of the squared summable sequences denoted $x = (x_1, x_2, \cdots)$; $Y = \bar{R}$; $f(x) = \sum_i -2^{-i}x_i$; $F(x) = \sum_i 2^{-5i}x_i^2$; and C is the set of nonpositive reals. Obviously, $S_{ad}P_C = \{0\}$ and $\inf P_C = 0$. Taking the sequence $x^{(j)}$, $j = 1, 2, \cdots$, where $x_i^{(j)} = 2^{2i}$ for $i = j$ and vanishes for $i \neq j$, we have $f(x^{(j)}) = -2^j$ and $F(x^{(j)}) = 2^{-j}$. In consequence of the compactness of $\bar{X}$, the sequence $\{x^{(j)}\}$ has a cluster point $\bar{x} \in \bar{X}$, for which we have $f(\bar{x}) = -\infty$ and $F(\bar{x}) = 0$; thus $\inf GP_C = -\infty$.

3. A characterization of $\inf GP_C$ and $\operatorname{Arginf} GP_C$ by the level sets. In what follows, we will consider the classical problem P_C to be nontrivial. For $a \in \bar{R}$ and $V \in \mathcal{U}_Y$, we denote the level set of the problem P_C: $\operatorname{lev}(a, V) = \{x \in X; f(x) \leqq a, F(x) \in V(C)\}$. By means of these level sets, a characterization of the value and of the set of the solutions of the generalized problem is given in Theorems 2 and 3, respectively.

THEOREM 2. *The collection* $\{\operatorname{lev}(a, V); a > a_0, V \in \mathcal{U}_Y\}$ *is a filter base on X iff* $a_0 \geqq \inf GP_C$.

Proof. For $a_0 < a < \inf f(X)$ the set $\operatorname{lev}(a, V)$ is empty, and therefore the collection in question is not a filter base. Now, consider $\inf f(X) \leqq a_0 < \inf GP_C$. Then the sets $\{\bar{x} \in \bar{X}; f(\bar{x}) \leqq a_0\}$ and $\{\bar{x} \in \bar{X}; F(\bar{x}) \in \bar{C}\}$ are disjoint and compact in $\bar{X}$, hence they have disjoint open neighbourhoods A_1, A_2, respectively. As $\bar{X} \setminus A_1$ is compact and f is l.s.c. on $\bar{X}$, $\inf f(\bar{X} \setminus A_1) > a_0$. As $\bar{X} \setminus A_2$ is compact, $F(\bar{X} \setminus A_2)$ is compact and disjoint with $\bar{C}$. Thus we see that for sufficiently small $\varepsilon > 0$ and $\bar{V} \in \mathcal{U}_{\bar{Y}}$, the sets $\{\bar{x} \in \bar{X}; f(\bar{x}) \leqq a_0 + \varepsilon\} \subset A_1$ and $\{\bar{x} \in \bar{X}; F(\bar{x}) \in \bar{V}(\bar{C})\} \subset A_2$ are disjoint, as well. Taking $V = \bar{V} \cap Y^2$, $\operatorname{lev}(a_0 + \varepsilon, V)$ is empty; thus the collection investigated cannot be a filter base.

On the other hand, for $a_0 \geqq \inf GP_C$, the considered level sets are nonempty, as it has been proved already in Theorem 1, and it is easy to see that they form a filter base. □

THEOREM 3. $\mathcal{N}(\operatorname{Arginf} GP_C) = \{U(\operatorname{lev}(a, V));\ U \in \mathcal{U}_X,\ V \in \mathcal{U}_Y,\ a > \inf GP_C\}$.

Proof. Due to Theorem 2 the collection in the right-hand side is a filter on X. First, consider any $A = U(\operatorname{lev}(a, V))$. Note that we may choose $U \in \mathcal{U}_X^*$ and $V \in \mathcal{U}_Y^*$ without any enlargement of A; cf. § 1. Taking symmetric $\bar{U} \in \mathcal{U}_{\bar{X}}$, $\bar{V} \in \mathcal{U}_{\bar{Y}}$, and a_1 such that $\bar{U}^2 \cap X^2 \subset U$, $\bar{V}^3 \cap Y^2 \subset V$, $\inf GP_C < a_1 < a$, we investigate the set $B = \bar{U}(\{\bar{x} \in \bar{X};\ f(\bar{x}) \leqq a_1,\ F(\bar{x}) \in \bar{V}(\bar{C})\})$. Clearly, $B \supset \bar{U}(\operatorname{Arginf} GP_C)$. Let $x \in B \cap X$. Then there is $\bar{x} \in \bar{X}$: $x \in \bar{U}(\bar{x})$, $f(\bar{x}) \leqq a_1$, $F(\bar{x}) \in \bar{V}(\bar{C})$. Due to the continuity of F on $\bar{X}$ and the definition of the extension of f, $\exists x_1 \in X$: $x_1 \in \bar{U}(\bar{x})$, $f(x_1) \leqq a$, $F(x_1) \in \bar{V}^2(\bar{C})$, and also $F(x_1) \in \bar{V}^3(C)$, hence $x \in A$. In other words, $B \cap X \subset A$ and B is a neighbourhood of $\operatorname{Arginf} GP_C$, thus $A \in \mathcal{N}(\bar{x})$ for every $\bar{x} \in \operatorname{Arginf} GP_C$.

Conversely, let $A \in \mathcal{N}(\operatorname{Arginf} GP_C)$. Then $\bar{A} = \operatorname{cl}_{\bar{X}} A$ is a neighbourhood of $\operatorname{Arginf} GP_C$. Since $\operatorname{Arginf} GP_C$ is compact, for a sufficiently small $\bar{U} \in \mathcal{U}_{\bar{X}}$, we have $\bar{A} \supset \bar{U}^2(\operatorname{Arginf} GP_C)$ and $A \supset \bar{U}^2(\operatorname{Arginf} GP_C) \cap X$. To prove the last inclusion, let us suppose the contradiction, i.e. that for every $\bar{U} \in \mathcal{U}_{\bar{X}}$ the set $N_{\bar{U}} = (X \setminus A) \cap \bar{U}^2(\operatorname{Arginf} GP_C)$ is nonempty. Consider the set $M_{\bar{U}} = \bar{U}^2(N_{\bar{U}}) \cap \operatorname{Arginf} GP_C$. For a symmetric entourage $\bar{U}$, $M_{\bar{U}}$ is nonempty. Note that it is generally not true if $\bar{U}$ is not symmetric. Furthermore, $\{M_{\bar{U}};\ \bar{U}$ be symmetric $\in \mathcal{U}_{\bar{X}}\}$ is the base of a filter on $\operatorname{Arginf} GP_C$ which has a cluster point $\bar{x} \in \operatorname{Arginf} GP_C$, because $\operatorname{Arginf} GP_C$ is compact. For this $\bar{x}$, $(X \setminus A) \cap \bar{U}^3(\bar{x})$ is nonempty for every $\bar{U} \in \mathcal{U}_{\bar{X}}$, which shows that $A \notin \mathcal{N}(\bar{x})$. This is, however, the desired contradiction.

To go on with the proof, we consider the sets $M = \bar{X} \setminus \bar{U}(\operatorname{Arginf} GP_C)$ and $L(\varepsilon, \bar{V}) = \{\bar{x} \in \bar{X};\ f(\bar{x}) \leqq \inf GP_C + \varepsilon,\ F(\bar{x}) \in \bar{V}(\bar{C})\}$. We may suppose $\bar{U}$ to be open, hence M is compact. We will prove that $\exists \varepsilon > 0$ and $\bar{V} \in \mathcal{U}_{\bar{Y}}$: $M \cap L(\varepsilon, \bar{V}) = \varnothing$. If it does not hold, $\{M \cap L(\varepsilon, \bar{V});\ \varepsilon > 0,\ \bar{V} \in \mathcal{U}_{\bar{Y}}\}$ will be the base of a filter on M which has a cluster point in M, say $\bar{x}$. For this $\bar{x}$ we have $f(\bar{x}) \leqq \inf GP_C$ and $F(\bar{x}) \in \bar{C}$, thus $\bar{x} \notin M$, which is, however, the contradiction. Hence for sufficiently small $\varepsilon > 0$ and $\bar{V} \in \mathcal{U}_{\bar{Y}}$ we have $L(\varepsilon, \bar{V}) \subset \bar{U}(\operatorname{Arginf} GP_C)$. Therefore $A \supset \bar{U}^2(\operatorname{Arginf} GP_C) \cap X \supset \bar{U}(L(\varepsilon, \bar{V})) \cap X = U(\operatorname{lev}(\inf GP_C + \varepsilon, V))$ with $U = \bar{U} \cap X^2 \in \mathcal{U}_X$, $V = \bar{V} \cap Y^2 \in \mathcal{U}_Y$. □

Remark. The filters $\mathcal{N}(\bar{x})$, being maximal $\mathcal{U}_X$-round, can be "constructed" either by means of the axiom of choice, at least if $\bar{x} \notin X$, or by means of the uniformity $\mathcal{U}_X^*$ (see [6, Thm. 4]), which is again of a nonconstructive nature. On the other hand, the characterization of $\mathcal{N}(\operatorname{Arginf} GP_C)$ by Theorem 3 may be considered as constructive because it does not use either the axiom of choice or the uniformity $\mathcal{U}_X^*$. Thus Theorem 3 gives certain "constructive" information about $\operatorname{Arginf} GP_C$, while there does not exist any possibility to give such a characterization for the element of $\operatorname{Arginf} GP_C$.

4. Some convergence results for the perturbed classical problem. In this section, it is shown that the value and the set of the approximate solutions of the perturbed classical problem $P_{\tilde{C}}$ converge just to the value and to the set of the generalized solutions of the problem GP_C, respectively. Thus the generalized problem GP_C may be considered as a more natural setting of the investigated optimization problem than its classical formulation P_C. The convergence results are summarized in the following theorem.

THEOREM 4. *Let $\delta > 0$ and $\bar{U} \in \mathcal{U}_{\bar{X}}$ be given. Then for sufficiently small $\varepsilon > 0$ and $V \in \mathcal{U}_Y$ it holds*

$\inf GP_C \geqq \inf P_{\tilde{C}} \geqq \inf GP_C - \delta$,

$\operatorname{Arginf}_\varepsilon P_{\tilde{C}} \subset \bar{U}(\operatorname{Arginf} GP_C)$,

$\operatorname{Arginf} GP_C \subset \bar{U}(\operatorname{Arginf}_\varepsilon P_{\tilde{C}})$, *where* $\tilde{C} = V(C)$.

Proof. Due to Theorem 1, we have $\inf GP_C \geqq \inf P_{\tilde{C}}$. As f is l.s.c. on $\bar{X}$ and $S_{ad} GP_C$ is compact, the set $\{\bar{x} \in \bar{X};\ f(\bar{x}) \geqq \inf GP_C - \delta\}$ is a neighbourhood of $S_{ad} GP_C$. Hence

there is an open entourage $\bar{U}_0 \in \mathcal{U}_{\bar{X}}$ such that $f(\bar{U}_0(S_{ad}GP_C)) \geqq \inf GP_C - \delta$. Using the compactness of $\bar{X} \setminus \bar{U}_0(S_{ad}GP_C)$ in the same manner as in the proof of Theorem 2, we obtain $\bar{V}_1 \in \mathcal{U}_{\bar{Y}}$ such that $\{\bar{x} \in \bar{X};\ F(\bar{x}) \in \bar{V}_1(\bar{C})\} \subset \bar{U}_0(S_{ad}GP_C)$. Taking $V_1 = \bar{V}_1 \cap Y^2$ and $\tilde{C} = V_1(C)$, we have $S_{ad}P_{\tilde{C}} \subset \bar{U}_0(S_{ad}GP_C)$, hence $\inf P_{\tilde{C}} \geqq \inf GP_C - \delta$.

Taking $\bar{V}_2 \in \mathcal{U}_{\bar{Y}}$ and $\varepsilon > 0$ such that $L(\varepsilon, \bar{V}_2) \subset \bar{U}(\operatorname{Arginf} GP_C)$, see the proof of Theorem 3, we obtain $\operatorname{Arginf}_\varepsilon P_{\tilde{C}} \subset \bar{U}(\operatorname{Arginf} GP_C)$ for $\tilde{C} = V_2(C)$ with $V_2 = \bar{V}_2 \cap Y^2$.

It is clear that the first two assertions of Theorem 4 are fulfilled if we choose $V \in \mathcal{U}_Y$ such that $V \subset V_1 \cap V_2$. Moreover, for a sufficiently small V, we have $\inf P_{\tilde{C}} \geqq \inf GP_C - \varepsilon/2$. Consider $\bar{x} \in \operatorname{Arginf} GP_C$ and $\bar{U}_1 \in \mathcal{U}_{\bar{X}}$, $\bar{U}_1 \subset \bar{U}^{-1}$. There is $x \in X$ such that $x \in \bar{U}_1(\bar{x})$, $F(x) \in V(C)$ and $f(x) \leqq \inf GP_C + \varepsilon/2$, and also $f(x) \leqq \inf P_{\tilde{C}} + \varepsilon$. Thus $\bar{x} \in \bar{U}(x)$ and $x \in \operatorname{Arginf}_\varepsilon P_{\tilde{C}}$, which completes the proof. □

5. The penalty-function method. In this section, a general setting of the well-known penalty-function method is investigated. Denoting φ a function $Y \to \bar{R}$ and $\{K_n\}$ an increasing sequence of positive constants tending toward the infinity, we can introduce the sequence of the unconstrained classical optimization problems, denoted $P_{C,n}$, which is created by applying the penalty-function method to the original problem P_C:

$$P_{C,n} \qquad \begin{array}{l} \text{minimize } f_n(x), \\ \text{with } f_n = f + K_n \cdot \varphi \circ F, \\ \text{subject to } x \in X. \end{array}$$

The expression $(-\infty)+(+\infty)$ in the definition of f_n will be avoided due to the assumptions in Theorem 5. As usual, we denote the value $\inf P_{C,n} = \inf f_n(X)$ and the set of the approximate classical solutions $\operatorname{Arginf}_\varepsilon P_{C,n} = \{x \in X;\ f_n(x) \leqq \inf P_{C,n} + \varepsilon\}$. In the classical framework, very little can be said about the convergence when $n \to \infty$, see e.g. [3]. The generalized problem GP_C, however, enables us to give some convergence results. Recall that P_C is supposed to be nontrivial.

THEOREM 5. *Let f be bounded from below by a constant $b > -\infty$; and let $\varphi : Y \to \bar{R}$ fulfill the assumptions: φ is uniformly continuous, $\varphi(C) = 0$, and $\forall V \in \mathcal{U}_Y \exists \delta > 0$: $\varphi(Y \setminus V(C)) \geqq \delta$. Then* $\inf P_{C,n} \nearrow \inf GP_C$ *(monotone convergence). Furthermore, let $x_n \in \operatorname{Arginf}_{\varepsilon(n)} P_{C,n}$ with $\varepsilon(n) \searrow 0$ for $n \to +\infty$. Then the sequence $\{x_n\}$ has a cluster point and each of such cluster points belongs to* $\operatorname{Arginf} GP_C$.

Proof. The sequence $\{\inf P_{C,n}\}$ is nondecreasing, because $f_n \leqq f_m$ for $n \leqq m$, and thus there exists its limit $\lim_{n\to\infty} \inf P_{C,n}$. Clearly, the assumptions imposed on φ guarantee that $\varphi(\bar{y}) = 0$ for $\bar{y} \in \bar{C}$ and $\varphi(\bar{y}) > 0$ for $\bar{y} \notin \bar{C}$. Let $\bar{x} \in \operatorname{Arginf} GP_C$. Then $f(\bar{x}) = \inf GP_C$ and $\varphi(F(\bar{x})) = 0$. Since $\varphi \circ F$ is continuous on $\bar{X}$, there is $\bar{U} \in \mathcal{U}_{\bar{X}}$ (depending on $\varepsilon > 0$) such that $\varphi(F(\bar{U}(\bar{x}))) \leqq \varepsilon$. Also there is $x \in X \cap \bar{U}(\bar{x})$ for which $f(x) \leqq \inf GP_C + \varepsilon$. Therefore, we have $f_n(x) \leqq \inf GP_C + \varepsilon(1 + K_n)$. Thus for every n, $\inf P_{C,n} \leqq \inf GP_C$, because ε has been arbitrarily positive, and obviously $\lim_{n\to\infty} \inf P_{C,n} \leqq \inf GP_C$.

Since $\bar{X}$ is compact, the sequence $\{x_n\}$ has at least one cluster point, say $\bar{x}$. We can easily obtain the estimate $\varphi(F(x_n)) \leqq (\inf P_C - b + \varepsilon(n))/K_n$, which gives $\varphi(F(x_n)) \to 0$ for $n \to \infty$. Thus $\varphi(F(\bar{x})) = 0$, and also $F(\bar{x}) \in \bar{C}$; in other words, $\bar{x} \in S_{ad}GP_C$. Obviously, $\inf P_{C,n} \geqq f(x_n) - \varepsilon(n)$, hence $\lim_{n\to\infty} \inf P_{C,n} \geqq f(\bar{x}) \geqq \inf GP_C$. This shows that $\lim_{n\to\infty} \inf P_{C,n} = \inf GP_C$ and $f(\bar{x}) = \inf GP_C$, i.e. $\bar{x} \in \operatorname{Arginf} GP_C$. □

6. Connections to the classical concepts. In this section two special cases are studied. In both of them, X is a normed linear space (not necessarily complete) with the norm $\|\cdot\|_X$. Naturally, the uniformity $\mathcal{U}_X$ is the norm uniformity, i.e. with the base $\{\{(x_1, x_2) \in X^2;\ \|x_1 - x_2\|_X \leqq \varepsilon\};\ \varepsilon > 0\}$.

6.1. Y is an ordered normed linear space. Let Y be an ordered normed linear space with the norm $\|\cdot\|_Y$ and the ordering $\leqq$. We suppose that the nonpositive cone $C=\{y\in Y;\ y\leqq 0\}$ has a nonempty interior; and $y>0$ will mean that $-y\in \operatorname{int} C$ (this notation is due to [3]). Of course, $\mathcal{U}_Y$ is the norm uniformity on Y.

In this framework the classical optimization problem P_C has the form

$$P_C^{(1)} \qquad \begin{array}{l} \text{minimize } f(x) \\ \text{subject to } x\in X, \qquad F(x)\leqq 0. \end{array}$$

Since for $p>0$ $C+p$ is a uniform neighbourhood of C (i.e. $C+p=V(C)$ for some $V\in\mathcal{U}_Y$), the perturbed classical problem may have the form

$$P_{C+p}^{(1)} \qquad \begin{array}{l} \text{minimize } f(x) \\ \text{subject to } x\in X, \qquad F(x)\leqq p, \end{array}$$

with $p>0$. Note that for every $V\in\mathcal{U}_Y$ we can choose a sufficiently small $p>0$ such that $C+p\subset V(C)$. Thus due to Theorem 3, the filter $\mathcal{N}(\operatorname{Arginf} GP_C^{(1)})$ has the base $\{A_{\varepsilon,p};\ \varepsilon>0, p>0\}$ with $A_{\varepsilon,p}=\{x\in X;\ \exists x_1\in X,\ \|x-x_1\|_X\leqq\varepsilon,\ f(x_1)\leqq\inf GP_C^{(1)}+\varepsilon,\ F(x_1)\leqq p\}$. Furthermore, Theorem 4 gives the convergence results for $p\searrow 0$ and $\varepsilon\searrow 0$, namely $\inf P_{C+p}^{(1)}\to\inf GP_C^{(1)}$ and $\operatorname{Arginf}_\varepsilon P_{C+p}^{(1)}\to\operatorname{Arginf} GP_C^{(1)}$ (in the sense of Theorem 4).

Now, we can investigate the penalty-function method in more detail. In the function spaces discussed in the sequel, C will be the cone of nonpositive functions on a measurable domain Ω in an Euclidean space. Using the usual quadratic penalty function and the projection onto the cone C, we obtain a special form of the function φ used in § 5, namely $\varphi: y\mapsto\inf_{y_1\leqq 0}\|y-y_1\|_Y^2$. Obviously, this function is uniformly continuous as a function $Y\to\bar{R}$, $\varphi(y)=0$ for $y\leqq 0$ and $\forall p>0\exists\delta>0$; $y\leqq p$ or $\varphi(y)\geqq\delta$. Thus we may apply Theorem 5 to obtain the convergence results. It should be emphasized that for the evaluation of φ we need not know the projection of y onto C. For example, for the case $Y=L_\infty(\Omega)$, i.e. the Banach space of the essentially bounded measurable functions, we can explicitly evaluate $\varphi(y)=((\operatorname{ess\,sup} y)^+)^2$ where $a^+=\max(a,0)$, while the projection of y onto C is generally nonunique. The same situation appears in the space $C^0(\Omega)$, i.e. the Banach space of the continuous functions on a compact domain Ω. On the other hand, the evaluation of φ may be very difficult even if the projection onto C is unique. This is the case $Y=H^1([0,1])$, i.e. the Sobolev space (of the Hilbert type) on the interval $[0,1]$, in which the projection is of very complicated nature, as it was shown in [5].

Remark. In applications, the interior of the cone C is often empty. For example, this is the case $Y=L_2(\Omega)$, i.e. the Hilbert space of the squared integrable functions. In such cases the notation used in this section cannot be employed because $C+p$ is not a uniform neighbourhood of C. However, the general theory may be applied. Using the function $\varphi: y\mapsto\inf_{y_1\in C}\|y-y_1\|_Y^2$, the filter $\mathcal{N}(\operatorname{Arginf} GP_C^{(1)})$ can be generated by the base $\{A_\varepsilon;\ \varepsilon>0\}$ with $A_\varepsilon=\{x\in X;\ \exists x_1\in X, \max(\|x-x_1\|_X; f(x_1)-\inf GP_C^{(1)};\ \varphi(F(x_1)))\leqq\varepsilon\}$. The perturbed problem can take the set $\tilde{C}=\{y\in Y;\ \varphi(y)\leqq\varepsilon\}$ with $\varepsilon>0$. Note that the sets $\tilde{C}$, having nonempty interiors, are quite large in comparison with C, but even such large perturbations are reasonable from the point of view of our theory because they still guarantee the convergence (in the sense of Theorem 4) of the perturbed problems. The penalty-function method can be applied in the same manner as in the case when C has a nonempty interior. If $Y=L_2(\Omega)$, then we can even evaluate $\varphi(y)=\|y^+\|_{L_2(\Omega)}^2$.

6.2. Classical dual problems. Now it will be shown that $\inf GP_C$ equals just to the extremal value of the classical dual problem obtained by the usual augmented-Lagrangian method (for a survey of the duality theory in optimization see [4]). Here C is a nonempty subset of a normed linear space Y endowed with a norm $\|\cdot\|_Y$. The dual problem can be constructed by the general duality theory of Lindberg when considered the generalized pairing $\psi: Y\times R^+\times Y^*\to R$ defined by $\psi(p, r, v)=\langle v, p\rangle + r\cdot\|p\|_Y^a$ where R^+ is the set of nonnegative reals, Y^* is the dual of Y, $\langle\cdot,\cdot\rangle$ is the pairing between Y^* and Y, and $a>0$. The perturbed objective $\Phi: X\times Y\to\bar R$ of the problem P_C is defined as $\Phi(x, p)=f(x)+\delta_C(F(x)-p)$ where $\delta_C: Y\to\bar R$ is the usual indicator function of C. The (augmented) Lagrangian $L: X\times R^+\times Y^*\to\bar R$ is then defined by

$$L(x, r, v)=\inf_{p\in Y}(\Phi(x, p)+\psi(p, r, v))=f(x)+\inf_{p\in C}\psi(F(x)-p, r, v),$$

and the dual objective $G: R^+\times Y^*\to\bar R$ has the form

$$G(r, v)=\inf_{x\in X}L(x, r, v)=\inf_{p\in Y}(h(p)+\psi(p, r, v)),$$

where $h: Y\to\bar R$ is the extremal-value function defined by

$$h(p)=\inf_{x\in X}\Phi(x, p).$$

The extremal value of the dual problem (i.e. the problem: maximize G over the domain $R^+\times Y^*$) is denoted β (i.e. $\beta=\sup G(R^+, Y^*)$).

We will prove that $\beta=\inf GP_C$ provided that f is bounded from below as required in Theorem 5. Denote

$$\varepsilon(r, v)=\inf_{p\in Y}\psi(p, r, v).$$

Clearly, $\varepsilon(r, v)\leqq 0$ and $\lim_{r\to+\infty}\varepsilon(r, v)=0$ for any $v\in Y^*$. As $\psi(p, r_1, v_1)+\psi(p, r_2, v_2)=\psi(p, r_1+r_2, v_1+v_2)$, for $r\geqq r_0\geqq 0$ we have $G(r, v)\geqq G(r_0, v_0)+\varepsilon(r-r_0, v-v_0)$. For any $\delta>0$ we can take r_0, v_0 such that $G(r_0, v_0)\geqq\beta-\delta/2$ and r such that $\varepsilon(r-r_0, -v_0)\geqq -\delta/2$, hence $G(r, 0)\geqq\beta-\delta$. Since the function $r\mapsto G(r, 0)$ is nondecreasing, we see that $\lim_{r\to+\infty}G(r, 0)=\beta$. However, using the notation of § 5 with $\varphi(y)=\inf_{y_1\in C}\|y-y_1\|_Y^a$, we have clearly $\inf P_{C,n}=G(K_n, 0)$, and employing Theorem 5, we come to $\inf GP_C=\beta$.

The problem P_C is called ψ-normal (with respect to the perturbations Φ) iff $\beta=\inf P_C$, see [4]. Thus we may use the usual criteria for ψ-normality which in our situation ensures that $\inf P_C=\inf GP_C$ and, due to Theorem 1, $\operatorname{Arginf} P_{\mathrm{cl}_Y C}=\operatorname{Arginf} GP_C\cap X$.

6.3. $\bar R$-valued constraints. Let I be an arbitrary index set. We will study the optimization problem

$$P_C^{(2)}\qquad \begin{aligned}&\text{minimize } f(x)\\ &\text{subject to } x\in X,\quad F_i(x)\leqq 0,\quad i\in I,\end{aligned}$$

where $F_i: X\to\bar R$ are uniformly continuous. To employ the general framework used above, we set $Y=\bar R^I$, $F=(F_i)_{i\in I}$ and $C=\{(y_i)_{i\in I}\in\bar R^I;\ \forall i\in I\ y_i\leqq 0\}$. If the uniformity on Y is considered as the product of the uniformities on $\bar R$ (i.e. $\mathcal{U}_Y=\prod_{i\in I}\mathcal{U}_{\bar R}$), then F is uniformly continuous, C is closed and, moreover, Y is compact. Therefore $\bar Y=Y$ and $\bar C=C$, which simplifies the generalized problem which can be explicitly written

in the form

$$GP_C^{(2)} \quad \begin{array}{l} \text{minimize } f(\bar{x}) \\ \text{subject to } \bar{x} \in \bar{X}, \quad F_i(\bar{x}) \leqq 0, \quad i \in I. \end{array}$$

The perturbed classical problem can now have the form

$$P_{\varepsilon,J}^{(2)} \quad \begin{array}{l} \text{minimize } f(x) \\ \text{subject to } x \in X, \quad F_i(x) \leqq \varepsilon, \quad i \in J, \end{array}$$

with $\varepsilon > 0$ and J is a finite subset of I. Obviously, the sets in the form $A_{\varepsilon,J} = \{x \in X; \exists x_1 \in X, \|x - x_1\|_X \leqq \varepsilon, f(x_1) \leqq \inf GP_C^{(2)} + \varepsilon, F_i(x_1) \leqq \varepsilon \text{ for } i \in J\}$ generate the filter $\mathcal{N}(\text{Arginf } GP_C^{(2)})$ when ε ranges the positive reals and J ranges the finite subsets of I. Furthermore, Theorem 4 gives the interesting result that $\inf GP_C^{(2)}$ and $\text{Arginf } GP_C^{(2)}$ can be obtained with prescribed accuracy when one fulfills (with certain accuracy) only a finite number of the constraints.

Supposing I to be finite, we can obtain the usual penalized problem by means of the function $\varphi: (y_i)_{i \in I} \mapsto \sum_{i \in I} (y_i^+)^2$ which clearly fulfills the properties required in Theorem 5. It is interesting that the penalty-function method can be generalized to the case when I is countable, say $I = \{1, 2, \cdots\}$. Then the penalized problem can be constructed by means of the function $\varphi: (y_i)_{i \in I} \mapsto \sum_{i \in I} 2^{-i} y_i^+ / (1 + y_i^+)$. The verification of the properties required for φ is again a simple matter.

Remark. The widely appearing case of a finite number of the constraints $F_i: X \to R$ can be treated by both of the mentioned special cases; however, the former case (§ 6.1) requires F_i to be uniformly continuous with respect to the additive uniformity on R, while for the later case F_i may be only uniformly continuous with respect to the uniformity induced from $\bar{R}$.

Remark. Let X be an Euclidean space, C be closed in Y, and the optimization problem P_C have certain coercive structure, namely either $f(x) \to +\infty$ for $\|x\|_X \to +\infty$ or $\exists V \in \mathcal{U}_Y$ such that the set $\{x \in X; F(x) \in V(C)\}$ is bounded. Then we can restrict our investigation to a sufficiently large closed ball B in X. Since B is compact, we have clearly $B \cap S_{ad} GP_C = B \cap S_{ad} P_C$, $\text{Arginf } GP_C = \text{Arginf } P_C$ and $\inf GP_C = \inf P_C$. Thus our theory is not too interesting for the coercive problems on a finite-dimensional domain.

7. Concluding remarks. If we accept the generalized problems as a realistic setting of the "technical" optimization problems, and thus avoid the classical problems, we can introduce also the perturbed generalized problems, which might be denoted as $GP_{\tilde{C}}$, for which we can investigate the convergence $\inf GP_{\tilde{C}} \to \inf GP_C$ and $\text{Arginf}_\varepsilon GP_{\tilde{C}} \to \text{Arginf } GP_C$ in a similar manner as it was made in § 4. Also the penalized classical problems can be extended on $\bar{X}$ in consequence of the uniform continuity of φ. Thus we can obtain the problems which might be denoted as $GP_{C,n}$, and similar convergence properties as in § 5 can be established.

Also, it should be pointed out that our concept is closely related with the notion of a proximity space, see e.g. [2]. For readers familiar with the proximity-space theory we recall that $\mathcal{U}_X^*$ induces not only the same topology as $\mathcal{U}_X$, but also the same proximity as $\mathcal{U}_X$. Moreover, $\mathcal{U}_X^*$ is the only precompact uniformity inducing this proximity. At the same time, $\mathcal{U}_X^*$ is the coarsest uniformity inducing the proximity in question. $\bar{X}$ is isomorphic to the so-called Smirnov compactification of X with respect to the proximity induced by $\mathcal{U}_X$.

Acknowledgment. The author would like to thank to J. Jarušek of the Institute of Information Theory and Automation, Czechoslovak Academy of Sciences for helpful advice and remarks in the course of preparing of this paper.

REFERENCES

[1] N. BOURBAKI, *General Topology*, Hermann, Paris, 1966.
[2] Á. CSÁSZÁR, *General Topology*, Akadémiai Kiadó, Budapest, 1978.
[3] D. LUENBERGER, *Optimization by Vector Space Methods*, John Wiley, New York, 1968.
[4] J. V. OUTRATA AND J. JARUŠEK, *Duality theory in mathematical programming and optimal control*, Supplement to Kybernetika, Vols. 20 (1984) and 21 (1985).
[5] J. V. OUTRATA AND Z. SCHINDLER, *An augmented Lagrangian method for a class of convex continuous optimal control problems*, Problems Control Inform. Theory, 10 (1981), pp. 67–81.
[6] T. ROUBÍČEK, *A Generalized Solution of a Nonconvex Minimization Problem and Its Stability*, Kybernetika 22 (1986), to appear.
[7] J. WARGA, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, London, 1972.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 5, September 1986

005

SIMILARITY AND REDUCTION FOR TIME VARYING LINEAR SYSTEMS WITH WELL-POSED BOUNDARY CONDITIONS*

I. GOHBERG† AND M. A. KAASHOEK‡

Abstract. The classical theory of similarity and reduction of causal and anti-causal systems is generalized and extended to time varying linear systems with well-posed boundary conditions. Two main problems are solved. The first is to determine to what extent a system with boundary conditions can be simplified by similarity and reduction. The second problem is to find the invariants of such a simplification. Special attention is given to the time invariant case.

Key words. time varying systems with boundary conditions, controllability, observability, irreducibility, weighting patterns, similarity, dilation, structure theorems, integral operators, semi-separable kernels

AMS(MOS) subject classifications. 93B10, 93B05, 93B07, 93B20, 45B05, 45E99, 93E10

0. Introduction and summary.

0.1. Introduction. On a finite time interval a multivariate time varying linear system with boundary conditions has the following state space representation:

$$\begin{aligned} &\dot{x}(t) = A(t)x(t) + B(t)u(t), && a \leqq t \leqq b, \\ &y(t) = C(t)x(t) + D(t)u(t), && a \leqq t \leqq b, \\ &N_1 x(a) + N_2 x(b) = 0. \end{aligned} \tag{0.1}$$

Here $A(t): X \to X$, $B(t): Z \to X$, $C(t): X \to Y$ and $D(t): Z \to Y$ are linear operators acting between finite dimensional linear spaces. As a function of t the main coefficient $A(t)$ is assumed to be integrable on $a \leqq t \leqq b$, the input coefficient $B(t)$ and the output coefficient $C(t)$ are square integrable and the external coefficient $D(t)$ is measurable and essentially bounded. The boundary conditions of the system (0.1), which are given in terms of two linear operators N_1 and N_2 acting on the state space X, are assumed to be well-posed, which means that $\det(N_1 + N_2 U(b)) \neq 0$. Here $U(t): X \to X$, $a \leqq t \leqq b$, is the fundamental operator of the system (0.1), i.e.,

$$\dot{U}(t) = A(t)U(t), \qquad a \leqq t \leqq b, \qquad U(a) = I_X.$$

The well-posedness of the boundary conditions implies that the input/output map of the system (0.1) can be written in the form of an integral operator:

$$y(t) = D(t)u(t) + \int_a^b k(t, s)u(s)\, ds, \qquad a \leqq t \leqq b,$$

the kernel $k(t, s)$ being given by

$$k(t, s) = \begin{cases} C(t)U(t)(N_1 + N_2 U(b))^{-1} N_1 U(s)^{-1} B(s), & a \leqq s < t \leqq b, \\ -C(t)U(t)(N_1 + N_2 U(b))^{-1} N_2 U(b) U(s)^{-1} B(s), & a \leqq t < s \leqq b. \end{cases}$$

Time varying linear systems with boundary conditions in state space representation appear for the first time explicitly in the work of A. J. Krener [17], [18], where he used such systems to model the problem of boundary value regulation. Later he analysed in this way the smoothing of non-Markovian linear processes [19]. More recently, also

* Received by the editors July 24, 1984, and in revised form March 20, 1985.

† School of Mathematical Studies, Tel Aviv University, Ramat Aviv, Tel Aviv, Israel.

‡ Department of Mathematics and Computer Science, Vrije Universiteit, Postbus 7161, 1007 MC Amsterdam, The Netherlands.

in the linear estimation theory of stochastic processes governed by time varying systems with boundary conditions, further progress was made by M. B. Adams, A. S. Willsky and B. C. Levy [1], [2], [3]. Systems with boundary conditions appear implicitly in several earlier papers on linear estimation theory written by T. Kailath and his co-authors (see [11] and the references given there). In Kailath's papers the systems are hidden and only their input/output maps appear in the form of integral operators with semi-separable kernels. Let us mention in particular the paper of B. D. O. Anderson and T. Kailath [4], where these connections are seen more clearly.

Together with H. Bart the present authors came to systems with boundary conditions via an analysis of Wiener-Hopf integral equations and related convolution equations [6], [7]. To illustrate this, consider the equation

$$y(t) = u(t) - \int_0^\tau k(t-s)u(s)\,ds, \qquad 0 \leqq t \leqq \tau, \tag{0.2}$$

and assume that the matrix function $k(t)$, which is defined on $[-\tau, \tau]$, admits an extension to a function on the full real line which has a rational Fourier transform $\hat{k}(\lambda)$ with $\hat{k}(\infty) = 0$. Then, using the classical realization theory for rational matrix functions, $\hat{k}(\lambda)$ can be written in the form $\hat{k}(\lambda) = C(A-\lambda)^{-1}B$, and it turns out that the integral equation (0.2) describes the input/output map of the following boundary value system:

$$\begin{aligned} \dot{x}(t) &= -iAx(t) + iBu(t), \qquad 0 \leqq t \leqq \tau, \\ y(t) &= -Cx(t) + u(t), \qquad 0 \leqq t \leqq \tau, \\ (I-P)x(0) &+ Px(\tau) = 0, \end{aligned} \tag{0.3}$$

where P is the spectral projection corresponding to the eigenvalues of A in the upper half plane. The connection between Wiener-Hopf equations and systems with boundary conditions allows one to obtain explicit formulas for the solutions of the equations and for the factorizations of the corresponding symbols. These results are not restricted to systems with finite dimensional state spaces and are of interest also for various classes of infinite dimensional problems. In particular, in this way it is possible to solve for the non-conservative case the linear problem of energy transport in a semi-infinite or finite medium [5, Chap. 6], [16], [21]. Let us mention that for the transport problems the parameter t in the systems (0.1) and (0.3) is not the time but a spatial variable which characterizes the deepness in the medium.

We feel that the time has come now to find out the basic theory for systems with boundary conditions and to bring it up to the level of the classical theory of time varying causal systems which was developed in the sixties by R. E. Kalman, D. C. Youla, L. Weiss and others (see, e.g., [14], [24], [25] and the books [8], [12], [15], [22]). In particular, this means that a similarity and reduction theory for systems with boundary conditions has to be developed. To do this is the main aim of the present paper.

0.2. Summary. The main results about causal systems which we would also like to obtain for systems with boundary conditions can be described shortly as follows. First of all, a causal time varying system θ is irreducible if and only if it is controllable and observable. Here θ is called irreducible if among all causal time invariant systems with the same weighting pattern as θ the state space of θ is of minimal dimension. By definition the weighting pattern is the finite rank kernel

$$k_1(t,s) = C(t)U(t)U(s)^{-1}B(s), \qquad a \leqq t \leqq b, \quad a \leqq s \leqq b.$$

Secondly, after similarity any causal system can be reduced to a controllable and observable one in a canonical way and without changing the weighting pattern, and, thirdly, two irreducible causal time varying systems with the same external coefficient and the same weighting pattern are similar. From these results it is clear that the weighting pattern is the full invariant for similarity and reduction.

In our case for time varying systems with noncausal boundary conditions it is not always possible to transform a system by similarity and reduction into a controllable and observable one. Further, examples show that the weighting pattern is not the full invariant. In fact, in the general case the role of the weighting pattern is taken over by a sequence of finite rank kernels:

$$k_r(t, s) = C(t)U(t)P^{r-1}U(s)^{-1}B(s), \qquad a \leqq t \leqq b, \quad a \leqq s \leqq b,$$

where $r = 1, 2, \cdots$ and $P = (N_1 + N_2U(b))^{-1}N_2U(b)$. This sequence of kernels (or their associated finite rank operators) will be called the *sequence of weighting patterns* of the system (0.1). In the causal case (i.e., $N_1 = I$, $N_2 = 0$ and hence $P = 0$) all kernels in the sequence are identically zero except the first one which is the classical weighting pattern.

To characterize the noncausal irreducible systems, we need to introduce for each noncausal system an extended system which is obtained from the original one by adding in a standard way a number of new inputs and new outputs. For the system (0.1) the extension we have in mind is defined as follows:

$$\begin{aligned} &\dot{x}(t) = A(t)x(t) + \tilde{B}(t)\tilde{u}(t), && a \leqq t \leqq b, \\ &\tilde{y}(t) = \tilde{C}(t)x(t), && a \leqq t \leqq b, \\ &N_1x(a) + N_2x(b) = 0, \end{aligned} \tag{0.4}$$

where

$$\begin{aligned} \tilde{B}(t) &= [B(t)\ U(t)PU(t)^{-1}B(t) \cdots U(t)P^{n-1}U(t)^{-1}B(t)], \\ \tilde{C}(t) &= [C(t)\ C(t)U(t)PU(t)^{-1} \cdots C(t)U(t)P^{n-1}U(t)^{-1}]^T \end{aligned}$$

with $P = (N_1 + N_2U(b))^{-1}N_2U(b)$ and n equal to the state space dimension. (Here, as well as in the sequel, the symbol $[\cdots]^T$ means that one has to take the block transpose.) We prove that the original system (0.1) is irreducible if and only if the extended system (0.4) is controllable and observable, which means that the following block matrices have full rank:

$$[\mathscr{C} \quad P\mathscr{C} \quad \cdots \quad P^{n-1}\mathscr{C}], \qquad [\mathscr{O} \quad \mathscr{O}P \quad \cdots \quad \mathscr{O}P^{n-1}]^T,$$

where $\mathscr{C}$ and $\mathscr{O}$ are the controllability and observability Gramians of (0.1), respectively, i.e.,

$$\mathscr{C} = \int_a^b U(t)^{-1}B(t)B(t)^*(U(t)^{-1})^*\,dt, \qquad \mathscr{O} = \int_a^b U(t)^*C(t)^*C(t)U(t)\,dt.$$

Irreducibility of (0.1) means now that among all time varying systems with the same sequence of weighting patterns as (0.1) the state space of (0.1) is of minimal dimension. As in the causal case a canonical procedure allows one to reduce (after similarity) any time varying system with boundary conditions to an irreducible one. Further, we show that irreducible systems with the same external coefficient and the same sequence of weighting patterns are similar.

We specify and extend the theory of similarity and reduction for time invariant systems. The system (0.1) is time invariant if the coefficients $A=A(t)$, $B=B(t)$, $C=C(t)$ and $D=D(t)$ do not depend on t. In the time invariant case irreducibility means that the following block matrices have full rank:

$$[B\ AB\ \cdots\ A^{n-1}B\ PB\ PAB\ \cdots\ PA^{n-1}B\ \cdots\ P^{n-1}B\ P^{n-1}AB\ \cdots\ P^{n-1}A^{n-1}B],$$

$$[C\ CA\ \cdots\ CA^{n-1}\ CP\ CAP\ \cdots\ CA^{n-1}P\ \cdots\ CP^{n-1}\ CAP^{n-1}\ \cdots\ CA^{n-1}P^{n-1}]^{T},$$

where, as before, n is equal to the dimension of the state space. For time invariant systems with the extra property that the operator P commutes with the main coefficient A, which is true obviously in the causal case, our results concerning similarity and reduction resemble strongly the classical theorems for causal time invariant systems.

Let us make a few remarks about the extended system defined by (0.4). In the paper we actually work with a more refined type of extension which has a few additional properties. Namely, if θ denotes the original system and $E(\theta)$ the extension, then $E(E(\theta))=E(\theta)$ and $E(\theta)=\theta$ whenever θ is causal, anti-causal or θ is controllable and observable. In our opinion this notion of extension deserves to be studied further from the system theoretic point of view.

The theory of similarity and reduction developed in this paper has a natural analogue for discrete time systems with boundary conditions which we have set out fully in [10, Chap. IV]. In [10] one also finds a more detailed exposition of the present paper.

1. Preliminaries. This section has a preparatory character. We introduce some necessary terminology and we recall for systems with boundary conditions the notions of similarity, dilation and irreducibility.

In what follows the time varying boundary value system (0.1) will be denoted by

$$\theta=(A(t), B(t), C(t), D(t); N_1, N_2)_a^b. \tag{1.1}$$

Here $A(t):X\to X$, $B(t):Z\to X$, $C(t):X\to Y$ and $D(t):Z\to Y$ are linear operators acting between finite dimensional linear spaces. As a function of t the *main coefficient* $A(t)$ is integrable on $a\leqq t\leqq b$, the *input coefficient* $B(t)$ and the *output coefficient* $C(t)$ are square integrable, and the *external coefficient* $D(t)$ is measurable and essentially bounded. The spaces X, Y and Z are finite dimensional vector spaces over $\mathbb{C}$ endowed with a norm and an inner product. The boundary conditions of (1.1) are given in terms of two linear operators N_1 and N_2 acting on the *state space* X. In (1.1) the indices a, b will be omitted when it is clear on which time interval the system θ has to be considered. By definition the *fundamental operator* of θ is the unique absolutely continuous solution $U(t):X\to X$, $a\leqq t\leqq b$, of the operator differential equation

$$\dot{U}(t)=A(t)U(t),\quad a\leqq t\leqq b,\quad U(a)=I_X. \tag{1.2}$$

The boundary conditions of θ are said to be *well-posed* if the operator $N_1+N_2U(b)$ is invertible, in which case they can be rewritten in the following equivalent form:

$$(I-P)x(a)+PU(b)^{-1}x(b)=0, \tag{1.3}$$

where $P=(N_1+N_2U(b))^{-1}N_2U(b)$. The operator P is called the *canonical boundary value operator* of θ. For a system θ with well-posed boundary conditions the input-

output operator $T_\theta: L_2([a, b], Z) \to L_2([a, b], Y)$ is the bounded integral operator

$$(T_\theta\varphi)(t) = D(t)\varphi(t) + C(t)U(t)\Big\{(I-P)\int_a^t U(s)^{-1}B(s)\varphi(s)\,ds - P\int_t^b U(s)^{-1}B(s)\varphi(s)\,ds\Big\}, \qquad a \leqq t \leqq b.$$

Controllability and observability of a system with well-posed boundary conditions do not depend on the particular form of the boundary conditions (see [18]) and they are the same as for causal systems (i.e., $N_1 = I$, $N_2 = 0$). The operators

$$\mathscr{C}(\theta) = \int_a^b U(t)^{-1}B(t)B(t)^*(U(t)^{-1})^*\,dt, \qquad \mathscr{O}(\theta) = \int_a^b U(t)^*C(t)^*C(t)U(t)\,dt$$

are called the *controllability Gramian* and *observability Gramian* of θ, respectively, and as in the causal case a system θ with well-posed boundary conditions is controllable (resp., observable) if and only if $\mathscr{C}(\theta)$ (resp., $\mathscr{O}(\theta)$) is invertible.

Two time varying systems θ_1 and θ_2,

$$\theta_\nu = (A_\nu(t), B_\nu(t), C_\nu(t), D_\nu(t); N_1^{(\nu)}, N_2^{(\nu)}), \qquad \nu = 1, 2, \tag{1.4}$$

with state spaces X_1 and X_2, respectively, are called *similar* (notation: $\theta_1 \simeq \theta_2$) if θ_1 and θ_2 have the same external coefficient and there exist an invertible operator $E: X_1 \to X_2$ and an absolutely continuous function $S(t): X_1 \to X_2$, $a \leqq t \leqq b$, of which the values are invertible operators, such that

$$A_2(t) = S(t)A_1(t)S(t)^{-1} + \dot{S}(t)S(t)^{-1}, \tag{1.5}$$

$$B_2(t) = S(t)B_1(t), \tag{1.6}$$

$$C_2(t) = C_1(t)S(t)^{-1}, \tag{1.7}$$

$$N_1^{(2)} = EN_1^{(1)}S(a)^{-1}, \qquad N_2^{(2)} = EN_2^{(1)}S(b)^{-1}, \tag{1.8}$$

almost everywhere on $a \leqq t \leqq b$. This notion of similarity appears in a natural way when in (0.1) the state $x(t)$ is replaced by $z(t) = S(t)x(t)$ (see [9, § I.5]). We shall refer to $S(t)$, $a \leqq t \leqq b$, as a *similarity transformation* between θ_1 and θ_2. Formula (1.5) implies that the fundamental operators $U_1(t)$ and $U_2(t)$ of θ_1 and θ_2, respectively, are related in the following way:

$$U_2(t) = S(t)U_1(t)S(a)^{-1}, \qquad a \leqq t \leqq b. \tag{1.9}$$

Well-posedness of the boundary conditions is preserved under a similarity transformation and similar systems with well-posed boundary conditions have similar canonical boundary value operators. In fact, if P_1 and P_2 are the canonical boundary value operators of θ_1 and θ_2, respectively, then the above formulas imply that $P_2 = S(a)P_1S(a)^{-1}$. It follows that similar systems with well-posed boundary conditions have the same input-output operators.

For a system with well-posed boundary conditions we define $\theta_\square$ to be the system

$$\theta_\square = (0, U(t)^{-1}B(t), C(t)U(t), D(t); I-P, P)_a^b. \tag{1.10}$$

Here $U(t)$ is the fundamental operator of θ and P is its canonical boundary value operator. Obviously, $\theta \simeq \theta_\square$; in fact, to obtain $\theta_\square$ one applies to θ the similarity transformation $S(t) = U(t)^{-1}$, $a \leqq t \leqq b$. Note that, $(\theta_\square)_\square = \theta_\square$.

A system $\theta = (A(t), B(t), C(t), D(t); N_1, N_2)_a^b$ will be called a *dilation* of the system $\theta_0 = (A_0(t), B_0(t), C_0(t), D_0(t); N_1^{(0)}, N_2^{(0)})_a^b$ if $D(t) = D_0(t)$ a.e. on $a \leqq t \leqq b$ and the state space X of θ admits a decomposition, $X = X_1 \oplus X_0 \oplus X_2$, such that relative to this decomposition the coefficients of θ and its boundary value operators are partitioned in the following way:

$$A(t) = \begin{pmatrix} * & * & * \\ 0 & A_0(t) & * \\ 0 & 0 & * \end{pmatrix}, \qquad B(t) = \begin{pmatrix} * \\ B_0(t) \\ 0 \end{pmatrix}, \tag{1.11}$$

$$C(t) = (0 \quad C_0(t) \quad *), \tag{1.12}$$

$$N_\nu = E \begin{pmatrix} * & * & * \\ 0 & N_\nu^{(0)} & * \\ 0 & 0 & * \end{pmatrix}, \qquad \nu = 1, 2. \tag{1.13}$$

Here (1.11) and (1.12) hold a.e. on $a \leqq t \leqq b$. The operator E appearing in (1.13) is some invertible operator on X. The symbols $*$ denote unspecified entries. The dilation is said to be *proper* whenever $\dim X > \dim X_0$. If the dilation θ has well-posed boundary conditions, then the same is true for θ_0 and the systems θ and θ_0 have the same input-output operator. To see this, one uses the fact that the fundamental operators $U(t)$ and $U_0(t)$ and the canonical boundary value operators P and P_0 of θ and θ_0 are related in the following way:

$$U(t) = \begin{pmatrix} * & * & * \\ 0 & U_0(t) & * \\ 0 & 0 & * \end{pmatrix}, \qquad P = \begin{pmatrix} * & * & * \\ 0 & P_0 & * \\ 0 & 0 & * \end{pmatrix}. \tag{1.14}$$

The system θ_0 is called a *reduction* of θ if θ is a dilation of θ_0, and θ_0 is said to be a *proper* reduction if, in addition, the state space dimension of θ_0 is strictly less than the state space dimension of θ. It is one of the classical results about causal systems that for a causal system θ the system $\theta_\square$ (see (1.10)) has a controllable and observable reduction (cf., [14], [24], [25]). Since $\theta \simeq \theta_\square$, it follows that a causal time varying system is similar to a dilation of a controllable and observable system (which is the main part of Kalman's canonical structure theorem [14]).

A time varying linear system θ with well-posed boundary conditions is called *irreducible* if none of the systems similar to θ admits a proper reduction. Since the system $\theta_\square$ has a proper reduction whenever θ (or any system similar to θ) has a proper reduction, one may conclude that θ is irreducible if and only if $\theta_\square$ does not have a proper reduction. For a causal system irreducibility is equivalent to controllability and observability (see [14]), but for systems with noncausal boundary conditions this is not true. In fact, in contrast to controllability and observability the notion of irreducibility depends heavily on the boundary conditions.

2. Main theorems. To state our main theorems, we use a standard procedure to add to a system new inputs and new outputs without changing the state space. Let $\theta = (A(t), B(t), C(t), D(t); N_1, N_2)_a^b$ be a system with well-posed boundary conditions. Let $U(t)$ be the fundamental operator of θ and P its canonical boundary value operator. For r and q positive numbers we define the (r, q)-*extension* of θ (notation: $E_{rq}(\theta)$) to be the system

$$E_{rq}(\theta) \quad \begin{cases} \dot{x}(t) = A(t)x(t) + B^{(r)}(t)\tilde{u}(t), & a \leqq t \leqq b, \\ \tilde{y}(t) = C^{(q)}(t), & a \leqq t \leqq b, \\ N_1 x(a) + N_2 x(b) = 0, \end{cases}$$

of which the input space is equal to $Z^{(r)}$ (i.e., the direct sum of r copies of the input space Z of θ), the output space is $Y^{(q)}$ (i.e., the direct sum of q copies of the outspace Y of θ) and

$$B^{(r)}(t)=[B(t) \quad U(t)PU(t)^{-1}B(t) \quad \cdots \quad U(t)P^{r-1}U(t)^{-1}B(t)],$$
$$C^{(q)}(t)=[C(t) \quad C(t)U(t)PU(t)^{-1} \quad \cdots \quad C(t)U(t)P^{q-1}U(t)^{-1}]^{T}.$$

The definition of $E_{rq}(\theta)$ does not involve the external coefficient of θ. The system $E_{rq}(\theta)$ has well-posed boundary conditions, and $E_{rq}(\theta)$ and θ have the same canonical boundary value operator. The controllability and observability Gramians of $E_{rq}(\theta)$ are the following operators:

$$\mathscr{C}(E_{rq}(\theta))=\sum_{\nu=1}^{r} P^{\nu-1}\mathscr{C}(\theta)(P^{*})^{\nu-1}, \tag{2.1}$$

$$\mathscr{O}(E_{rq}(\theta))=\sum_{\nu=1}^{q} (P^{*})^{\nu-1}\mathscr{O}(\theta)P^{\nu-1}, \tag{2.2}$$

where $\mathscr{C}(\theta)$ and $\mathscr{O}(\theta)$ are the controllability and observability Gramians of θ, respectively. Obviously,

$$\operatorname{Im}\mathscr{C}(E_{rq}(\theta))=\operatorname{Im}[\mathscr{C}(\theta) \quad P\mathscr{C}(\theta) \quad \cdots \quad P^{r-1}\mathscr{C}(\theta)], \tag{2.3}$$

$$\operatorname{Ker}\mathscr{O}(E_{rq}(\theta))=\bigcap_{j=1}^{q} \operatorname{Ker}\mathscr{O}(\theta)P^{j-1}. \tag{2.4}$$

Let m and l be the smallest positive integers such that both $\mathscr{C}(E_{ml}(\theta))$ and $\mathscr{O}(E_{ml}(\theta))$ are of maximal rank. Note that the dimension of the right-hand side of (2.3) is maximal for $r=d$, where d is the degree of the minimal polynomial of P (which is less than the dimension of the state space of θ). It follows that $1\leqq m\leqq d$. Similarly, $1\leqq l\leqq d$. The system $E_{ml}(\theta)$ will be called the *standard extension* of θ (notation: STE (θ)). If θ is causal (i.e., $N_1=I$, $N_2=0$ and hence $P=0$) or if θ is anti-causal (i.e., $N_1=0$, $N_2=I$ and hence $P=I$), then the degree of the minimal polynomial of P is 1 and hence in those cases STE $(\theta)=\theta$. The equality STE $(\theta)=\theta$ also holds if the original system θ is controllable and observable. Finally, if the standard extension procedure is applied to the system STE (θ), then nothing new is obtained, i.e.,

$$\text{STE (STE }(\theta))=\text{STE }(\theta). \tag{2.5}$$

THEOREM 2.1. *Let $\theta=(A(t), B(t), C(t), D(t); N_1, N_2)_a^b$ be a system with well-posed boundary conditions, and let $U(t)$ be its fundamental operator. The following statements are equivalent*:

(i) *θ is irreducible.*

(ii) *The standard extension of θ is controllable and observable.*

(iii) *The following block matrices have full rank*:

$$[\mathscr{C}(\theta) \quad P\mathscr{C}(\theta) \quad \cdots \quad P^{d-1}\mathscr{C}(\theta)], \qquad [\mathscr{O}(\theta) \quad \mathscr{O}(\theta)P \quad \cdots \quad \mathscr{O}(\theta)P^{d-1}]^{T};$$

(iv) $\operatorname{rank}[W_{j+k-1}(\theta)]_{j,k=1}^{d}=\dim X$.

Here $\mathscr{C}(\theta)$ and $\mathscr{O}(\theta)$ are the controllability and observability Gramians of θ, respectively, X is the state space of θ, the operator P is the canonical boundary value operator of θ, the number d is the degree of the minimal polynomial of P and $W_j(\theta)$ is the finite rank integral operator:

$$(W_j(\theta)\varphi)(t)=C(t)U(t)P^{j-1}\int_a^b U(s)^{-1}B(s)\varphi(s)\,ds, \qquad a\leqq t\leqq b. \tag{2.6}$$

The operator $W_j(\theta)$ defined by (2.6), which acts from $L_2([a, b], Z)$ into $L_2([a, b], Y)$, will be called the *jth weighting pattern* of the system $\theta (j = 1, 2, \cdots)$. For a causal system all weighting patterns are zero except the first one which is the classical weighting pattern. In case P is a projection only the first two weighting patterns are of interest. If the coefficients of θ are analytic functions in the time parameter t and P is a projection, then the weighting patterns are uniquely determined by the input-output operator and, conversely, the input-output operator is uniquely determined by (the first two) weighting patterns. The weighting patterns do not change under similarity, dilation and reduction.

THEOREM 2.2. *Two irreducible systems are similar if and only if they have the same external coefficient and the same sequence of weighting patterns, and in that case the similarity transformation is unique.*

THEOREM 2.3. *A time varying system θ with well-posed boundary conditions is irreducible if and only if among all time varying systems with well-posed boundary conditions and with the same sequence of weighting patterns as θ the system θ has the smallest state space dimension.*

THEOREM 2.4. *A time varying system θ with well-posed boundary conditions is similar to a dilation of an irreducible system. More precisely, let $\theta = (A(t), B(t), C(t), D(t); N_1, N_2)_a^b$ have well-posed boundary conditions, let $U(t)$ be the fundamental operator of θ and P its canonical boundary value operator, and choose a direct sum decomposition, $X = X_1 \oplus X_0 \oplus X_2$, of the state space X of θ such that $X_1 = \operatorname{Ker} \mathcal{O}$ and X_0 is a direct complement of $X_1 \cap \operatorname{Im} \mathcal{C}$ in $\operatorname{Im} \mathcal{C}$, where $\mathcal{C}$ and $\mathcal{O}$ are the controllability and observability Gramians of the standard extension of θ, respectively. Then relative to the decomposition $X = X_1 \oplus X_0 \oplus X_2$ the following partitionings hold true:*

$$U(t)^{-1}B(t) = \begin{pmatrix} * \\ B_0(t) \\ 0 \end{pmatrix}, \qquad C(t)U(t) = (0 \quad C_0(t) \quad *), \qquad a \leqq t \leqq b, \quad a.e.,$$

$$P = \begin{pmatrix} * & * & * \\ 0 & P_0 & * \\ 0 & 0 & * \end{pmatrix},$$

the system $\theta_0 = (0, B_0(t), C_0(t), D(t); I - P_0, P_0)_a^b$ is irreducible and θ is similar to a dilation of θ_0.

COROLLARY 2.5. *Two time varying systems θ_1 and θ_2 with well-posed boundary conditions are similar to dilations of the same (irreducible) system if and only if θ_1 and θ_2 have the same external coefficient and the same sequence of weighting patterns.*

The theory developed in this section yields a general procedure to transform an arbitrary time varying system θ with well-posed boundary conditions into a controllable and observable system. First one applies (as in Theorem 2.4) a similarity and a reduction to transform θ into an irreducible system θ_0. Next, one adds to θ_0 inputs and outputs using the standard extension procedure defined in the beginning of this section. According to Theorem 2.1 the resulting system STE (θ_0) is controllable and observable.

In an earlier version of this paper (see [10]) we used the terms p-controllability and q-observability to describe irreducibility. Here p-controllability and q-observability of θ mean that the operators

$$\sum_{j=1}^{p} P^{j-1}\mathcal{C}(\theta)(P^*)^{j-1}, \qquad \sum_{j=1}^{q} (P^*)^{j-1}\mathcal{O}(\theta)P^{j-1}$$

are invertible. Thus θ is p-controllable and q-observable if and only if the (p, q)-extension $E_{pq}(\theta)$ is controllable and observable, and according to Theorem 2.1 the system θ is irreducible if and only if θ is p-controllable and q-observable for p and q sufficiently large.

3. Proofs of the main theorems. In this section we prove Theorems 2.1–2.4 and Corollary 2.5. We begin with a few general observations. In what follows $\theta = (A(t), B(t), C(t), D(t); N_1, N_2)_a^b$ is a system with well-posed boundary conditions, $U(t)$ is the fundamental operator of θ and P its canonical boundary value operator.

LEMMA 3.1. *Assume that θ admits a reduction, i.e., the state space X of θ admits a decomposition $X = X_1 \oplus X_0 \oplus X_2$ such that relative to this decomposition* (1.11), (1.12) *and* (1.13) *hold true. Then*

$$\text{Im } P^{r-1}\mathscr{C}(\theta) \subset X_0 \oplus X_1, \quad \text{Ker } \mathscr{O}(\theta)P^{r-1} \supset X_1, \quad r \geqq 1. \tag{3.1}$$

Proof. From formula (1.11) it follows that the fundamental operator $U(t)$ of θ can be represented as in (1.14). This implies that $\text{Im } U(t)^{-1}B(t) \subset X_1 \oplus X_0$ and $\text{Ker } C(t)U(t) \supset X_1$ for almost all $a \leqq t \leqq b$. But then it is clear from the definitions of the Gramians $\mathscr{C}(\theta)$ and $\mathscr{O}(\theta)$ that $\text{Im } \mathscr{C}(\theta) \subset X_1 \oplus X_0$ and $\text{Ker } \mathscr{O}(\theta) \supset X_1$. Next, use the fact that P can be written as in (1.14). So the spaces $X_1 \oplus X_0$ and X_1 are invariant under P and (3.1) follows. □

Let $E_{ml}(\theta)$ be the standard extension of θ. Denote by $\mathscr{C}$ and $\mathscr{O}$ the controllability and observability Gramians of $E_{ml}(\theta)$. From the definition of the standard extension it follows (see (2.3) and (2.4)) that

$$\text{Im } \mathscr{C} = \text{Im } [\mathscr{C}(\theta) \quad P\mathscr{C}(\theta) \quad \cdots \quad P^{r-1}\mathscr{C}(\theta)], \qquad r \geqq m, \tag{3.2}$$

$$\text{Ker } \mathscr{O} = \bigcap_{j=1}^{q} \text{Ker } \mathscr{O}(\theta)P^{j-1}, \qquad q \geqq l. \tag{3.3}$$

This implies that $\text{Im } \mathscr{C}$ and $\text{Ker } \mathscr{O}$ are invariant under P.

Put $X_1 = \text{Ker } \mathscr{O}$, and let X_0 be a direct complement of $X_1 \cap \text{Im } \mathscr{C}$ in $\text{Im } \mathscr{C}$. Furthermore, let X_2 be a direct complement of $X_1 + X_0$ in the state space X of θ. Then $X = X_1 \oplus X_0 \oplus X_2$. The fact that P leaves invariant the spaces X_1 and $X_1 + X_0$ implies that relative to the decomposition $X = X_1 \oplus X_0 \oplus X_2$ the operator P admits the following partitioning:

$$P = \begin{pmatrix} * & * & * \\ 0 & P_0 & * \\ 0 & 0 & * \end{pmatrix}. \tag{3.4}$$

Let us consider the partitioning of $U(t)^{-1}B(t)$ and $C(t)U(t)$ relative to the decomposition $X = X_1 \oplus X_0 \oplus X_2$:

$$U(t)^{-1}B(t) = \begin{pmatrix} B_1(t) \\ B_0(t) \\ B_2(t) \end{pmatrix}, \qquad C(t)U(t) = (C_1(t) \quad C_0(t) \quad C_2(t)). \tag{3.5}$$

LEMMA 3.2. *In* (3.5) *the entries $B_2(t)$ and $C_1(t)$ are zero almost everywhere on $a \leqq t \leqq b$.*

Proof. Introduce the following auxiliary operators:

$$\Gamma: L_2([a, b], Z) \to X, \qquad \Gamma\varphi = \int_a^b U(s)^{-1}B(s)\varphi(s)\, ds, \tag{3.6}$$

$$\Lambda: X \to L_2([a, b], Y), \qquad (\Lambda x)(t) = C(t)U(t)x. \tag{3.7}$$

Note the $\Gamma\Gamma^* = \mathscr{C}(\theta)$ and $\Lambda^*\Lambda = \mathcal{O}(\theta)$. In particular, $\operatorname{Im}\Gamma = \operatorname{Im}\mathscr{C}(\theta)$ and $\operatorname{Ker}\Lambda = \operatorname{Ker}\mathcal{O}(\theta)$. Since $\operatorname{Im}\mathscr{C} \subset X_1 \oplus X_0$, we have $\operatorname{Im}\mathscr{C}(\theta) \subset X_1 \oplus X_0$ (see (3.2)), and it follows that $\Pi\Gamma = 0$, where Π is the projection of X along $X_1 \oplus X_0$ onto X_2. In other words

$$\int_a^b \Pi U(s)^{-1} B(s)\varphi(s)\, ds = 0 \qquad \forall \varphi \in L_2([a, b], Z).$$

But then we may conclude that $\Pi U(s)^{-1}B(s) = 0$ on $a \leqq s \leqq b$. This proves that $B_2(t) = 0$ a.e. Next observe that (3.3) implies that $\operatorname{Ker}\Lambda \subset X_1$, and thus $C(t)U(t)x = 0$ a.e. on $a \leqq t \leqq b$ for each $x \in X_1$. So $C_1(t) = 0$ a.e. □

Proof of Theorem 2.1. (i)⇒(ii). Assume θ is irreducible. Put

$$\theta_0 = (0, B_0(t), C_0(t), D(t); I - P_0, P_0)_a^b, \tag{3.8}$$

where $B_0(t)$ and $C_0(t)$ are defined by (3.5) and P_0 by (3.4). The triangular form of P (see (3.4)) and the fact that in (3.5) the entries $B_2(t)$ and $C_1(t)$ are zero a.e. imply that θ_0 is a reduction of the system $\theta_\square$ (which is defined by (1.10)). It follows that θ is similar to a dilation of θ_0. But θ is irreducible. So $X = X_0$, and we may conclude that $\operatorname{Ker}\mathcal{O} = (0)$ and $\operatorname{Im}\mathscr{C} = X$. In other words, the standard extension of θ is controllable and observable.

(ii)⇒(i). Assume that the standard extension of θ is controllable and observable. Let θ_1 be a system that is similar to θ. We want to prove that θ_1 does not have a proper reduction. First we show that the standard extension of θ_1 is also controllable and observable. Let X and X_1 be the state spaces of θ and θ_1, respectively, and let $S(t): X \to X_1$, $a \leqq t \leqq b$, be the similarity transformation between θ and θ_1. Then for any pair of positive numbers r and q the transformation $S(t)$ is a similarity between $E_{rq}(\theta)$ and $E_{rq}(\theta_1)$. This implies that the standard extension of θ and θ_1 are similar. Since controllability and observability are preserved under similarity, it follows that the standard extension of θ_1 is controllable and observable.

But then it suffices to show that θ does not have a proper reduction. Assume, as in Lemma 3.1, that θ has a reduction induced by the decomposition $X = X_1 \oplus X_0 \oplus X_2$. Then (3.1) implies that $\operatorname{Im}\mathscr{C} \subset X_0 \oplus X_1$ and $\operatorname{Ker}\mathcal{O} \supset X_1$. But θ is controllable and observable. So $\operatorname{Im}\mathscr{C} = X$ and $\operatorname{Ker}\mathcal{O} = (0)$. It follows that $X_0 = X$, and the reduction is not proper. So θ is irreducible.

(ii)⇔(iii). Let d be the degree of the minimal polynomial of P. Assume $E_{ml}(\theta)$ is the standard extension of θ. Then we know that $1 \leqq m \leqq d$ and $1 \leqq l \leqq d$. In particular (see (3.2) and (3.3)) we have

$$\operatorname{Im}\mathscr{C} = \operatorname{Im}[\mathscr{C}(\theta) \quad P\mathscr{C}(\theta) \quad \cdots \quad P^{d-1}\mathscr{C}(\theta)], \qquad \operatorname{Ker}\mathcal{O} = \bigcap_{j=1}^{d} \operatorname{Ker}\mathcal{O}(\theta)P^{j-1}. \tag{3.9}$$

Thus the standard extension of θ is controllable and observable if and only if

$$X = \operatorname{Im}[\mathscr{C}(\theta) \quad P\mathscr{C}(\theta) \quad \cdots \quad P^{d-1}\mathscr{C}(\theta)], \qquad \bigcap_{j=1}^{d} \operatorname{Ker}\mathcal{O}(\theta)P^{j-1} = (0). \tag{3.10}$$

This proves the equivalence of the statements (ii) and (iii).

(iii)⇔(iv). Using the auxiliary operators Γ and Λ defined by (3.6) and (3.7), respectively, one can rewrite the jth weighting pattern in the form: $W_j(\theta) = \Lambda P^{j-1}\Gamma$.

It follows that

$$[W_{j+k-1}(\theta)]_{j,k=1}^{d} = \begin{pmatrix} \Lambda \\ \Lambda P \\ \vdots \\ \Lambda P^{d-1} \end{pmatrix} [\Gamma \quad P\Gamma \quad \cdots \quad P^{d-1}\Gamma]. \tag{3.11}$$

Since $\Gamma\Gamma^* = \mathscr{C}(\theta)$ and $\Lambda^*\Lambda = \mathscr{O}(\theta)$, it is easily seen that

$$\operatorname{Im}[\mathscr{C}(\theta) \quad P\mathscr{C}(\theta) \quad \cdots \quad P^{d-1}\mathscr{C}(\theta)] = \operatorname{Im}[\Gamma \quad P\Gamma \quad \cdots \quad P^{d-1}\Gamma],$$

$$\bigcap_{j=1}^{d} \operatorname{Ker} \mathscr{O}(\theta)P^{j-1} = \bigcap_{j=1}^{d} \operatorname{Ker} \Lambda P^{j-1},$$

and thus the rank of the left-hand side of (3.11) is equal to dim X if and only if (3.10) holds. This proves the equivalence of (iii) and (iv). □

Proof of Theorem 2.4. Lemma 3.2 and the paragraph preceding this lemma show that $U(t)^{-1}B(t)$, $C(t)U(t)$ and P have the desired block matrix structure. It follows that the system $\theta_0 = (0, B_0(t), C_0(t), D(t); I - P_0, P_0)_a^b$ is a reduction of the system $\theta_\square$ (see (1.10)), and thus θ is similar to a dilation of θ_0. It remains to prove that θ_0 does not admit a proper reduction.

Let Γ and Λ be the operators defined by (3.6) and (3.7), and consider the following auxiliary operators

$$\Gamma_0 : L_2([a, b], Z) \to X_0, \qquad \Gamma_0 = \int_a^b B_0(s)\varphi(s)\, ds,$$

$$\Lambda_0 : X_0 \to L_2([a, b], Y), \qquad (\Lambda_0 x)(t) = C_0(t)x.$$

From the partitioning of $U(t)^{-1}B(t)$, $C(t)U(t)$ and P we may conclude that

$$P^{r-1}\Gamma = \begin{pmatrix} * \\ P_0^{r-1}\Gamma_0 \\ 0 \end{pmatrix}, \quad \Lambda P^{r-1} = (0 \quad \Lambda_0 P_0^{r-1} \quad *), \quad r \geqq 1. \tag{3.12}$$

Now recall that $X_0 \subset \operatorname{Im} \mathscr{C} = \operatorname{Im}[\Gamma \; P\Gamma \; \cdots \; P^{d-1}\Gamma]$, where d is the degree of the minimal polynomial of P. So the first identity in (3.12) implies that $X_0 = \operatorname{Im}[\Gamma_0 \; P_0\Gamma_0 \; \cdots \; P_0^{d-1}\Gamma_0]$. Furthermore, $X_1 = \bigcap_{j=1}^{d} \operatorname{Ker} \Lambda P^{j-1}$, and so the second identity in (3.12) shows that $\bigcap_{j=1}^{d} \operatorname{Ker} \Lambda_0 P_0^{j-1} = (0)$. Now apply Theorem 2.1 to conclude that θ_0 is irreducible. □

Proof of Theorem 2.2. It is easily seen that the weighting patterns are preserved under similarity. It follows that two similar systems have the same external coefficient and the same sequence of weighting patterns. We have to prove that for irreducible systems the converse statement is also true.

For $\nu = 1, 2$ let $\theta_\nu = (A_\nu(t), B_\nu(t), C_\nu(t), D_\nu(t); N_1^{(\nu)}, N_2^{(\nu)})_a^b$ be an irreducible system. Assume that $D_1(\cdot) = D_2(\cdot)$ and that θ_1 and θ_2 have the same sequence of weighting patterns. We want to prove that θ_1 and θ_2 are similar. For $\nu = 1, 2$ let X_ν be the state space of θ_ν, let $U_\nu(t)$ be the fundamental operator of θ_ν and P_ν its canonical boundary value operator. Take n to be the maximum of dim X_1 and dim X_2. Put

$$\tilde{B}_\nu(t) = [B_\nu(t) \quad U_\nu(t)P_\nu U_\nu(t)^{-1}B_\nu(t) \quad \cdots \quad U_\nu(t)P_\nu^n U_\nu(t)^{-1}B_\nu(t)],$$

$$\tilde{C}_\nu(t) = [C_\nu(t) \quad C_\nu(t)U_\nu(t)P_\nu U_\nu(t)^{-1} \quad \cdots \quad C_\nu(t)U_\nu(t)P_\nu^n U_\nu(t)^{-1}]^T,$$

and consider the causal system $\Delta_\nu = (A_\nu(t), \tilde{B}_\nu(t), \tilde{C}_\nu(t), 0; I, 0)_a^b$, where $\nu = 1, 2$. The

irreducibility of θ_1 and θ_2 implies that Δ_1 and Δ_2 are controllable and observable. Note that the first weighting pattern of Δ_ν is given by

$$W_1(\Delta_\nu) = [W_{j+k-1}(\theta_\nu)]_{j,k=1}^{n+1}.$$

Since θ_1 and θ_2 have the same sequence of weighting patterns, we conclude that Δ_1 and Δ_2 have the same first weighting pattern. But then we can apply the classical similarity theorem for (observable and controllable) causal systems (see, e.g., [25]) to show that $\Delta_1 \simeq \Delta_2$. So there exists an absolutely continuous function $S(t): X_1 \to X_2$, of which the values are invertible operators, such that

$$A_2(t) = S(t)A_1(t)S(t)^{-1} + \dot{S}(t)S(t)^{-1}, \tag{3.13}$$

$$\tilde{B}_2(t) = S(t)\tilde{B}_1(t), \qquad \tilde{C}_2(t) = \tilde{C}_1(t)S(t)^{-1}, \tag{3.14}$$

almost everywhere on $a \leqq t \leqq b$. Obviously, (3.14) implies that

$$B_2(t) = S(t)B_1(t), \qquad C_2(t) = C_1(t)S(t)^{-1}. \tag{3.15}$$

So to prove that θ_1 and θ_2 are similar, it remains to show that $P_2 = S(a)^{-1}P_1S(a)$. To do this, introduce

$$\hat{B}_\nu(t) = [B_\nu(t) \quad U_\nu(t)P_\nu U_\nu(t)^{-1}B_\nu(t) \quad \cdots \quad U_\nu(t)P^{n-1}U_\nu(t)^{-1}B_\nu(t)],$$

$$\hat{C}_\nu(t) = [C_\nu(t) \quad C_\nu(t)U_\nu(t)P_\nu U_\nu(t)^{-1} \quad \cdots \quad C_\nu(t)U_\nu(t)P^{n-1}U_\nu(t)^{-1}]^T,$$

for $\nu = 1, 2$. From (3.14) we may conclude that

$$\hat{B}_2(t) = S(t)\hat{B}_1(t), \tag{3.16}$$

$$U_2(t)P_2U_2(t)^{-1}\hat{B}_2(t) = S(t)(U_1(t)P_1U_1(t)^{-1})\hat{B}_1(t). \tag{3.17}$$

Now use that (3.13) implies that $U_2(t)S(a) = S(t)U_1(t)$, and insert this latter identity in the right-hand side of (3.17). So with (3.16) and (3.17) we come to:

$$(P_2 - S(a)P_1S(a)^{-1})U_2(t)\hat{B}_2(t) = 0, \quad a \leqq t \leqq b, \quad \text{a.e.}$$

But then

$$(P_2 - S(a)P_1S(a)^{-1})\left(\sum_{j=1}^{n} P_2^{j-1}\mathscr{C}(\theta_2)(P_2^*)^{j-1}\right)$$

$$= (P_2 - S(a)P_1S(a)^{-1})\int_a^b U_2(t)^{-1}\hat{B}_2(t)\hat{B}_2(t)^*(U_2(t)^{-1})^*\,dt = 0,$$

and we can use the controllability of the standard extension of θ_2 to show that $P_2 - S(a)P_1S(a)^{-1} = 0$.

The uniqueness of the similarity follows directly from the corresponding statement for causal systems. Indeed, let $S_0(t): X_1 \to X_2$, $a \leqq t \leqq b$, be a similarity transformation between θ_1 and θ_2. Then $U_2(t)S_0(a) = S_0(t)U_1(t)$ for $a \leqq t \leqq b$ and $P_2S_0(a) = S_0(a)P_1$. It follows that $S_0(t)$ is also a similarity transformation between Δ_1 and Δ_2. But Δ_1 and Δ_2 are causal systems, which are controllable and observable. Hence the similarity transformation is unique. □

Proof of Corollary 2.5. Similarity and dilation do not change the external coefficient and the sequence of weighting patterns. So we have only to prove the "if part" of Corollary 2.5.

To do this, assume that θ_1 and θ_2 have the same external coefficient and the same sequence of weighting patterns. We know (Theorem 2.4) that θ_1 is similar to a dilation

of an irreducible system, θ_{10} say. Also θ_2 is similar to a dilation of an irreducible system, θ_{20} say. Since similarity and dilation do not change the external coefficient and the sequence of weighting patterns, we conclude that θ_{10} and θ_{20} have the same external coefficient and the same sequence of weighting patterns. But then we can apply Theorem 2.2 to show that θ_{10} and θ_{20} are similar. Put $\theta_0 = \theta_{10}$. Then θ_0 is irreducible and θ_1 is similar to a dilation of θ_0. Further, since $\theta_{20} \simeq \theta_0$, we can use the next lemma to show that θ_2 is similar to a dilation of θ_0. □

LEMMA 3.3. *Let θ be a dilation of θ_0, and let $\tilde{\theta}_0$ be a system similar to θ_0. Then θ is similar to a dilation of $\tilde{\theta}_0$.*

Proof. Let $\theta = (A(t), B(t), C(t), D(t); N_1, N_2)_a^b$ and $\theta_0 = (A_0(t), B_0(t), C_0(t), D_0(t); N_1^{(0)}, N_2^{(0)})_a^b$. Further, let $S_0(t): X_0 \to \tilde{X}_0$, $a \leqq t \leqq b$, be a similarity between θ_0 and $\tilde{\theta}_0$. Thus the boundary value operators of $\tilde{\theta}_0$ are of the form

$$\tilde{N}_1^{(0)} = F_0 N_1^{(0)} S_0(a)^{-1}, \qquad \tilde{N}_2^{(0)} = F_0 N_2^{(0)} S_0(b)^{-1},$$

where $F_0: X_0 \to \tilde{X}_0$ is some invertible operator. Since θ is a dilation of θ_0, we may assume that (1.11), (1.12) and (1.13) hold true. Put $S(t) = I_{X_1} \oplus S_0(t) \oplus I_{X_2}$ and $F = I_{X_1} \oplus F_0 \oplus I_{X_2}$. Then $F, S(t): X_1 \oplus X_0 \oplus X_2 \to X_1 \oplus \tilde{X}_0 \oplus X_2$, $a \leqq t \leqq b$, are invertible operators. Consider $S(t)$, $a \leqq t \leqq b$, as a similarity transformation. When applied to θ, this transformation yields a new system $\tilde{\theta}$ of which the boundary value operators are given by

$$\tilde{N}_1 = FE^{-1} N_1 S(a)^{-1}, \qquad \tilde{N}_2 = FE^{-1} N_2 S(b)^{-1},$$

where E is the invertible operator appearing in (1.13). So $\theta \simeq \tilde{\theta}$, and it is easily checked that $\tilde{\theta}$ is a dilation of $\tilde{\theta}_0$. □

Proof of Theorem 2.3. Assume that among all systems with the same sequence of weighting patterns as θ the dimension of the state space of θ is as small as possible. Let θ_0 be the system introduced in Theorem 2.4. Since θ is similar to a dilation of θ_0, we know that θ and θ_0 have the same weighting patterns. So our hypotheses imply that the dimension of the state space X_0 of θ_0 is larger than or equal to the dimension of the state space X of θ. But from the construction of θ_0 it is clear that $X_0 \subset X$. Thus $X_0 = X$. But then $\theta \simeq \theta_0$, and we can use the irreducibility of θ_0 to conclude that θ is irreducible.

To prove the converse, assume that θ is irreducible. Let θ_1 be a system with well-posed boundary conditions and with the same sequence of weighting patterns as θ. Without loss of generality we assume that θ and θ_1 have the same external coefficient. Then, by Corollary 2.5, the systems θ and θ_1 are similar to dilations of the same irreducible system θ_0. The fact that θ is irreducible implies that $\theta \simeq \theta_0$. So the dimension of the state space of θ is equal to the dimension of the state space of θ_0 and the latter dimension is less than or equal to the dimension of the state space of θ_1. □

4. The time invariant case. In this section we review and specify the theory developed in §§ 2 and 3 for a time invariant system with boundary conditions:

$$\theta \quad \begin{cases} \dot{x}(t) = Ax(t) + Bu(t), & a \leqq t \leqq b, \\ y(t) = Cx(t) + Du(t), & a \leqq t \leqq b. \\ N_1 x(a) + N_2 x(b) = 0. \end{cases}$$

Such a system is characterized by the fact that the main coefficient, the input and output coefficients and the external coefficient do not depend on the time variable t. For brevity we write $\theta = (A, B, C, D; N_1, N_2)_a^b$.

The fundamental operator of the time invariant system θ is equal to $U(t) = e^{(t-a)A}$, $a \leqq t \leqq b$. Hence θ has well-posed boundary conditions if and only if $\det(N_1 e^{aA} + N_2 e^{bA}) \neq 0$, and in that case the canonical boundary value operator P of θ is given by

$$P = e^{aA}(N_1 e^{aA} + N_2 e^{bA})^{-1} N_2 e^{(b-a)A}. \tag{4.1}$$

For a time invariant system $\theta = (A, B, C, D; N_1, N_2)_a^b$ with well-posed boundary conditions, the (r, q)-extension $E_{rq}(\theta)$ is the system:

$$E_{rq}(\theta) \quad \begin{cases} \dot{x}(t) = Ax(t) + B^{(r)}(t)\tilde{u}(t), & a \leqq t \leqq b, \\ \tilde{y}(t) = C^{(q)}(t)x(t), & a \leqq t \leqq b, \\ N_1 x(a) + N_2 x(b) = 0, \end{cases}$$

where

$$B^{(r)}(t) = [B \quad e^{(t-a)A}Pe^{(a-t)A}B \quad \cdots \quad e^{(t-a)A}P^{r-1}e^{(a-t)A}B],$$
$$C^{(q)}(t) = [C \quad Ce^{(t-a)A}Pe^{(a-t)A} \quad \cdots \quad Ce^{(t-a)A}P^{q-1}e^{(a-t)A}]^T,$$

with P equal to the canonical boundary value operator of θ (see (4.1)). It follows that in general the (r, q)-extensions of θ are not time invariant, but depend on time. (An important special case when the (r, q)-extensions are time invariant systems will be discussed in the next section.) The fact that the (r, q)-extensions of θ (which include the standard extension) are not time invariant is probably one of the reasons that in many respects a time invariant system with well-posed boundary conditions behaves like a time varying system.

The rth weighting pattern of a time invariant system $\theta = (A, B, C, D; N_1, N_2)_a^b$ is the finite rank integral operator

$$(W_r(\theta)\varphi)(t) = Ce^{(t-a)A}P^{r-1}\int_a^b e^{(a-s)A}B\varphi(s)\,ds.$$

By expanding $e^{(t-a)A}$ and $e^{(a-s)A}$ into power series in $t-a$ and $s-a$, respectively, one sees that the weighting patterns are uniquely determined by the operators

$$CA^iP^{r-1}A^jB, \qquad i, j, r-1 = 0, 1, 2, \cdots. \tag{4.2}$$

The triple sequence (4.2) will be called the *sequence of moments* of θ.

Next we review the theorems of § 2 for the time invariant case. Let $\theta = (A, B, C, D; N_1, N_2)_a^b$ be a time invariant system with well-posed boundary conditions, and let P be its canonical boundary value operator. The condition of irreducibility for θ in Theorem 2.1(iii) may be replaced by the condition that the following block matrices have full rank:

$$[B \quad AB \quad \cdots \quad A^{n-1}B \quad PB \quad PAB \quad \cdots \quad PA^{n-1}B \quad \cdots \quad P^{d-1}B \quad P^{d-1}AB \quad \cdots \quad P^{d-1}A^{n-1}B],$$
$$[C \quad CA \quad \cdots \quad CA^{n-1} \quad CP \quad CAP \quad \cdots \quad CA^{n-1}P \quad \cdots \quad CP^{d-1} \quad CAP^{d-1} \quad \cdots \quad CA^{n-1}P^{d-1}]^T.$$

Here n is the dimension of the state space and (as before) d is the degree of the minimal polynomial of P.

In Theorem 2.2 (for time invariant systems) the weighting patterns can be replaced by the moments. Note, however, that the similarity transformation between two similar irreducible time invariant systems does not have to be time independent. For example,

consider the following two systems:

$$\theta_1 \quad \begin{cases} \dot{x}_1 = 0, \quad \dot{x}_2 = u, \quad 0 \leqq t \leqq 1, \\ y = x_2, \\ x_1(0) - x_2(0) + x_2(1) = 0, \qquad x_1(0) - x_1(1) - x_2(1) = 0; \end{cases}$$

$$\theta_2 \quad \begin{cases} \dot{x}_1 = x_1, \quad \dot{x}_2 = u, \quad 0 \leqq t \leqq 1, \\ y = x_2, \\ x_1(0) - x_2(0) + x_2(1) = 0, \qquad x_1(0) - x_1(1) - x_2(1) = 0. \end{cases}$$

One checks that

$$S(t) = \begin{pmatrix} e^t & 0 \\ 0 & 1 \end{pmatrix}, \qquad 0 \leqq t \leqq 1,$$

is a similarity transformation which transforms θ_1 into θ_2. Since the main coefficient of θ_1 is zero and the main coefficient of θ_2 is different from zero, there is no time independent similarity transforming θ_1 into θ_2. It is readily checked (using Theorem 2.1) that θ_1 and θ_2 are both irreducible.

In Theorem 2.3 for a time invariant system one may replace the weighting patterns by the moments. The same remark holds true for Corollary 2.5.

When applied to a time invariant system the similarity and reduction procedure of Theorem 2.4 does not always yield an irreducible system which is also time invariant. In fact, it may happen that a system θ is time invariant and that there is no irreducible time invariant system with the same sequence of weighting patterns as θ. For example, take

$$\Delta \quad \begin{cases} \dot{x}_1 = u, \quad \dot{x}_2 = x_1, \quad \dot{x}_3 = 0, \quad 0 \leqq t \leqq 1, \\ y = x_3, \\ x_1(0) = x_2(0) = 0, \qquad x_2(1) - x_3(1) = 0. \end{cases}$$

The first weighting pattern of the system Δ is the zero operator and all other weighting patterns are equal to the rank one integral operator

$$(W\varphi)(t) = \int_0^1 (s-1)\varphi(s)\, ds, \qquad 0 \leqq t \leqq 1.$$

It is not difficult to check that the following system Δ_0 has the same sequence of weighting patterns as Δ:

$$\Delta_0 \quad \begin{cases} \dot{x}_1 = (1-t)u, \quad \dot{x}_2 = (1-t)u, \quad 0 \leqq t \leqq 1, \\ y = -x_1 + x_2, \\ x_2(0) = 0, \qquad x_1(1) = 0. \end{cases}$$

Note that the state space of Δ has dimension three and that of Δ_0 is two-dimensional. Thus (see Theorem 2.3) the system Δ is not irreducible. On the other hand by applying Theorem 2.1 one easily sees that Δ_0 is an irreducible system. However, Δ_0 is time varying and one shows without difficulty that there is no time invariant system with a two-dimensional state space that has the same sequence of weighting patterns as Δ and Δ_0. So for the time invariant system Δ there is no irreducible time invariant system with the same sequence of weighting patterns as Δ.

5. Displacement systems. Deviating slightly from the terminology introduced in [9, § 1.6] we call a time invariant system θ with well-posed boundary conditions a

displacement system if the canonical boundary value operator P of θ commutes with the main coefficient A of θ. Causal (or anti-causal) time invariant systems are displacement systems. For a displacement system the kernel k of the input-output operator is a displacement kernel which means (see [13]) that k depends only on the difference of the arguments, i.e., $k(t, s) = h(t-s)$ for some function h. A similar remark holds true for the kernels of the weighting patterns of a displacement system. In fact, for a displacement system the kernel of the rth weighting pattern has the form

$$C e^{(t-a)A} P^{r-1} B, \qquad a \leqq t \leqq b, \quad a \leqq s \leqq b.$$

and its moments can be written as $CA^k P^{r-1} B$.

Let $\theta = (A, B, C, D; N_1, N_2)_a^b$ be a displacement system, and let P be its canonical boundary value operator. The (r, q)-extension $E_{rq}(\theta)$ of θ is the following system:

$$E_{rq}(\theta) \quad \begin{cases} \dot{x}(t) = Ax(t) + B^{(r)}\tilde{u}(t), & a \leqq t \leqq b, \\ \tilde{y}(t) = C^{(q)}x(t), & a \leqq t \leqq b, \\ N_1 x(a) + N_2 x(b) = 0, \end{cases}$$

where

$$B^{(r)} = [B \quad PB \quad \cdots \quad P^{r-1}B],$$
$$C^{(q)} = [C \quad CP \quad \cdots \quad CP^{q-1}]^T.$$

It follows that $E_{rq}(\theta)$ is again a displacement system. In particular, we see that the (r, q)-extensions (including the standard extension) of θ are time invariant. As a consequence for displacement systems the theory of similarity and reduction of § 2 resembles strongly the classical theory of causal time invariant systems.

THEOREM 5.1. *A similarity transformation between two irreducible displacement systems is time independent.*

Proof. For $\nu = 1, 2$ let $\theta_\nu = (A_\nu, B_\nu, C_\nu, D_\nu; N_1\ N_2)_a^b$ be an irreducible displacement system with state space X_ν. Let $S(t): X_1 \to X_2$, $a \leqq t \leqq b$, be a similarity transformation which transforms θ_1 into θ_2. We want to show that $S(t)$ does not depend on t. Put $F = S(a)$. Then $S(t) = e^{tA_2} F e^{-tA_1}$ and $P_2 F = F P_1$, where P_1 and P_2 are the canonical boundary value operators of θ_1 and θ_2, respectively. Since $P_1 A_1 = A_1 P_1$ and $P_2 A_2 = A_2 P_2$, it follows that

$$S(t) P_1 = P_2 S(t), \qquad a \leqq t \leqq b. \tag{5.1}$$

For $\nu = 1, 2$ consider the causal system $\Delta_\nu = (A_\nu, \tilde{B}_\nu, \tilde{C}_\nu, 0; I, 0)_a^b$, where

$$\tilde{B}_\nu = [B_\nu \quad P_\nu B_\nu \quad \cdots \quad P_\nu^{d-1} B_\nu],$$
$$\tilde{C}_\nu = [C_\nu \quad C_\nu P_\nu \quad \cdots \quad C_\nu P_\nu^{d-1}]^T.$$

Here d is the degree of the minimal polynomial of P_ν (which does not depend on ν). From (5.1) it follows that $S(t)$, $a \leqq t \leqq b$, is a similarity which transforms Δ_1 into Δ_2. The irreducibility of θ_1 and θ_2 implies that Δ_1 and Δ_2 are controllable and observable systems. Since, in addition, Δ_1 and Δ_2 are causal, we can use the classical theory of causal systems to show that $S(t)$ does not depend on t (see, e.g., [25, Thm. 2]). □

THEOREM 5.2. *A displacement system is a dilation of an irreducible displacement system.*

Proof. Let $\theta = (A, B, C, D; N_1, N_2)_a^b$ be a displacement system. Let P be the canonical boundary value operator of θ, and let n be the dimension of the state space

X of θ. Put

$$X_1 = \bigcap_{j,r=0}^{n-1} \operatorname{Ker} CA^jP^r, \qquad \tilde{X} = \bigvee_{j,r=0}^{n-1} \operatorname{Im} P^rA^jB, \tag{5.2}$$

where $\bigvee_{j=0}^{m} Z_j$ stands for the linear hull of the spaces $Z_0, \cdots, Z_m$. The operator P^n is a linear combination of the operators $I, P, \cdots, P^{n-1}$. So, if $x \in X_1$, then $CA^jP^nx = 0$ for $j = 0, \cdots, n-1$, and $Px \in X_1$. Also, $\operatorname{Im} P^nA^jB \subset \tilde{X}$, which implies that $P\tilde{X} \subset \tilde{X}$. Hence X_1 and $\tilde{X}$ are invariant under P. Since A commutes with P, we may interchange in (5.2) the positions of A and P. It follows that the spaces X_1 and $\tilde{X}$ are also invariant under A.

Let X_0 be a direct complement of $X_1 \cap \tilde{X}$ in $\tilde{X}$. Further, let X_2 be a direct complement of $X_1 \oplus X_0$ in X. So $X = X_1 \oplus X_0 \oplus X_2$ and $X_0 \subset \tilde{X} \subset X_1 \oplus X_0$. It follows that relative to the decomposition $X = X_1 \oplus X_0 \oplus X_2$ the operators A and P admit the following partitioning:

$$A = \begin{pmatrix} * & * & * \\ 0 & A_0 & * \\ 0 & 0 & * \end{pmatrix}, \qquad P = \begin{pmatrix} * & * & * \\ 0 & P_0 & * \\ 0 & 0 & * \end{pmatrix}. \tag{5.3}$$

Note that $\operatorname{Im} B \subset X_1 \oplus X_0$ and $X_1 \subset \operatorname{Ker} C$. So

$$B = \begin{pmatrix} * \\ B_0 \\ 0 \end{pmatrix}, \qquad C = [0 \quad C_0 \quad *] \tag{5.4}$$

relative to the decomposition $X = X_1 \oplus X_0 \oplus X_2$. Put

$$\theta_0 = (A_0, B_0, C_0, D; I - P_0, P_0 e^{-(b-a)A_0})_a^b.$$

Then θ is a dilation of θ_0. The fact that A commutes with P implies that A_0 commutes with P_0. Note that P_0 is the canonical boundary value operator of θ_0. Thus θ_0 is a displacement system. To prove that θ_0 is irreducible, it suffices to show that

$$\bigcap_{j,r=0}^{n-1} \operatorname{Ker} C_0A_0^jP_0^r = (0), \qquad \bigvee_{j,r=0}^{n-1} P_0^rA_0^jB_0 = X_0.$$

But these identities follow readily from the partitionings of A, P, B and C (see formulas (5.3) and (5.4)). □

Note that Theorem 5.2 implies that a displacement system is irreducible if and only if it does not admit a proper reduction. The next result is an immediate corollary of Theorem 5.2 and Theorem 2.2.

COROLLARY 5.3. *Two displacement systems θ_1 and θ_2 are dilations of similar irreducible systems if and only if θ_1 and θ_2 have the same external coefficient and the same sequence of moments.*

The results of this section can be extended to certain classes of time invariant systems with well-posed boundary conditions that are nondisplacement systems. For example, if $AP - PA = \alpha P + \beta A$ for some complex numbers α and β, then one can use the results of [20] (cf. [23]) to get the desired generalizations.

Acknowledgment. It is a pleasure to thank L. Lerer for useful discussions on the subject of this paper.

REFERENCES

[1] M. B. ADAMS, *Linear estimation of boundary value stochastic processes*, Ph.D. thesis, LIDS-TH-1295, Massachusetts Institute of Technology, Cambridge, MA, 1983.

[2] M. B. ADAMS, A. S. WILLSKY AND B. C. LEVY, *Linear estimation of boundary value stochastic processes, Part* I: *The role and construction of complementary models*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 803-811.

[3] ———, *Linear estimation of boundary value stochastic processes, Part* II: 1-D *smoothing problems*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 811-821.

[4] B. D. O. ANDERSON AND T. KAILATH, *Some integral equations with nonsymmetric separable kernels*, SIAM J. Appl. Math., 20 (1971), pp. 659-669.

[5] H. BART, I. GOHBERG AND M. A. KAASHOEK, *Minimal factorization of matrix and operator functions*, Operator Theory: Advances and Applications, Vol. 1, Birkhäuser Verlag, Basel, 1979.

[6] ———, *Wiener-Hopf integral equations, Toeplitz matrices and linear systems*, in Toeplitz Centennial, I. Gohberg, ed., Operator Theory: Advances and Applications, Vol. 4, Birkhäuser Verlag, Basel, 1982, pp. 85-135.

[7] ———, *Convolution equations and linear systems*, Integral Equations and Operator Theory, 5 (1982), pp. 283-340.

[8] H. D'ANGELO, *Linear Time Varying Systems*, Allyn and Bacon, Boston, MA, 1970.

[9] I. GOHBERG AND M. A. KAASHOEK, *Time varying linear systems with boundary conditions and integral operators*, I. *The transfer operator and its properties*, Integral Equations and Operator Theory, 7 (1984), pp. 325-391.

[10] ———, *Time varying linear systems with boundary conditions and integral operators*, II. *Similarity and reduction*, Report nr. 261, Dept. Mathematics and Computer Science, Vrije Universiteit, Amsterdam, 1984.

[11] T. KAILATH, *A view of three decades of linear filtering theory*, IEEE Trans. Inform. Theory, IT-20 (1974), pp. 146-181.

[12] ———, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[13] T. KAILATH, L. LJUNG AND M. MORF, *Generalized Krein-Levinson equations for the efficient computation of Fredholm resolvents of nondisplacement kernels*, in Topics in Functional Analysis, I. Gohberg and M. Kac, eds., Academic Press, New York, NY, 1978, pp. 169-184.

[14] R. E. KALMAN, *Mathematical description of linear dynamical systems*, this Journal, 1 (1963), pp. 152-192.

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

[16] H. G. KAPER, C. G. LEKKERKERKER AND J. HEJTMANEK, *Spectral methods in linear transport theory*, Operator Theory: Advances and Applications, Vol. 5, Birkhäuser Verlag, Basel, 1982.

[17] A. J. KRENER, *Acausal linear systems*, Proc. 18th IEEE Conference on Decision and Control, Ft. Lauderdale, FL, 1979.

[18] ———, *Boundary value linear systems*, Astérisque, 75/76 (1980), pp. 149-165.

[19] ———, *Smoothing of stationary cyclic processes*, Proc. MTNS, Santa Monica, CA, 1981, pp. 154-157.

[20] S. LEVIN, *Linear dynamical systems with partial derivatives*, Integral Equations Operator Theory, 7 (1984), pp. 118-137.

[21] C. V. M. VAN DER MEE, *Semigroup and factorization methods in transport theory*, Ph.D. thesis, Vrije Universiteit, Amsterdam, 1981, Mathematical Centre Tracts 146, Mathematical Centre, Amsterdam, 1981.

[22] H. H. ROSENBROCK, *State Space and Multivariable Theory*, Nelson, London, 1970.

[23] L. WAXMAN, *On characteristic operator-functions of Lie-algebras*, Kharkovskogo Univ., USSR Kharkov, 83 (1972), pp. 42-45; Integral Equations and Operator Theory, 6 (1983), pp. 312-318.

[24] L. WEISS, *On the structure theory of linear differential systems*, this Journal, 6 (1968), pp. 659-680.

[25] D. C. YOULA, *The synthesis of linear dynamical systems from prescribed weighting patterns*, SIAM J. Appl. Math., 14 (1966), pp. 527-549.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 5, September 1986

006

LARGE DEVIATIONS ESTIMATES FOR SYSTEMS WITH SMALL NOISE EFFECTS, AND APPLICATIONS TO STOCHASTIC SYSTEMS THEORY*

PAUL DUPUIS† AND HAROLD J. KUSHNER‡

Abstract. For typical stochastic systems (e.g., tracking systems), estimates of the behavior are hard to get. If the noise effects are small, then asymptotic methods are appealing (to get, for example, estimates of times required to lose track, etc.). In this paper, systems with small noise effects and wide bandwidth noise inputs are analyzed via large deviations methods. Inputs to many systems in control, communication or in physics are not "white noise", but of certain "wide bandwidth" types. Since large deviations results can be sensitive to the actual noise model used, working with a model that is close to the "physical" form is important. Several such models are dealt with, where the bandwidth is large, but the "intensity" small. For the models chosen, the action functionals turn out to be the same as for the "small Gaussian white noise" models. Thus, it is actually feasible to do computations with them. Estimates of the probability that the path lies in various sets are obtained. The formula for the mean escape time of the system from a set in which the "average dynamics" are stable is given, as are results on the most likely escape routes and the likely locations of the path on long time intervals. Such quantities, already available with small white noise model, are very helpful for understanding the long term systems behavior. The methods are applied to a detailed analysis of a phase locked loop tracking system (a special form of a nonlinear filter).

Key words. large deviations, phase locked loops, escape times, asymptotic analysis, wide bandwidth noise inputs, small noise effects

AMS(MOS) subject classifications. 60F10, 93E03, 94A05

1. Introduction.

The most common model used for large deviations (or small noise effects) asymptotic analysis in physics and engineering is the "small Gaussian white noise" model

$$dx^\varepsilon = b(x^\varepsilon)\,dt + \sqrt{\varepsilon}\sigma(x^\varepsilon)\Sigma^{1/2}\,dw, \qquad x \in R^n, \tag{1.1}$$

[1], [2], [3], [13], [14]. We insert the (nonsingular) matrix Σ in anticipation of future developments. Typically, the asymptotic analysis is used to estimate escape times from certain sets of interest, or the probability of inclusion of the path in a given set over some given time interval.

Suppose, in particular, that θ is an asymptotically stable point for $\dot{x} = b(x)$ and G is a neighborhood of θ, with a smooth boundary. Then we might be interested in an estimate of the probability of escape of $x^\varepsilon(\cdot)$ from G on a time interval $[0, T]$. Let $C_x[0, T]$ denote the set of continuous R^n-valued functions on $[0, T]$, with initial value x. Define the *H-function* (for (1.1)) and its *dual L* by

$$H(\alpha, x) = \alpha' b(x) + \alpha'\sigma(x)\Sigma\sigma'(x)\alpha/2, \tag{1.2}$$

$$L(\beta, x) = \sup_\alpha [\beta'\alpha - H(\alpha, x)] = \sup_\alpha [\alpha'(\beta - b(x)) + \alpha'\sigma(x)\Sigma\sigma'(x)\alpha/2]. \tag{1.3}$$

* Received by the editors November 1, 1984, and in revised form June 21, 1985.

† Division of Applied Mathematics, Lefschetz Center for Dynamical Systems, Brown University, Providence, Rhode Island 02912. The research of this author was supported in part by the Army Research Office under grant DAAG 29-84-K-0082 and the Office of Naval Research under grant N00014-83-K-0542.

‡ Division of Applied Mathematics, Lefschetz Center for Dynamical Systems, Brown University, Providence, Rhode Island 02912. The research of the author was supported in part by the Air Force Office of Scientific Research under grant 81-0116, the National Science Foundation under grant ECS 82-11476, and the Office of Naval Research under grant N00014-83-K-0542.

If $\sigma(x)\Sigma\sigma'(x) = \tilde{\Sigma}(x)$ is invertible, then

$$L(\beta, x) = (\beta - b(x))'\tilde{\Sigma}^{-1}(x)(\beta - b(x))/2. \tag{1.4a}$$

If

$$\tilde{\Sigma} = \begin{bmatrix} \tilde{\Sigma}_1 & 0 \\ 0 & 0 \end{bmatrix},$$

with $\tilde{\Sigma}_1$ invertible then (write $\beta = (\beta_1, \beta_2)$, $b = (b_1, b_2)$)

$$L(\beta, x) = \begin{cases} (\beta_1 - b_1(x))'\tilde{\Sigma}_1^{-1}(x)(\beta_1 - b_1(x))/2 & \text{for } \beta_2 = b_2(x), \\ \infty & \text{otherwise.} \end{cases} \tag{1.4b}$$

Define $\tau_G(\phi) = \inf\{t: \phi(t) \notin G\}$ for $\phi(\cdot) \in C_x[0, \infty)$. All $\phi(\cdot)$ below are in either $C_x[0, T]$ or $C_x[0, \infty)$. Define the action functional $S(\phi, T)$ by

$$S(\phi, T) = \begin{cases} \int_0^T L(\dot{\phi}(s), \phi(s))\, ds & \text{for } \phi(\cdot) \text{ absolutely continuous,} \\ \infty & \text{otherwise.} \end{cases}$$

Finally, define $S(\phi) = S(\phi, \tau_G(\phi))$. Let $A \subset C_x[0, T]$ with A^0 and $\bar{A}$ denoting the interior and closure of A, respectively. Let A_G denote the set of functions $\phi(\cdot)$ in $C_\theta[0, T]$ such that $\phi(t) \in \partial G$ for some $t < \infty$, and define $\tau_G^\varepsilon = \inf\{t: x^\varepsilon(t) \notin G\}$.

Typical results for (1.1) are (under broad conditions in G, $b(\cdot)$ *and* $\sigma(\cdot)$, there is equality in (1.5b)), for $x \in G$,

$$\lim_\varepsilon \varepsilon \log E_x \tau_G^\varepsilon = \inf_{\phi \in A_G} S(\phi), \tag{1.5a}$$

$$\begin{aligned} -\inf_{\phi \in A^0} S(\phi, T) &\leqq \underline{\lim}_\varepsilon \varepsilon \log P_x\{x^\varepsilon(\cdot) \in A\} \\ &\leqq \overline{\lim}_\varepsilon \varepsilon \log P_x\{x^\varepsilon(\cdot) \in A\} \leqq -\inf_{\phi \in \bar{A}} S(\phi, T). \end{aligned} \tag{1.5b}$$

Thus, obtaining the estimates in (1.5) requires solving a variational problem. With model (1.1), the integrand $L(\dot{\phi}, \phi)$ can be written explicitly, which is not usually the case for models which use other than "white noise". This simplicity underlies much of the popularity of (1.1).

By defining $u = (\sigma(\phi)\Sigma^{1/2})^{-1}(\dot{\phi} - b(\phi))$ in (1.3), we see that the variational problem of $\inf_{\phi \in A} \int_0^T L(\dot{\phi}, \phi)\, ds$ is equivalent to the "optimal control" problem:

$$\inf_{\phi \in A} \int_0^T |u(t)|^2\, dt, \qquad \dot{\phi} = b(\phi) + \sigma(\phi)\Sigma^{1/2}u. \tag{1.5c}$$

This equivalence will be useful in § 4.

From the point of view of applications to physical problems, the model (1.1) has several deficiencies. Owing to the small noise effects and the stability of $\dot{x} = b(x)$, the escape phenomena of interest requires a long time to occur, and escape depends on a burst of "unusual" noise. No physical noise is actually "white Gaussian", and the estimates can be quite sensitive to the actual noise model; e.g., let

$$\dot{x} = b(x^\varepsilon) + \sigma(x^\varepsilon)\xi^\varepsilon,$$

where $\int_0^t \xi^\varepsilon(s)\, ds/\sqrt{\varepsilon}$ converges to a Wiener process. Under quite reasonable conditions, the estimates for the quantities in (1.5) for this model can differ considerably from

those obtained for (1.1) [4]. See also [5] which presents some continuity results on the estimates with respect to the statistics of $\xi^\varepsilon(\cdot)$.

In this paper, we replace $w(\cdot)$ in (1.1) by more realistic noise processes—which are approximations to a Wiener process in an appropriate sense. We choose models such that the above action functional $S(\phi, T)$ can be used; i.e., the small Gaussian white noise approximation is valid. The treatment of applications with realistic noise processes (in, for example, stochastic systems theory) seems (to date) to require the type of analysis which we do here. The noise models are of the type which arise in numerous applications in control and communications theory, and in physics. A detailed application to a phase locked loop (a special form of nonlinear filter) problem is given in § 5. Owing to the possible sensitivity of the results to the model, we treat the system without simplifying it—and then show that the "Gaussian white noise" model for a simpler form is indeed valid, under the stated assumptions on the input noise.

We work with systems of the form

$$\dot{x}^\rho = b(x^\rho) + \varepsilon\sigma(x^\rho)\xi^\gamma(t) \tag{1.6}$$

where $\rho = (\varepsilon, \gamma)$, and $b(\cdot)$ and $\sigma(\cdot)$ are bounded, together with their first partial derivatives. Let τ_G^ρ denote the escape time of $x^\rho(\cdot)$ from G. Two specific stationary models will be used for the $\xi^\gamma(\cdot)$; in either case the integral of the driving noise $\int_0^t \xi^\gamma(s)\,ds$ converges weakly to a Wiener process $w(\cdot)$. In the phase locked loop example, we illustrate how the model might be easily modified to handle a variety of slightly different cases.

We will take limits as $\varepsilon \to 0$ and $\gamma \to 0$ simultaneously (i.e., as $\rho \to 0$). Thus $\xi^\gamma(t) \to$ "white noise" (loosely speaking) and its "weight" ε converges to zero. This allows a more realistic modelling than we could get with only a single parameter, since it allows us to vary the "bandwidth" and "intensity" of the noise more or less independently.

In [1, p. 132], the system $x^\varepsilon = b(x^\varepsilon, \varepsilon\xi)$ is dealt with, where $\xi(\cdot)$ is a fixed Gaussian process, and the analogue of (1.5a) is obtained, but the result or technique are not appropriate for our model or for applications such as those in § 5. In such cases, we are concerned with "small noise effects", and these can be caused by either (a) small noise (multiply a *fixed* process by ε), or (b) by using a process $\xi(t/\gamma)$, for small γ (wide bandwidth), or by a *combination* of both effects (e.g., use $\varepsilon\xi(t/\gamma)/\sqrt{\gamma}$). Such combinations are important in applications.

In § 2, we define the specific noise models which are to be used, and in § 3, the appropriate H-functionals are obtained. The basic limit theorems are stated and proved in § 4. Analogues of both (1.5a) and (1.5b) are obtained, as well as estimates of the locations of the points of escape from G.

The proofs depend on several (large derivations type) estimates for the noise processes—and for related dynamical systems, and these are derived in the appendices. We also obtain some results on the "jumping" of the process from invariant set to invariant set (of $\dot{x} = b(x)$), analogous to the results in [1, Chap. 6]. These results are useful for the problem of the asymptotic distribution of $x^\varepsilon(\cdot)$, for large t and small ε, when the system $\dot{x} = b(x)$ has several stable points or invariant sets.

In § 5, the techniques of § 4 are applied to a phase locked loop problem, and it is shown, under reasonable conditions, that a "small white noise" model is appropriate, and that the "high frequency" terms which appear in the actual dynamical equations can be neglected, in the sense that the correct action functionals are actually of the "small white noise" type, without the "high frequency" terms. These results validate the types of approximations and simplifications made in the PLL model analyzed in [3].

2. The noise models. The "wide-bandwidth" Gaussian noise process of Model I is a standard one in many applications. With Model II, we attempt to get close to the "physical" noise in many applications in control and communication theory, and in applications in physics. The idea is that the noise is "close to" impulsive but has a "short memory". Thus, we use a scaled and filtered impulsive noise process. Consider, for example, "shot" or "impulsive" noise in an electrical circuit. Owing to the sensitivity of the large deviations estimates to the noise model (see [4], for example), it is important to use a model which is as close to the "physical" situation as possible—even though for standard applications (not of a "large deviations" or "small noise" type) a Gaussian process approximation is usually adequate. With scalings other than that used in Model II below, the limit action functional would not be of the "small white noise" type.

Model I. Let $\xi^\gamma(t) = \xi(t/\gamma)/\sqrt{\gamma}$, where

$$d\xi = A\xi\, dt + B\, dw, \quad A \text{ stable}, \quad w(\cdot) \text{ standard}. \tag{2.1}$$

Then

$$W^\gamma(t) \equiv \int_0^t \xi^\gamma(s)\, ds$$

converges weakly to a Wiener process with covariance matrix

$$\tilde{\Sigma}_1 + \tilde{\Sigma}_1' = \Sigma_1 = \int_{-\infty}^{\infty} E\xi^\gamma(t)\xi^\gamma(0)\, dt = \int_{-\infty}^{\infty} E\xi(t)\xi(0)\, dt$$

where

$$\tilde{\Sigma}_1 = -A^{-1}\int_0^\infty (\exp At)BB'(\exp A't)\, dt.$$

An integration by parts yields $\Sigma_1 = (A^{-1}B)(A^{-1}B)'$.

Model II. $\xi^\gamma(t) = \bar{\xi}^\gamma(t/\gamma)/\gamma$, where

$$d\bar{\xi}^\gamma = A\bar{\xi}^\gamma\, dt + d\bar{J}^\gamma, \tag{2.2}$$

where $\bar{J}^\gamma(\cdot)$ is a jump Markov process with jump rate $\mu_\gamma\gamma$, and the jumps have the distribution of a random variable ψ^γ, where $E\psi^\gamma = 0$, $\operatorname{var}\psi^\gamma = \nu_\gamma$ and $\nu_\gamma\mu_\gamma = C_0$, a constant matrix. In order to obtain the desired "Gaussian" limit form for the H-functional an additional condition on the higher moments of ψ^γ, and on the relation between ε and γ is required: for some $\alpha > 0$ and $k < \infty$

$$\begin{aligned} &\mu_\gamma = O(\gamma^{-2-\alpha}), \qquad \nu_\gamma = O(\gamma^{2+\alpha})C_0, \\ &E|\psi_i^\gamma|^{2n} \leqq n!k^n|v_\gamma|^n, \\ &\gamma^{1+\alpha/2}/\varepsilon \to 0 \quad \text{as } \rho \to 0. \end{aligned} \tag{2.3}$$

Define

$$\tilde{\Sigma}_2 = \int_0^\infty (\exp At)\, dt \int_0^\infty (\exp As)C_0 \exp A's\, ds.$$

Then

$$W^\gamma(t) \equiv \int_0^t \xi^\gamma(s)\, ds$$

converges weakly to a Wiener process with covariance

$$\Sigma_2 = \int_{-\infty}^{\infty} E\xi^\gamma(t)\xi^\gamma(0)'\,dt = \tilde{\Sigma}_2 + \tilde{\Sigma}_2' = A^{-1}C_0(A')^{-1}.$$

Large deviations results for these noise models cannot be obtained directly from existing works. Freidlin and Ventsel [1] use (1.1), Ventsel [6] requires a Markov property for $x^\rho(\cdot)$, and other conditions which are not necessarily satisfied here. Azencott and Ruget [9] also require a Markov property on $x^\rho(\cdot)$. In Freidlin [7], a bounded noise process (and a single scaling parameter) is used. Except for a special case, we concentrate on the analogue of (1.5b) for $x^\rho(\cdot)$, and also obtain estimates of the type

$$\lim_\rho \varepsilon^2 \log E_x \tau_G^\rho = \inf_{\phi \in A_G} S(\phi). \tag{2.4}$$

The usual arguments for the equality in (2.4) (as in [1]) depend heavily on a Markov property and the use of Markov stopping times. In order to prove (2.4), we need various "large deviations" type estimates for $W^\gamma(\cdot)$, uniform in the initial condition $\xi^\gamma(0)$ in certain sets. Regrettably, this complicates the development.

3. The H-functions for the noise models.

Model I. Define

$$H(\alpha, x) = \alpha' b(x) + H_0(\alpha, x), \tag{3.1}$$

where H_0 is defined by (for any $t>0$ and with the *definition* $\lambda_\rho = \varepsilon^2$),

$$\begin{aligned} tH_0(\alpha, x) &= \lim_\rho \lambda_\rho \log E \exp \varepsilon \int_0^t \alpha'\sigma(x)\xi^\gamma(s)\,ds/\lambda_\rho \\ &= \alpha'\sigma(x)\int_{-\infty}^{\infty} E\xi^\gamma(s)\xi^\gamma(0)'\,ds\,\sigma'(x)\alpha/2 \\ &= \alpha'\sigma(x)\Sigma_1\sigma'(x)\alpha/2. \end{aligned} \tag{3.2}$$

If $\alpha(\cdot)$ is a piecewise constant function, then for each $T>0$,

$$\int_0^T H_0(\alpha(s), x)\,ds = \lim_\rho \lambda_\rho \log E \exp \varepsilon \int_0^T \alpha'(s)\sigma(x)\xi^\gamma(s)\,ds/\lambda_\rho. \tag{3.3}$$

The normalization sequence $\{\lambda_\rho\}$ does not depend on γ.

Model II. $H(\alpha, x)$ is defined by (3.1), and $H_0(\alpha, x)$ by the limit in (3.2), where also $\lambda_\rho = \varepsilon^2$; but here we have (for any $t>0$)

$$\begin{aligned} tH_0(\alpha, x) &= \lim_\rho \lambda_\rho \log E \exp \varepsilon\alpha' \int_0^t \sigma(x)\xi^\gamma(s)\,ds/\lambda_\rho \\ &= \lim_\rho \lambda_\rho \log E \exp \varepsilon\alpha'\sigma(x)\int_0^{t/\gamma} ds \int_\tau^{t/\gamma} \exp A(s-\tau)\,d\bar{J}^\gamma(\tau)/\lambda_\rho \\ &= \lim_\rho \lambda_\rho \log E \exp \varepsilon\alpha'\sigma(x)\Omega_0 \int_0^{t/\gamma} d\bar{J}^\gamma(\tau)/\lambda_\rho, \end{aligned} \tag{3.4}$$

where $\Omega_0 = \int_0^\infty \exp As\,ds = -A^{-1}$. Continuing, we have

$$H_0(\alpha, x) = \lim_\rho \lambda_\rho \mu_\gamma [E \exp \varepsilon\alpha'\sigma(x)\Omega_0\psi^\gamma/\lambda_\rho - 1]. \tag{3.5}$$

Under (2.3),

$$H_0(\alpha, x) = \alpha'\sigma(x)A^{-1}(\mu_\gamma\nu_\gamma)(A')^{-1}\sigma'(x)\alpha/2 = \alpha'\sigma(x)\Sigma_2\sigma'(x)\alpha/2, \tag{3.6}$$

(3.6) is of the Gaussian form, and it is the form which one would obtain if εdW were used in (1.6), where $W(\cdot)$ is the Wiener process limit (weak convergence sense) of $\{W^\gamma(\cdot)\}$. $H_0(\alpha, x)$ also satisfies (3.3) for any piecewise constant function $\alpha(\cdot)$.

Remark. By the above calculations, the H, L and S functionals are those of § 1, where the Σ there is either Σ_1 or Σ_2, depending on the noise model.

4. The limit theorems. Recall the definition of $L(\cdot,\cdot)$ in (1.3) and $S(\phi, T)$ in § 1.

THEOREM 4.1. *Under the assumptions in* §§ 1 *and* 2 *on the noise models and on* $\sigma(\cdot)$ *and* $b(\cdot)$,

$$\begin{aligned} -\inf_{\phi\in A^0} S(\phi, T) &\leqq \underline{\lim}_\rho \lambda_\rho \log P_x\{x^\rho(\cdot)\in A\} \\ &\leqq \overline{\lim}_\rho \lambda_\rho \log P_x\{x^\rho(\cdot)\in A\} \\ &\leqq -\inf_{\phi\in\bar{A}} S(\phi, T). \end{aligned} \tag{4.1}$$

Further, if $0<\delta<\frac{1}{2}$, *then* (4.1) *holds uniformly for* x *in a compact set and* $|\xi(0)|\leqq\gamma^{-\delta}\varepsilon^{-1}$ (*Model* I) *or* $|\bar{\xi}^\gamma(0)|\leqq\varepsilon^{-1}\gamma^{-\delta+1/2}$ (*Model* II).

The proof of this theorem is long, and is provided in Appendix I (Theorem A1.1 there).

Define $S^* = \inf_{\phi,T} S(\phi, T)$, where the infimum is over all T, $\phi\in C_\theta[0, T]$ satisfying $\phi(T)\in\partial G$. By (1.5c), this is equivalent to the optimal control problem

$$\inf \int_0^{\tau(\phi)} |u(t)|^2\, dt, \tag{4.2a}$$

$$\dot{\phi} = b(\phi)+\sigma(\phi)\Sigma^{1/2}u, \qquad \phi(0)=x, \tag{4.2b}$$

where $\tau(\phi)$ is the first hitting time of ∂G.

The mean exit time problem. We next prove a result for the mean exit time from a set G. The proof is an adaptation of that for the "white noise" case in [1], but is more complex since $x^\rho(\cdot)$ is neither Markovian, nor is the noise bounded. The proof in [1] requires that $\sigma(x)\Sigma\sigma'(x)$ be uniformly positive definite (the "nondegenerate" case). This is a very serious restriction in applications. In order to avoid it, we introduce the following controllability assumption for (4.2).

Assumption 4.1. There is an $M_1<\infty$ such that for small ε_1 and each x, $y\in N_{\varepsilon_1}(\theta)$ (ε_1-neighborhood of θ) there is a $u(\cdot)$ such that $|u(t)|\leqq M_1$ and for the *corresponding trajectory* (4.2b), $\phi(0)=x$, $\phi(t_1)=y$, *where* $t_1\to 0$ as $\varepsilon_1\to 0$.

This assumption always holds in the nondegenerate case. The assumption was used in [5, § 4 and elsewhere] to construct "approximately" optimal trajectories for certain "degenerate" problems.

THEOREM 4.2. *Let* θ *be an asymptotically stable equilibrium position of* $\dot{x}=b(x)$, *and assume* (*bounded*) G *is attracted to* θ. *Furthermore, assume that the boundary* ∂G *is smooth and* $(b(x), n(x))<0$ *for* $x\in\partial G$, *where* $n(x)$ *is the exterior normal of the boundary of* G, *and let* $S^*<\infty$. *Then, under Assumption* 4.1, *there exists a set with probability greater than* $1-\exp(-M_\rho/\lambda_\rho)$, *where* $M_\rho\to\infty$ *as* $\rho\to 0$ *such that*

$$\lim_\rho \lambda_\rho \log E_x\tau^\rho I_{A_\rho} = S^* \tag{4.3}$$

where τ^ρ *is the first exit time of* $x^\rho(\cdot)$ *from* G.

Remark. The proof of Theorem 4.2 is an adaptation of that for the "white noise" case in [1], but is more complex, since $x^\rho(\cdot)$ is not Markovian. Equation (4.3) differs slightly from the "usual" result, due to the presence of the "exceptional" set $\Omega - A_\rho$, of "exponentially small probability". This is hardly a restriction in applications.

Proof. Part I. First we show that for each $d>0$, there is a set A_ρ and a $\rho_0>0$ such that $P\{A_\rho\} \geqq 1-\exp(-M_\rho/\lambda_\rho)$, where $M_\rho \to \infty$ as $\rho \to 0$ and for $\rho < \rho_0$

$$\lambda_\rho \log E_x I_{A_\rho} \tau^\rho \leqq S^* + d \tag{4.4}$$

(d does not depend on $\xi(0)$ for $|\xi(0)| \leqq \gamma^{-\delta}\varepsilon^{-1}$, $\delta \in (0, \frac{1}{2})$).

As is discussed in [1, p. 124], we may choose positive μ, h, T_1, T_2, $k<1$, such that the following conditions hold:

(a) All solutions of $\dot{x} = b(x)$ starting in $G \cup \partial G$ satisfy

$$|\theta - x(t)| \leqq k\mu$$

for $t \geqq T_1/2$.

(b) For every point in the set

$$D = \{x: |\theta - x| \leqq \mu\}$$

there is $\phi^x(\cdot) \in C_x[0, \infty)$ such that $\phi^x(0) = x$, $\phi^x(t)$ reaches the exterior of the h-neighborhood of G at time $T(x) \leqq T_2$ and

$$S(\phi^x, T(x)) < S^* + \frac{d}{2}.$$

In [1], the $\phi^x(\cdot)$ do not hit the $k\mu$-neighborhood of θ after exit from D, but this is not needed.

The construction of such $\phi^x(\cdot)$ is clear in the nondegenerate case. For the degenerate case, the controllability assumption is needed if the d-optimal for S^* are tangent to ∂G at the first hitting time. It also guarantees that the particular points in D to which the trajectories (in the cycles constructed below) return are irrelevant in the analysis—since we can move "quickly and cheaply" between any point in D for small μ. Henceforth, we simply assume the existence of the paths specified in (a) and (b) above.

We prove the theorem for Model I. The proof for Model II is entirely analogous. Let $0 < \delta_1 < \delta_2 < \frac{1}{2}$. By Theorem 4.1 there is $\rho_0 > 0$ such that $\rho < \rho_0$ implies, uniformly in $y \in D$ and in $|\xi(0)| \leqq \varepsilon^{-1}\gamma^{-\delta_1}$,

$$P_{y,\xi(0)}\{\sup_{0 \leqq t \leqq T(y)} |x^\rho(t) - \phi^y(t)| < h\} \geqq \exp(-(S^* + d)/\varepsilon^2). \tag{4.5}$$

which implies that

$$P_{y,\xi(0)}\{\tau^\rho < T_2\} \geqq \exp(-(S^* + d)/\varepsilon^2).$$

In order to complete the proof, we need the following auxiliary lemma.

LEMMA 4.1. *There is $\rho_0 > 0$ such that for $\rho \leqq \rho_0$ and uniformly in $x \in G$ and in $|\xi(0)| \leqq \gamma^{-\delta_1}\varepsilon^{-1}$, we have*

$$P_{x,\xi(0)}\{\tau^\rho < T_1 + T_2\} \geqq \exp(-(S^* + d)/\varepsilon^2).$$

Proof of Lemma. Let τ_1 denote the time of first entrance into a μ-neighborhood of θ. For any $M < \infty$ and small ρ

$$P_{x,\xi(0)}\{\tau^\rho \leqq T_1 + T_2\} \geqq E_{x,\xi(0)}[P_{x^\rho(\tau_1)}\{\tau^\rho < T_2\} I_{\{\tau_1 \leqq T_1\}} \big| |\xi(\tau_1/\gamma)| \leqq \gamma^{-\delta_2}\varepsilon^{-1}](1 - e^{-M/\varepsilon^2})$$
$$- P_{x,\xi(0)}\{\sup_{0 \leqq t \leqq T_1/\gamma} |\xi(t)| \geqq \gamma^{-\delta_2}\varepsilon^{-1}\}.$$

By Lemma A2.1, uniformly in $|\xi(0)| \leqq \gamma^{-\delta_1}\varepsilon^{-1}$, the second term on the right is less than $\exp(-M/\varepsilon^2)$ for any given M if ρ is small enough. By (4.5), and the fact that the conditional probability that $\{\tau_1 \leqq T_1\}$ goes to unity as $\rho \to 0$, the first term is greater than (for small ρ)

$$\tfrac{1}{2}\exp(-(S^*+d/2)/\varepsilon^2), \tag{4.6}$$

and the lemma is proved.

Define $T = T_1 + T_2$ and the sets $B_\rho = \{\omega: \sup_{0\leqq t\leqq \exp(S^*+d)/\varepsilon^2} |\xi(t)| < \gamma^{-\delta_1}\varepsilon^{-1}\}$ and $B_\rho(n) = \{\omega: \sup_{0\leqq t\leqq nT} |\xi(t)| < \gamma^{-\delta_1}\varepsilon^{-1}\}$.

Select (by Lemma A2.1) $\rho_0 > 0$ such that for $\rho \leqq \rho_0$,

$$P\{\Omega - B_\rho\} \leqq \exp(-M/\varepsilon^2). \tag{4.7}$$

By the Markov property of $(x^\rho(\cdot), \xi^\gamma(\cdot))$, for $nT \leqq \exp(S^*+d)/\varepsilon^2$ we have

$$\begin{aligned} E_x P\{\tau^\rho \geqq nT\} I_{B_\rho} &\leqq E_x I_{\{\tau^\rho \geqq nT\}} I_{\{\tau^\rho \geqq nT-T\}} I_{B_\rho(n-1)} \\ &= E_x P_x\{\tau^\rho \geqq nT | \tau^\rho \geqq nT - T, \omega \in B_\rho(n-1)\} I_{\{\tau^\rho \geqq nT-T\}} I_{B_\rho(n-1)}. \end{aligned}$$

We have

$$\begin{aligned} P_x\{\tau^\rho \geqq nT | \tau^\rho \geqq nT - T, \omega \in B_\rho(n-1)\} &\leqq \sup_{\substack{y\in G \\ |\xi(0)|\leqq \varepsilon^{-1}\gamma^{-\delta_1}}} P_{y,\xi(0)}\{\tau^\rho \geqq nT\} \\ &= 1 - \inf_{\substack{y\in G \\ |\xi(0)|\leqq \varepsilon^{-1}\gamma^{-\delta_1}}} P_{y,\xi(0)}\{\tau^\rho < T\} \\ &\leqq 1 - \tfrac{1}{2}\exp(-(S^*+d/2)/\varepsilon^2). \end{aligned}$$

Then, iterating yields

$$E_x P\{\tau^\rho \geqq nT\} I_{B_\rho} \leqq (1 - \tfrac{1}{2}\exp(-(S^*+d/2)/\varepsilon^2)^n). \tag{4.8}$$

For $nT = \exp(S^*+d)/\varepsilon^2$, we find that for small ρ (so that $(\exp(d/2\varepsilon^2)/T \geqq 2M/\varepsilon^2)$)

$$\begin{aligned} E_x P\{\tau^\rho \geqq \exp(S^*+d)/\varepsilon^2\} I_B &\leqq [1 - \tfrac{1}{2}\exp(-(S^*+d/2)/\varepsilon^2)]^{[\exp(S^*+d)/\varepsilon^2]/T} \\ &\leqq \exp(-[\exp d/2\varepsilon^2]/2T) \leqq \exp - M/\varepsilon^2. \end{aligned}$$

Define $\tilde{B}_\rho = \{\omega: \tau^\rho \leqq \exp(S^*+d)/\varepsilon^2\}$ and $A_\rho = B_\rho \cap \tilde{B}_\rho$. Then

$$P\{A_\rho\} \geqq 1 - 2\exp(-M/\varepsilon^2)$$

and

$$\begin{aligned} E_x \tau^\rho I_{A_\rho} &\leqq T \sum_0^{\exp(S^*+d)/\varepsilon^2} P_x\{\tau^\rho \geqq nT\} \\ &\leqq T \sum_0^\infty [1 - \exp(-(S^*+d/2)/\varepsilon^2)]^n \qquad (4.9) \\ &\leqq \exp(S^*+d)/\varepsilon^2. \end{aligned}$$

Proof. Part II. We now prove the reverse inequality to (4.4), namely that *for each* $d > 0$, *there is a* $\rho_0 > 0$ *such that for* $\rho \leqq \rho_0$,

$$\lambda_\rho \log E_x \tau^\rho \geqq S^* - d, \qquad x \in G. \tag{4.10}$$

It is convenient to separate the proof into a few lemmas. We first prove

LEMMA 4.2. *Given $\delta > 0$, $M < \infty$ there is a $\rho_0 > 0$ such that for $\rho \leqq \rho_0$ and $x \in G$*

$$P_x\{\sup_{0 \leqq t \leqq \tau^\rho/\gamma} |\xi(t)| \geqq \gamma^{-\delta}\varepsilon^{-1}\} \leqq \exp(-M/\varepsilon^2). \tag{4.11}$$

If the probability in (4.11) *is conditioned on $\xi(0)$, then the result holds uniformly for $|\xi(0)| \leqq \gamma^{-\delta_1}\varepsilon^{-1}$ for any $\delta > \delta_1$.*

Proof of the lemma. By Part I of the proof, for any $M_1 < \infty$ and $h > 0$, there is a $\rho_0 > 0$ such that $\rho \leqq \rho_0$ implies that for $x \in G$, and $|\xi(0)| \leqq \gamma^{-\delta_1}\varepsilon^{-1}$,

$$P_{x,\xi(0)}\{\tau^\rho > \exp M_1/\varepsilon^2\} \leqq (\exp(-M_1/\varepsilon^2)) \exp(S^* + d)/\varepsilon^2.$$

It follows from Lemma A2.1 that for each $M_2 < \infty$ and small enough ρ (and uniformly in the desired $|\xi(0)|$ set),

$$P_{\xi(0)}\{\sup_{t \leqq (\exp M_1/\varepsilon^2)/\gamma} |\xi(t)| \leqq \gamma^{-\delta}\varepsilon^{-1}\} \leqq (\exp(-M_2/\varepsilon^2)) \exp M_1/\varepsilon^2.$$

Using the above two estimates and selecting M_1 and $M_2 - M_1$ large enough and ρ small enough yields the lemma

LEMMA 4.3. *Let $\delta \in (0, \frac{1}{2})$ and $\alpha > 0$ (and small). There are $c > 0$ and $T_0 < \infty$ such that for all $T < \infty$ and $x \in \bar{G} - N_\alpha(\theta)$ and $|\xi(0)| \leqq \gamma^{-\delta}\varepsilon^{-1}$ we have for small enough ε*

$$P_x\{\tau^\alpha > T\} \leqq \exp(-c(T - T_0)/\varepsilon^2), \tag{4.12}$$

where $\tau^\alpha = \inf\{t: x^\rho(t) \notin G - N_\alpha(\theta)\}$.

Remark on the proof. Using the bound on the noise $|\xi(t)| \leqq \gamma^{-\delta}\varepsilon^{-1}$ on $[0, \tau^\rho/\gamma]$ (w.p. $\geqq 1 - \exp - M/\varepsilon^2$), the proof is a simple modification of that of [1, Chap. 4, Lemma 2.2].

We are now ready to start the proof of (4.10). Following the idea in [1], fix small $\mu > 0$ and define $\Gamma_i = \{x: d(x, \theta) = i\mu\}$, for $i = 1$ and 2. Define the Markov times $\{\tau_i, \sigma_i\}$ by $\tau_0 = 0$ and

$$\sigma_i = \inf\{t > \tau_i: x^\rho(t) \in \Gamma_2\},$$
$$\tau_i = \inf\{t > \sigma_{i-1}: x^\rho(t) \in \Gamma_1 \cup \partial G\}.$$

Let $0 < \delta_1 < \delta_2 < \frac{1}{2}$ and let B denote the set $\{\xi: |\xi| \leqq \gamma^{-\delta_1}\varepsilon^{-1}\}$. For simplicity, we always will assume that the various initial conditions $\xi(t) \in B$ for $t < \tau^\rho$. This is true w.p. $\geqq 1 - \exp(-M/\varepsilon^2)$ for small ρ by Lemma 4.2. If (4.10) holds under this assumption, then it holds as stated.

For $x \in \Gamma_1$,

$$\begin{aligned} P_{x,\xi(0)}\{x^\rho(\tau_1) \in \partial G\} \leqq \max_{\substack{y \in \Gamma_2 \\ \xi(0) \in B}} [&P_{y,\xi(0)}\{\tau^\rho = \tau_1 < T, |\xi(\sigma_0/\gamma)| \leqq \gamma^{-\delta_2}\varepsilon^{-1}\} \\ &+ P_{x,\xi(0)}\{\tau^\rho = \tau_1 \geqq T, |\xi(\sigma_0/\gamma)| \leqq \gamma^{-\delta_2}\varepsilon^{-1}\} \\ &+ P_{\xi(0)}\{|\xi(\sigma_0/\gamma)| \geqq \gamma^{-\delta_2}\varepsilon^{-1}\}]. \end{aligned} \tag{4.13}$$

Since $\sigma_0 \leqq \tau^\rho$, Lemma 4.2 implies that, for any $M < \infty$, the last term on the right side of (4.13) is $\leqq \exp(-M/\varepsilon^2)$ for small ρ. By Lemma 4.3, there is a $T_1 < \infty$ such that for all $y \in \Gamma_2$ and small ρ,

$$P_{y,\xi(0)}\{\tau^\rho = \tau_1 \geqq T_1 | |\xi(\sigma_0/\gamma)| \leqq \gamma^{-\delta_2}\varepsilon^{-1}\} \leqq \exp(-M/\varepsilon^2). \tag{4.14}$$

By Theorem 4.1 (with the appropriate choice for the set A) we have, for all $y \in \Gamma_2$ and

sufficiently small ρ, and large enough T,

$$
\begin{aligned}
&P_y\{\tau^\rho = \tau_1 < T_1 \,|\, |\xi(\sigma_0/\gamma)| \leqq \gamma^{-\delta_2}\varepsilon^{-1}\} \\
&\qquad \leqq P_y\{\tau^\rho < T_1 \,|\, |\xi(\sigma_0/\gamma)| \leqq \gamma^{-\delta_2}\varepsilon^{-1}\} \leqq \exp\,(-(S^* - d/2)/\varepsilon^2).
\end{aligned} \tag{4.15}
$$

Combining (4.13) to (4.15) yields, for small ρ,

$$
P_{x,\xi(0)}\{x^\rho(\tau_1) \in \partial G\} \leqq \exp\,(-(S^* - d)/\varepsilon^2). \tag{4.16}
$$

Let v denote the smallest n for which $x^\rho(\tau_n) \in \partial G$. Then by the strong Markov property of $(x^\rho(\cdot),\ \xi^\gamma(\cdot))$

$$
\begin{aligned}
P_{x,\xi(0)}\{v > n\} &= P_{x,\xi(0)}\{x^\rho(\tau_i) \in \Gamma_1,\ i < n\} \\
&= E_{x,\xi(0)} P_{x,\xi(0)}\{x^\rho(\tau_n) \in \Gamma_1 \,|\, \xi(\tau_{n-1}/\gamma),\ x^\rho(\tau_{n-1})\} I_{\{v > n-1\}} \\
&\geqq \inf_{\substack{y \in \Gamma_2 \\ \xi \in B}} P_{y,\xi}\{x^\rho(\tau_1) \in \Gamma_1\} \cdot P_{x,\xi(0)}\{v > n-1\}. \qquad (4.17) \\
&\geqq [1 - \exp -(S^* - d)/\varepsilon^2] P_{x,\xi(0)}\{v > n-1\}. \\
&\geqq [1 - \exp -(S^* - d)/\varepsilon^2]^n.
\end{aligned}
$$

By Lemma 4.3, there is a K_1 such that $E_{x,\xi(0)}(\tau_1 - \sigma_0) \geqq K_1$ for $x \in G - N_\mu(\theta)$. This and (4.17) yields

$$
\begin{aligned}
E_x\tau^\rho &= \sum_1^\infty E_x I_{\{v \geqq n\}}(\tau_n - \tau_{n-1}) \\
&\geqq \sum_1^\infty E_x I_{\{v \geqq n\}}(\tau_n - \sigma_{n-1}) \\
&\geqq \sum_1^\infty P_x\{v \geqq n\} \inf_{\substack{y \in \Gamma \\ \xi(0) \in B}} E_{y,\xi}(\tau_1 - \sigma_0) \qquad (4.18) \\
&\geqq K_1 \exp\,(S^* - d)/\varepsilon^2.
\end{aligned}
$$

Thus, (4.10) holds for $x \in \Gamma_1$; hence for all $x \in N_\mu(\theta)$. The result holds in general, for any $x \in G$, since $P_x\{\tau^\rho > \tau_1\} \to 1$ for any $x \in G$. Q.E.D.

COROLLARY 4.3. *Assume the conditions of Theorem* 4.2 *except we allow part of the boundary* ∂G *to be a trajectory of* $\dot{x} = b(x)$ *(i.e.,* $(b(x),\ n(x)) \leqq 0$ *on* ∂G*) and assume nondegeneracy (i.e., in* (4.2b) $\sigma(x)\Sigma\sigma'(x) \geqq \alpha I$ *in* $\bar{G}$*, where* $\alpha > 0$*). Then* (4.3) *continues to hold.*

Remark on the proof. The proof follows the same lines as that of Theorem 4.2. Although we cannot necessarily find the required T_1 for x near ∂G, to get (4.9), we simply use the fact that

$$
\inf_{\phi, T} S(\phi, T) = S_x \to 0 \quad \text{as } x \to \partial G,
$$

where the inf is over $\{\phi\colon \phi(0) = x,\ \phi(T) \in \partial G\}$, and the fact that for any $d > 0$, $P_{x,\xi(0)}\{\tau^\rho \leqq T\} \geqq \exp -(S_x + d)/\varepsilon^2$ for $|\xi(0)| \leqq \varepsilon^{-1}\gamma^{-\delta}$ (Model I, and analogously for Model II) and small ε. To get (4.10), we use a set G_a, $a < 0$, (to which Theorem 4.2 can be applied), and the fact that the S^* for this converges to the S^* for G, as $a \to 0$.

Many of the other results derived by Freidlin and Ventsel in [1] for the "white Gaussian noise" case carry over to our noise models. We have

THEOREM 4.4. *Assume all the conditions of Theorem* 4.2, *and let there be points*

$y_1, \cdots, y_q \in \partial G$ *such that*

$$\inf_{T>0} \inf_{\phi \in A_i} S(\phi, T) = \inf_{T>0} \inf_{\phi \in A_0} S(\phi, T), \tag{4.19}$$

where $A_i = \{\phi \in C_\theta[0, T]: \phi(T) = y_i\}$ *and* $A_0 = \{\phi \in C_\theta[0, T]: \phi(T) \in \partial G\}$. *Then for each* $x \in G$ *and* $\delta > 0$,

$$\lim_{\rho \to 0} P_x\{d(x^\rho(\tau^\rho), \bigcup_1^q \{y_i\}) < \delta\} \to 1 \tag{4.20}$$

(*i.e., the points of exit of* $x^\rho(\cdot)$ *from* G *converge to* $\bigcup_1^q \{y_i\}$ *as* $\rho \to 0$).

Remark. We omit the proof. The method of proof closely parallels that of [1, Thm. 4.2.1], with modifications of the type used in Theorem 4.2 in order to account for the unboundedness of $\xi(\cdot)$. We only note that, for any $M > 0$ there are $T_\rho \to \infty$ such that $P_{x,\xi(0)}\{\tau^\rho \leqq T_\rho\} \to 1$ and also $|\xi(t)| \leqq \varepsilon^{-1}\gamma^{-\delta}$ (for $\delta \in (0, \frac{1}{2})$ and Model I, with an analogous estimate for Model II) for all $t \leqq T_\rho$, with the probability $\geqq 1 - \exp(-M/\varepsilon^2)$. When $\sigma(x)\Sigma\sigma'(x)$ is degenerate somewhere in $\bar{G}$, then we use the condition (A4.1) as an aid in constructing the "almost" optimal paths which the proof in [1] requires.

Although the details are many and tedious, it can be verified that Theorems 6.6.1 *and* 6.6.2 *of* [1] *hold for our noise models when the system is nondegenerate.* Since the statements of these theorems require the introduction of a great deal of new terminology, the reader is referred to [1]. These theorems give a fairly complete picture of the behavior of $x^\rho(\cdot)$ for large times, when $\dot{x} = b(x)$ has many invariant sets; in particular, they deal with the jump times from one collection of invariant sets to another.

5. The phase locked loop problem. From the point of view of asymptotic methods one of the most interesting systems in communication theory is the phase locked loop (PLL). In [3], the theory of large deviations was applied to estimate the mean time to "loss of track". The model was of the form (1.1)—a small white noise model, and the PLL system was reduced to "basehand", by dropping all terms with high frequency components.

The point of view here is a little more realistic. We use the wide bandwidth noise models I or II and show explicitly that the so-called high frequency terms can be neglected. For notational simplicity, we use the very simplest system, but the analysis and conclusions are applicable to general forms of the PLL and related systems.

Let ω^γ denote the carrier frequency, and suppose that the noise is wideband in an absolute sense but narrow band relative to ω^γ. In particular, we model the noise as follows: let $\xi_i(\cdot)$, $i = 1, 2$, be two mutually independent processes of the form of Model I or II. Let η_γ be such that $\eta_\gamma/\gamma \to^\gamma 0$. A standard way of describing wide bandwidth noise $n^\gamma(\cdot)$ of the desired type is via the formula (σ is a constant matrix here)

$$\begin{aligned} n^\gamma(t) &= \varepsilon u^\gamma(t), \\ u^\gamma(t) &= \sigma[\xi_1^\gamma(t) \sin(\omega_0 t/\eta_\gamma) + \xi_2^\gamma(t) \cos(\omega_0 t/\eta_\gamma)]. \end{aligned} \tag{5.1}$$

The bandwidth is $O(1/\gamma)$ and the "center" and "carrier" frequencies are $\omega_0/\eta_\gamma \equiv \omega^\gamma$. See Fig. 5.1 for a description of our system. The dotted box could contain a filter, but we omit it for simplicity until later. The VCO (voltage controlled oscillator) is an oscillator whose frequency deviation from a central (carrier) frequency is proportional to its input. The device is an essential component of many modern communications systems, and much effort has gone into its analysis [10], [11]. The purpose of the PLL is to provide an estimate $\hat{\theta}(t)$ of the unknown input phase θ. The statistics of escape

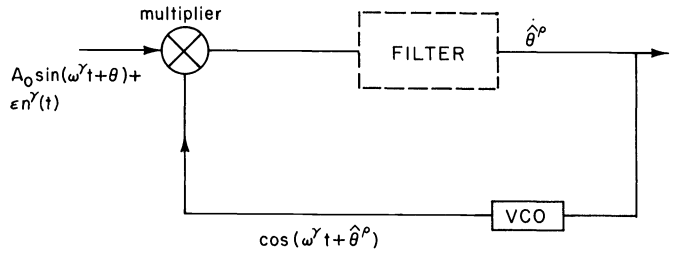


FIG. 5.1. *A phase locked loop.*

of the estimate from a neighborhood of θ are important in understanding the tracking ability of the system.

We first study the PLL system

$$\begin{aligned} &\dot{\hat{\theta}}^\rho = u^\rho, \\ &u^\rho = \cos(\omega^\gamma t+\hat{\theta}^\rho)[A_0 \sin(\omega^\gamma t+\theta)+\varepsilon n^\gamma(t)]. \end{aligned} \tag{5.2}$$

Thus

$$\begin{aligned} \dot{\hat{\theta}} = &\frac{A_0}{2}\sin(\theta-\hat{\theta}^\rho)+\frac{A_0}{2}\sin(\theta+\hat{\theta}^\rho+2\omega^\gamma t) \\ &+\frac{\varepsilon}{2}\sigma[\xi_1^\gamma(t)\cos\hat{\theta}^\rho-\xi_2^\gamma(t)\sin\hat{\theta}^\rho] \\ &+\frac{\varepsilon}{2}\sigma[\xi_1^\gamma(t)\cos(2\omega^\gamma t+\hat{\theta}^\rho)+\xi_2^\gamma(t)\sin(2\omega^\gamma t+\hat{\theta}^\rho)]. \end{aligned} \tag{5.3}$$

By a direct calculation using the definition of the H-functional, and the fact that $\cos(2\omega^\gamma t)$ and $\sin(2\omega^\gamma t)$ are scaled "faster" than θ_i^γ is, we can see that the H-functional for (5.3) is the same as for the system

$$\dot{x}^\rho = \frac{A_0}{2}\sin(\theta-x^\rho)+\frac{\varepsilon}{2}\sigma[\xi_1^\gamma(t)\cos x^\rho-\xi_2^\gamma(t)\sin x^\rho]. \tag{5.4}$$

The H-functional for (5.4) is of the "Gaussian white noise" form

$$H(x,\alpha) = \alpha\frac{A_0}{2}\sin(\theta-x)+\frac{\sigma\Sigma\sigma'\alpha^2}{8}, \tag{5.5}$$

where $\Sigma = \Sigma_1$ or Σ_2 depending on the noise model, and the normalizing sequence is $\lambda_\rho = \varepsilon^2$.

In fact, according to Appendix III, the large deviations formulas (4.1), (4.2) for (5.3) are the same as those for the model (5.4). The effects of these high frequency terms all disappear in the limit, from the point of view of the theory of large deviations, and these terms can be ignored in all the calculations.

If a filter is incorporated into the forward branch of the PLL, then the system (5.3) is replaced by

$$\dot{\nu}^\rho = A_1\nu^\rho + B_1 u^\rho, \qquad \dot{\hat{\theta}}^\rho = H_1\nu^\rho, \tag{5.6}$$

where A_1 is stable, and $H_1\nu^\rho$ is the filter output. The analysis is the same. Arguments such as that of Appendix III show that the "high frequency" terms can be dropped irrespective of the filtering action.

We next study the mean time required until $|\theta^\rho(t)|$ first reaches the critical level π, for a simple PLL with a first order filter.

Mean escape time for a PLL *with a simple filter.* As noted in Corollary 4.3, it is possible to generalize the conditions on the domain G in Theorem 4.2 if certain conditions hold near the boundary ∂G. The PLL treated below will be such a case. For specificity, let us use a first order filter in Fig. 5.1, although analogous results will hold if any order stable filter is used. Denote the filter by

$$\dot{v}_1 = -av_1 + b \cdot (\text{filter input})$$

$$\text{filter output} = cv_1, \quad a > 0, \quad b > 0, \quad c > 0.$$

Write $x = (x_1, x_2) = (v_1, \hat{\theta})$. For simplicity, let $\theta(t) \equiv 0$. This is not a restriction, since it is only the errors in the estimate which are important. Then

$$\begin{aligned} \dot{x}_1^\rho &= -ax_1^\rho + b \sin(-x_2^\rho) + \varepsilon n^\rho \\ \dot{x}_2^\rho &= cx_1^\rho \end{aligned} \qquad \equiv f(x^\rho) + \begin{pmatrix} \varepsilon n^\rho \\ 0 \end{pmatrix}. \tag{5.7}$$

The system is degenerate, but satisfies (A4.1), since the linearized system

$$\dot{x} = Ax - Bu, \quad A = \begin{bmatrix} -a & -b \\ c & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

is controllable.

In "PLL" parlance, a "cycle slip" refers to $\hat{\theta}(\cdot) = x_2(\cdot)$ reaching the level $\pm\pi$, or a movement away from the origin into the domain of attraction of another stable point. Refer to Fig. 5.2. The natural region of interest is G. A slip can occur if $|x_2(t)|$ exceeds π—or if $x(t)$ exists G through the curves C_1 or C_2, which are trajectories of the unperturbed system $\dot{x} = f(x)$. In the latter case, they will move quickly toward the stable points $(0, 2\pi)$ or $(0, -2\pi)$, at a "rapid" change in phase (often causing a noticeable effect on the behavior of the communications system with which the PLL is used). So, we are concerned with exit from G.

Near C_1 and C_2, the trajectories of $\dot{x} = f(x)$ are parallel to these curves, and the curve through x takes longer and longer to reach a neighborhood of the origin as $x \to C_1$ or C_2. Thus Theorem 4.2 cannot be used directly. But a slight modification of the proof yields (4.3). The method combines the ideas of Theorem 4.2 with the remarks after Corollary 4.3. We note the following facts. We restrict our attention to the point $(0, \pi)$ of C_1, although the identical remarks can be made about C_2 and $(0, -\pi)$.

(1) The system $\dot{x} = f(x) + Bu$ is controllable (in the sense of Assumption A4.1) in a neighborhood of $(0, \pi)$. To see this, we need only check controllability (in the usual sense for linear systems; see, e.g., [2]) for the linearized (about $(0, \pi)$) system

$$\dot{y} = Ay + Bu, \quad A = \begin{bmatrix} -a & b \\ c & 0 \end{bmatrix}, \quad B = \begin{bmatrix} b \\ 0 \end{bmatrix}.$$

(2) For any $\alpha > 0$ there is a neighborhood N^α of $(0, \pi)$ and positive δ such that if $x \in N^\alpha \cap G$, then there is a $u_x(\cdot)$ transferring x to $\partial N_\delta(G)$ at a time T_x with $\int_0^{T_x} |u_x(t)|^2 \, dt \leqq \alpha$.

(3) For α as in (2), there is a $\beta > 0$ and a neighborhood N^β of C_1 such that for $x \in N^\beta \cap G$,

$$\inf_T \inf_{\phi \in A_\beta} S(\phi, T) \leqq \alpha,$$

where $A_\beta = \{\phi(\cdot)\colon \phi(0) \in N^\beta, \ \phi(T) \in \partial N_\delta(G)\}$.

(4) For $x(0) \notin N^{\beta} \cap G$, and $u > 0$, there is a T_1^{β} such that the trajectory of $\dot{x} = f(x)$ reaches the u-neighborhood of the origin in time $\leqq T_1^{\beta}/2$.

By using these facts in the proof of Theorem 4.2 to treat points near to C_1 or C_3 (see also the remark after Corollary 4.3), we obtain (4.3).

Remark on "batches" of cycle slips. It is noted in applications that cycle slips (or what are taken to be cycle slips) often occur in "batches". Perhaps some insight into this *one possible cause* can be obtained from Fig. 5.2. If due to large noise, the path is pushed far out but between the upper and lower trajectories, it will quickly return to a neighborhood of a stable point, perhaps in an oscillatory manner, when the noise level drops, giving the impression that many cycle slips are occurring in a very short period of time. Similarly, if it returns to a stable point after reaching a neighborhood of the separatrices, the oscillations will give the appearance of a "batch" of cycle slips. An explanation of a possible alternative cause involving the crossing of several "π-levels" in a short time is given in [14, pp. 281–284].

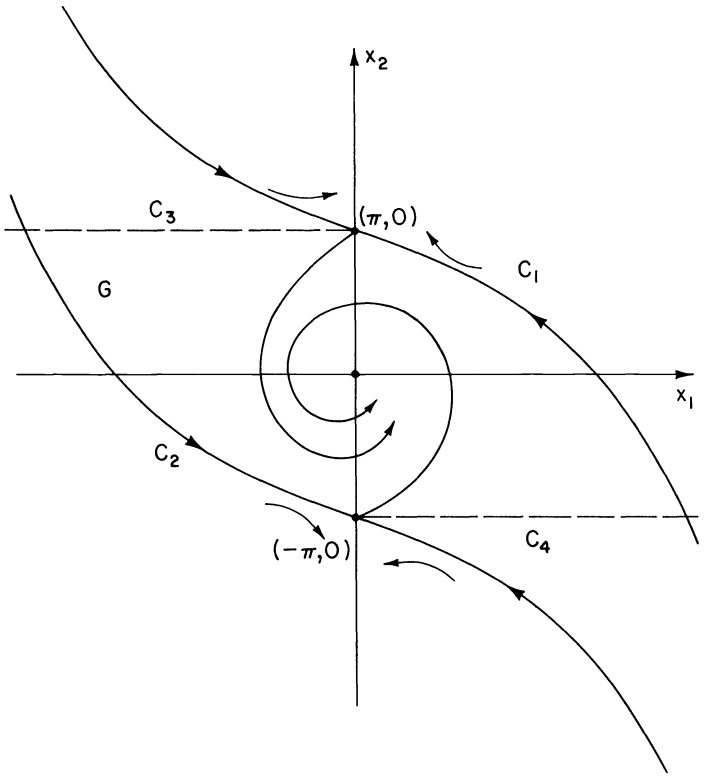


FIG. 5.2. *Trajectories for the noiseless phase locked loop.*

Appendix I. Proof of large deviations theorems. Here we state and prove the basic large deviations inequalities (A1.3), (A1.4) and indicate how these are used to prove (4.1). In order to satisfy the hypotheses of Theorem 4.2, it will be necessary to obtain our estimates uniformly in $x(0) \in G$, and $|\xi(0)| \leqq \gamma^{-\delta}\varepsilon^{-1}$ (Model I). The analogous uniformity that is required for Model II is for $|\bar{\xi}^{\gamma}(0)| \leqq \varepsilon^{-1}\gamma^{-\delta+1/2}$. Note that the uniformity for Model I is expressed in terms of $\xi(\cdot)$, while that of Model II is in terms of $\bar{\xi}^{\gamma}(\cdot)$. The reason for this difference is the different scaling required to get weak convergence to a Wiener process in the two cases. The main Theorem A1.1 is first stated. But before giving the proof, it is convenient to first obtain large deviations

results for the system,

$$y^\rho(t) = \varepsilon \int_0^t \xi^\gamma(s)\, ds + x, \tag{A1.1}$$

and this will be done in Lemmas A1.1 and A1.2. Following this, we modify the method which Varadhan [8] used to extend large deviations results for the process $\varepsilon w(\cdot)$ to system (1.1), in order to extend our results for (A1.1) to the system (1.6.).

THEOREM A1.1. *Consider system* (1.6), *the associated H, L, and S functionals, and the normalizing sequence* $\lambda_\rho = \varepsilon^2$. *Let A be a set in* $C_x[0, T]$. *Then for* $\frac{1}{2} > \delta > 0$, *the limits* (A1.2) *hold uniformly in* $|\xi(0)| \leqq \gamma^{-\delta}\varepsilon^{-1}$ (*noise Model* I) *and* $|\bar{\xi}^\gamma(0)| \leqq \gamma^{-\delta+1/2}\varepsilon^{-1}$ (*noise Model* II), *and for x in any compact set.*

$$\underline{\lim_\rho}\, \lambda_\rho \log P\{x^\rho(\cdot) \in A\} \geqq -\inf_{\phi \in A^0} S(\phi, T), \tag{A1.2a}$$

$$\overline{\lim_\rho}\, \lambda_\rho \log P\{x^\rho(\cdot) \in A\} \leqq -\inf_{\phi \in \bar{A}} S(\phi, T). \tag{A1.2b}$$

Definitions. By § 3, the *H*, *L* and *S* functionals for the process $y^\rho(\cdot)$ are defined by

$$H_y(\alpha) = \alpha'\Lambda\alpha/2, \tag{A1.3}$$

$$L_y(\beta) = \beta'\Lambda^{-1}\beta/2, \tag{A1.4}$$

$$S_y(\phi, T) = \int_0^T \dot{\phi}(s)'\Lambda^{-1}\dot{\phi}(s)\, ds/2. \tag{A1.5}$$

(Here, as elsewhere, $S_y(\phi, T) = \infty$ if ϕ is not absolutely continuous). For Model I, $\Lambda = \Sigma_1$, and for Model II, $\Lambda = \Sigma_2$ (see § 2 for the definition), and we can assume (w.l.o.g.) that Σ_i are nonsingular. Define $\Phi_s^y = \{\phi \in C_x[0, T]\colon S_y(\phi, T) \leqq s\}$. The set Φ_s is compact [7].

Before proving the main theorem, it is convenient to prove the following auxiliary lemmas.

LEMMA A1.1. *Let* $y^\rho(\cdot)$ *be given by* (A1.1). *Then given* $c > 0$, $h > 0$, $s > 0$, $\frac{1}{2} > \delta > 0$, *and* $\phi \in C_x[0, T]$, *there is* $\rho_0 > 0$ *such that for* $\rho < \rho_0$ *and* $x \in G$, $|\xi(0)| \leqq \gamma^{-\delta}\varepsilon^{-1}$ (*Model* I), *or* $|\bar{\xi}^\gamma(0)| \leqq \varepsilon^{-1}\gamma^{-\delta+1/2}$ (*Model* II),

$$P_{x,\xi(0)}\{d(y^\rho, \phi) < c\} \geqq \exp\left(-[S_y(\phi, T) + h]/\lambda_\rho\right), \tag{A1.6}$$

$$P_{x,\xi(0)}\{d(y^\rho, \Phi_s^y) > c\} \leqq \exp\left(-[s - h]/\lambda_\rho\right). \tag{A1.7}$$

Proof. First we prove the result for *Model* I. Via a change of variables $s \to s/\gamma$ and using $\xi(s)\, ds = A^{-1}A\xi(s)\, ds = A^{-1}\, d\xi(s) - A^{-1}B\, dw(s)$, we have

$$\begin{aligned} y^\rho(t) &= \varepsilon \int_0^t \xi(s/\gamma)\, ds/\sqrt{\gamma} + x \\ &= \sqrt{\gamma}\varepsilon A^{-1}(\xi(t/\gamma) - \xi(0)) - \sqrt{\gamma}\varepsilon A^{-1}B(w(t/\gamma) - w(0)) + x. \end{aligned}$$

By Lemma A2.1, for given $c > 0$, $M < \infty$, and uniformly in $|\xi(0)| \leqq \gamma^{-\delta}\varepsilon^{-1}$, for small ρ we have

$$P\{\sup_{0 \leqq t \leqq T} |\sqrt{\gamma}\varepsilon A^{-1}(\xi(t/\gamma) - \xi(0))| \geqq c\} \leqq \exp\left(-M/\varepsilon^2\right).$$

Hence the large deviations properties of (A1.1) are the same as those of

$$-\varepsilon A^{-1}B\sqrt{\gamma}[w(t/\gamma) - w(0)] + x.$$

As $-\sqrt{\gamma}[w(t/\gamma)-w(0)]=\tilde{w}(t)$ is a *standard Brownian motion*, the inequalities (A1.6), (A1.7) follow from the fact that Theorem A1.1 holds if the $x^\rho(\cdot)$ there is replaced by (define $A^{-1}B=C$)

$$\tilde{y}^\rho(t)=\varepsilon C\tilde{w}(t)+x$$

and the action functional

$$S_y(\phi, T)=\int_0^T \dot{\phi}(s)'(CC')^{-1}\dot{\phi}(s)\,ds/2,$$

and normalizing sequence $\lambda_\rho=\varepsilon^2$ are used [1], [8].

We now consider Model II. We have

$$\begin{aligned} y^\rho(t) &= \varepsilon\int_0^t \xi^\gamma(s)\,ds+x \\ &= \varepsilon\int_0^{t/\gamma} \bar{\xi}^\gamma(s)\,ds+x \\ &= \varepsilon\int_0^{t/\gamma} A^{-1}[d\bar{\xi}^\gamma(s)-d\bar{J}^\gamma(s)]+x \\ &= \varepsilon A^{-1}(\bar{\xi}^\gamma(t/\gamma)-\bar{\xi}^\gamma(0))-\varepsilon A^{-1}(\bar{J}^\gamma(t/\gamma)-\bar{J}^\gamma(0))+x. \end{aligned} \tag{A1.8}$$

As for Model I, we may ignore the first term on the right (Lemma A4.1). Several standard "continuity" methods can be used to establish the statement of Lemma A1.1 for the system

$$\tilde{y}^\rho(t)=-\varepsilon A^{-1}(\bar{J}^\gamma(t/\gamma)-\bar{J}^\gamma(0))+x \tag{A1.9}$$

with action functional

$$S_y(\phi, T)=\int_0^T \dot{\phi}(s)'\Sigma_2^{-1}\dot{\phi}(s)\,ds/2$$

and normalizing sequence $\lambda_\rho=\varepsilon^2$. (For example, one can use the technique in [7, Thm. 2.1], based on Gartner's theorem for a sampled system.) Q.E.D.

From Lemma A1.1, by a standard argument in large deviations literature (see, e.g., [7], or [9, Prop. 7.6]) we obtain the following:

LEMMA A1.2. *Let* $A\subset C_x[0, T]$, $\frac{1}{2}>\delta>0$. *Then the limits in* (A1.10) *hold, uniformly for x in any compact set, and for* $|\xi(0)|\leqq\gamma^{-\delta}\varepsilon^{-1}$ (*Model* I) *or* $|\bar{\xi}^\gamma(0)|\leqq\gamma^{-\delta+1/2}\varepsilon^{-1}$ (*Model* II).

$$\begin{aligned} -\inf_{\phi\in A^0} S_y(\phi, T) &\leqq \underline{\lim}_\rho \lambda_\rho \log P_{x,\xi(0)}\{y^\rho(\cdot)\in A\} \\ &\leqq \overline{\lim}_\rho \lambda_\rho \log P_{x,\xi(0)}\{y^\rho(\cdot)\in A\} \leqq -\inf_{\phi\in\bar{A}} S_y(\phi, T). \end{aligned} \tag{A1.10}$$

The proof of Theorem A1.1 uses a device of Varadhan [8]. To facilitate this, it is convenient to next prove a form of Lemma A1.2 or, Theorem A1.1 for a special auxiliary system—driven by $\xi^\gamma(\cdot)$. Let $\alpha>0$, and define $\pi_\alpha(t)=\alpha[t/\alpha]$, where $[t]$ is the largest integer less than or equal to t. Define the system

$$\begin{aligned} \dot{x}_\alpha^\rho(t) &= b(x_\alpha^\rho(\pi_\alpha(t)))+\varepsilon\sigma(x_\alpha^\rho(\pi_\alpha(t)))\xi^\gamma(t), \\ x_\alpha^\rho(0) &= x. \end{aligned} \tag{A1.11}$$

Thus the arguments of $b(\cdot)$ or $\sigma(\cdot)$ are constant on intervals of length α. For each $\alpha > 0$, there exists a continuous map $F_{\alpha,x}$ from $C_x[0, T]$ to $C_x[0, T]$ such that

$$x_\alpha^\rho(\cdot) = F_{\alpha,x}\left(\varepsilon \int_0^\cdot \xi^\gamma(s)\, ds\right). \tag{A1.12}$$

The $F_{\alpha,x}$ is continuous uniformly in x in any compact set, and in $|\xi(0)| \leqq \varepsilon^{-1}\gamma^{-\delta}$ for noise Model I and $|\bar{\xi}^\gamma(0)| \leqq \varepsilon^{-1}\gamma^{-\delta+1/2}$ for noise Model II. Large deviations results may be transferred from one system to another if they are connected by a continuous map. Let $S_\alpha(\phi, T)$ denote the action functional for $x_\alpha^\rho(\cdot)$. Then ([8, Remark 1, on p. 5], [1, Thm. 3.3.1])

$$S_\alpha(\phi, T) = \inf_{F_{\alpha,x}(f)=\phi} S_y(f, T).$$

If $\sigma(x)\Lambda\sigma'(x)$ is uniformly positive definite (where $\Lambda = \Sigma_1$ or Σ_2 as appropriate), we can evaluate S_α explicitly, since $F_{\alpha,x}$ is 1:1 and we get

$$S_\alpha(\phi, T) = \frac{1}{2}\int_0^T [\dot{\phi}(s) - b(\phi(\pi_\alpha(s)))]'[\sigma(\phi(\pi_\alpha(s)))\Lambda\sigma'(\phi(\pi_\alpha(s)))]^{-1}$$
$$\cdot [\dot{\phi}(s) - b(\phi(\pi_\alpha(s)))]\, ds.$$

In general, the integrand is $L_\alpha(s)$, where

$$L_\alpha(s) = \inf_\alpha [\alpha'(\dot{\phi}(s) - b(\phi(\pi_\alpha(s)))) - \alpha'\sigma(\phi(\pi_\alpha(s)))\Lambda\sigma'(\phi(\pi_\alpha(s)))\alpha/2].$$

The normalizing sequence is still $\lambda_\rho = \varepsilon^2$.

Thus, the large deviations result, Theorem A1.1 holds for system (A1.11), uniformly in the desired initial conditions of $(x, \xi^\gamma(0))$, with the use of S_α. In order to extend the result to (1.6), we need to bound $|x_\alpha^\rho(t) - x^\rho(t)|$, and this is done in the next lemma.

Let $\phi_\alpha \to \phi$ in $C_x[0, T]$. Then the "lower semi-continuity"

$$\underline{\lim}_\alpha S_\alpha(\phi_\alpha, T) \geqq S(\phi, T)$$

can be proved, but we omit the details.

LEMMA A1.3. *For given $c > 0$, $M < \infty$, $T < \infty$, there is an $\alpha_0 > 0$ such that for each $\alpha < \alpha_0$ there is a $\rho_\alpha > 0$ such that for $\rho \leqq \rho_\alpha$ and all $x \in G$ and $|\xi(0)| \leqq \gamma^{-\delta}\varepsilon^{-1}$ (Model I) and $|\bar{\xi}^\gamma(0)| \leqq \varepsilon^{-1}\gamma^{-\delta+1/2}$ (Model II)*

$$P\{\sup_{0 \leqq t \leqq T} |x^\rho(t) - x_\alpha^\rho(t)| > c\} \leqq \exp(-M/\lambda_\rho). \tag{A1.13}$$

Proof. For simplicity, we only prove the lemma for noise Model I. The details in the other case are essentially the same. By writing,

$$x_\alpha^\rho(t) = x + \int_0^t b(x_\alpha^\rho(s))\, ds + \varepsilon \int_0^t \sigma(x_\alpha^\rho(s))\xi^\gamma(s)\, ds + \beta_1^\rho(t),$$

where

$$\beta_1^\rho(t) = \int_0^t [b(x_\alpha^\rho(\pi_\alpha(s))) - b(x_\alpha^\rho(s))]\, ds$$
$$+ \varepsilon \int_0^t [\sigma(x_\alpha^\rho(\pi_\alpha(s)) - \sigma(x_\alpha^\rho(s))]\xi^\gamma(s)\, ds,$$

we have

$$\begin{aligned}\Delta_\alpha^\rho(t) &\equiv x^\rho(t) - x_\alpha^\rho(t) \\ &= \int_0^t [b(x^\rho(s)) - b(x_\alpha^\rho(s))]\, ds \\ &\quad + \varepsilon \int_0^t [\sigma(x^\rho(s)) - \sigma(x_\alpha^\rho(s))]\xi^\gamma(s)\, ds - \beta_1^\rho(t).\end{aligned} \tag{A1.14}$$

We next prove the assertion: *For $c_1 > 0$, $M < \infty$, $T < \infty$, and small enough α, there is a $\rho_0 > 0$ such that for $\rho \leqq \rho_0$ and $x \in G$ and $|\xi(0)| \leqq \gamma^{-\delta}\varepsilon^{-1}$ (Model* I)

$$P\{\sup_{0 \leqq t \leqq T} |\beta_1^\rho(t)| > c_1\} \leqq \exp(-M/\lambda_\rho).$$

In order to prove the assertion, first note that for small enough α and for $s \in [i\alpha, i\alpha + \alpha)$, we may assume that $|x_\alpha^\rho(s) - x_\alpha^\rho(i\alpha)|$ is as small as desired (uniformly in the appropriate initial conditions), save on a set of measure $\leqq \exp(-2M/\varepsilon^2)$. This follows from the decomposition (see the proof of Lemma A1.1)

$$\varepsilon \int_{i\alpha}^{s} \sigma(x_\alpha^\rho(\pi_\alpha(t)))\xi^\gamma(t)\, dt = \sqrt{\gamma}\varepsilon\sigma(x_\alpha^\rho(\pi_\alpha(\gamma t)))A^{-1}[\xi(t) - Bw(t)]|_{i\alpha/\gamma}^{s/\gamma},$$

Lemma A2.5, and the boundedness of $b(\cdot)$ and $\sigma(\cdot)$. In fact, by Lemma A2.5 and the above decomposition ($v < 1/2$)

$$P\{\sup_{i \leqq T/\alpha} \sup_{t \leqq \alpha} |x_\alpha^\rho(s) - x_\alpha^\rho(i\alpha)| > \alpha^v\} \leqq \exp(-2M/\varepsilon^2)$$

for small Δ and ρ. We thus assume that $|x_\alpha^\rho(s) - x_\alpha^\rho(i\alpha)| \leqq \alpha^v$, for $s \in [i\alpha, i\alpha + \alpha)$.

Owing to the Lipschitz property of $b(\cdot)$ and $\sigma(\cdot)$, we may now assume that the terms $[b(x_\alpha^\rho(\pi_\alpha(s))) - b(x_\alpha^\rho(s))]$, $[\sigma(x_\alpha^\rho(\pi_\alpha(s))) - \sigma(x_\alpha^\rho(s))]$ entering into the definition of $\beta_1^\rho(\cdot)$ are as small as desired save on a set of measure $\leqq \exp(-M/\lambda_\rho)$ for arbitrary M. The first consequence of this fact is that in proving the *assertion*, we may *ignore* the first integral in the definition of $\beta_1^\rho(t)$. After a change of scale, the use of $A\xi(s)\, ds = d\xi(s) - B\, dW(s)$, and an integration by parts (for the term involving $d\xi$), the second term in $\beta_1^\rho(t)$ becomes a sum of the following terms:

$$\begin{aligned}&\sqrt{\gamma}\varepsilon \int_0^{t/\gamma} [\sigma(x_\alpha^\rho(\pi_\alpha(\gamma s))) - \sigma(x_\alpha^\rho(\gamma s))]A^{-1}B\, dW_s, \\ &\sqrt{\gamma}\varepsilon[\sigma(x_\alpha^\rho(\pi_\alpha(\gamma s))) - \sigma(x_\alpha^\rho(\gamma s))]A^{-1}\xi(s)|_0^{t/\gamma}, \\ &\gamma^{3/2}\varepsilon \int_0^{t/\gamma} b(x_\alpha^\rho(\gamma s))'\sigma_x(x_\alpha^\rho(\gamma s))A^{-1}\xi(s)\, ds, \\ &\gamma\varepsilon^2 \int_0^{t/\gamma} \xi(s)'\sigma'(x_\alpha^\rho(\gamma s))\sigma_x(x_\alpha^\rho(\gamma s))A^{-1}\xi(s)\, ds, \\ &\sqrt{\gamma}\varepsilon \sum_{i=0}^{t/\alpha} [\sigma(x_\alpha^\rho(i\alpha + \alpha)) - \sigma(x_\alpha^\rho(i\alpha))]A^{-1}\xi((i\alpha + \alpha)/\gamma).\end{aligned} \tag{A1.15}$$

The last term arises due to the discontinuity in $\sigma(x_\alpha^\rho(\pi_\alpha(\gamma s)))$ at $\gamma s = i\alpha$ for each i. These terms are taken care of, with the required uniformity, by Lemmas A2.2, A2.1, A2.4, A2.3, and A2.5 (owing to which we can assume that the coefficients in the last sum are less than $K\alpha^v$ for $v \in (0, \frac{1}{2})$, and obtain the appropriate estimate for the sum). This finishes the proof of the *assertion.*

Now, return to the proof of Lemma A1.3 and consider the second term in the definition of $\Delta_\alpha^\rho(\cdot)$ in (A1.14). By a decomposition analogous to the one above, we may write that term as

$$\varepsilon\int_0^t [\sigma(x^\rho(s))-\sigma(x_\alpha^\rho(s))]\xi^\gamma(\alpha)\,ds$$
$$=-\sqrt{\gamma}\varepsilon\int_0^{t/\gamma}[\sigma(x^\rho(\gamma s))-\sigma(x_\alpha^\rho(\gamma s))]A^{-1}B\,dw(s)+\beta_2^\rho(t)$$

where $\beta_2^\rho(t)$ satisfies the same conditions as $\beta_1^\rho(t)$ does. (It can be represented as a sum of the same types of terms (A1.15).) Thus we can assume that $|\beta_2^\rho(t)|\leqq$ any c_2 w.p.$\geqq 1-\exp(-2M/\varepsilon^2)$. There exists a standard Wiener process $\tilde{w}(t)$ such that $x^\rho(\cdot)$ and $x_\alpha^\rho(\cdot)$ are nonanticipative with respect to $\tilde{w}(\cdot)$ and

$$-\sqrt{\gamma}\varepsilon\int_0^{t/\gamma}[\sigma(x^\rho(\gamma s))-\sigma(x_\alpha^\rho(\gamma s))]A^{-1}B\,dw(s)$$
$$=\varepsilon\int_0^t[\sigma(x^\rho(s))-\sigma(x_\alpha^\rho(s))]A^{-1}B\,d\tilde{w}(s).$$

Define $\beta_3^\rho(t)=\beta_2^\rho(t)-\beta_1^\rho(t)$, $\tilde{\Delta}_\alpha^\rho(t)=\Delta_\alpha^\rho(t)-\beta_3^\rho(t)$, and the terms

$$b^*(s)=\int_0^1 b_x(x_\alpha^\rho(s)+\tau\Delta_\alpha^\rho(s))\,d\tau,$$
$$\sigma^*(s)=\int_0^1\sigma_x(x_\alpha^\rho(s)+\tau\Delta_\alpha^\rho(s)\,d\tau.$$

Then, by Taylor's formula with remainder

$$\Delta_\alpha^\rho(t)=\int_0^t b^*(s)\Delta_\alpha^\rho(s)\,ds+\varepsilon\int_0^t\sigma^*(s)\Delta_\alpha^\rho(s)A^{-1}B\,d\tilde{w}(s)+\beta_3^\rho(t),$$

so we can write

$$\tilde{\Delta}_\alpha^\rho(t)=\int_0^t b^*(s)\tilde{\Delta}_\alpha^\rho(s)\,ds+\varepsilon\int_0^t\sigma^*(s)\tilde{\Delta}_\alpha^\rho(s)A^{-1}B\,d\tilde{w}(s)$$
$$+\int_0^t b^*(s)\beta_3^\rho(s)\,ds+\varepsilon\int_0^t\sigma^*(s)\beta_3^\rho(s)A^{-1}B\,d\tilde{w}(s). \tag{A1.16}$$

By what we have shown above, given any $\eta>0$, for small enough α and ρ we may assume that $|\beta_3^\rho(t)|<\eta$ (w.p.$\geqq 1-\exp(-2M/\varepsilon^2)$).

In order to estimate the solution to (A1.16), we now introduce a technique of Varadhan [8, § 6] and develop an argument paralleling his. Define $\tau=\inf\{s: |\tilde{\Delta}_\sigma^\rho(s)|\geqq c\}\wedge T$. Consider the function

$$g(z)=(\theta+|z|^2)^l$$

where θ and l will be chosen later on. By Itô's formula, and the boundedness of $b^*(\cdot)$, $\sigma^*(\cdot)$, there are constants k_i such that

$$dg(\tilde{\Delta}_\alpha^\rho(t))=\alpha_1(t)\,dt+\alpha_2(t)\,d\tilde{w}(t) \tag{A1.17}$$

where

$$\begin{aligned}\alpha_1(t) &\leqq k_1 l(\theta+|\tilde{\Delta}^\rho_\alpha(t)|^2)^{l-1}(\eta+|\tilde{\Delta}^\rho_\alpha(t)|)|\tilde{\Delta}^\rho_\alpha(t)| \\ &\quad + k_2\varepsilon^2 l(l-1)(\theta+|\tilde{\Delta}^\rho_\alpha(t)|^2)^{l-2}(\eta^2+|\tilde{\Delta}^\rho_\alpha(t)|^2)|\tilde{\Delta}^\rho_\alpha(t)|^2 \\ &\quad + k_3\varepsilon^2 l(\theta+|\tilde{\Delta}^\rho_\alpha(t)|^2)^{l-1}(\eta^2+|\tilde{\Delta}^\rho_\alpha(t)|^2) \\ &\leqq k_4 l(\theta+|\tilde{\Delta}^\rho_\alpha(t)|^2)^{l-1}(\eta^2+|\tilde{\Delta}^\rho_\alpha(t)|^2) \\ &\quad + k_5\varepsilon^2 l^2(\theta+|\tilde{\Delta}^\rho_\alpha(t)|^2)^{l-2}(\eta^2+|\tilde{\Delta}^\rho_\alpha(t)|^2)^2.\end{aligned}$$

Let $\theta=\eta^2$. Then for some constant k,

$$\alpha_1(t)\leqq kl(1+\varepsilon^2 l)g(\tilde{\Delta}^\rho_\alpha(t)).$$

Let $\lambda=cl(1+\varepsilon^2 l)$. Then

$$e^{-\lambda(\tau\wedge t)}g(\tilde{\Delta}^\rho_\alpha(\tau\wedge t))$$

is a nonnegative supermartingale. Let l/ε^2 replace l. By the supermartingale property,

$$E[e^{-\lambda\tau}(\eta^2+|\tilde{\Delta}^\rho_\alpha(\tau)|^2)^{l/\varepsilon^2}]\leqq\eta^{2l/\varepsilon^2}.$$

Hence

$$P\{\tau<T\}\leqq(\exp cl(l+1)T/\varepsilon^2)(\eta^2/(\eta^2+c^2))^{l/\varepsilon^2}.$$

Choose $l=1$ and η small enough to obtain

$$P\{\tau<T\}\leqq\exp(-2M/\varepsilon^2).$$

From this and $|\beta^\rho_3(t)|\leqq\eta$, we get

$$P\{\sup_{0\leqq t\leqq T}|\Delta^\rho_\alpha(t)|\geqq c+\eta\}\leqq\exp(-M/\varepsilon^2)$$

for small α and ρ. This proves Lemma A1.3. Q.E.D.

We are now ready for the

Proof of Theorem A1.1. Recall that $A\subset C_x[0,T]$, and let A have a nonempty interior. Define $A^c=$ complement of A and define the sets

$$A^a=\begin{cases}\{\phi: d(\phi,A^c)>-a\}, & a<0,\\ A, & a=0,\\ \{\phi: d(\phi,A)<a\}, & a>0.\end{cases}$$

First we show (A1.2a). Recall the definition of $S_\alpha(\phi,T)$ given below (A1.12). Let $h>0$, and choose $a<0$ so that there is $\phi_a\in A^a$ such that

$$S(\phi_a,T)\leqq\inf_{\phi\in A^0}S(\phi,T)+h/4. \tag{A1.18}$$

Since $S_\alpha(\phi,T)\to S(\phi,T)$ for each ϕ as $\alpha\to0$, we may assume that α is small enough so that

$$S_\alpha(\phi_a,T)\leqq S(\phi_a,T)+h/4. \tag{A1.19}$$

From Lemma A1.2, by taking α smaller if necessary we may assume that for fixed $M<\infty$,

$$P_x\{\sup_{p\leqq t\leqq T}|x^\rho(t)-x^\rho_\alpha(t)|>a\}\leqq\exp-M/\lambda_\rho.$$

By picking M large enough, for small ρ, we have

$$P_x\{x^\rho(\cdot)\in A\}\geqq P_x\{x^\rho_\alpha(\cdot)\in A^a\}-\exp(-M/\lambda_\rho)$$
$$\geqq \exp(-[\inf_{\phi\in A^0} S(\phi, T)+3h/4])-\exp(-M/\lambda_\rho)$$
$$\geqq \exp(-[\inf_{\phi\in A^0} S(\phi, T)+h]).$$

The second inequality above is due to the fact that Theorem A1.1 holds for the $x^\rho_\alpha(\cdot)$ process (with action functional S_α), and (A1.18), (A1.19). Since $h>0$ is arbitrary, the string of inequalities proves (A1.2a).

Before proving (A1.2b), it is convenient to prove

LEMMA A1.4. *Let $F\subset C_x[0, T]$ be closed. Then*

(i) $$\varliminf_{\alpha\to 0}\inf_{\phi\in F} S_\alpha(\phi, T)\geqq \inf_{\phi\in F} S(\phi, T).$$

(ii) $$\varliminf_{a\to 0^+}\inf_{\phi\in \bar F^a} S(\phi, T)=\inf_{\phi\in F} S(\phi, T).$$

Proof. (ii) is a consequence of the lower semicontinuity of $S(\cdot, T)$ and the compactness of each set $\{\phi: S(\phi, T)\leqq s\}$. We only show (i) for the case

$$\inf_{\phi\in F} S(\phi, T)<\infty.$$

Assume (i) is not true. Then there are ϕ_n, α_n and $h>0$ so that for all $n=1, 2, \cdots$

$$S_{\alpha_n}(\phi_n, T)\leqq \inf_{\phi\in F} S(\phi, T)-h.$$

It is easily seen that $\{\phi_n\}$ has compact closure. Hence there is ϕ^* such that

$$\phi_n\to\phi^*\in F.$$

By the "lower semicontinuity" result cited above Lemma A1.3,

$$S(\phi^*, T)\leqq \varliminf_n S_{\alpha_n}(\phi_n, T)\leqq \inf_{\phi\in F} S(\phi, T)-h,$$

a contradiction. Q.E.D.

We can now complete the proof of (A1.2b). Let $h>0$ be given. By Lemmas A1.3 and A1.4 we may choose $a>0$, $\delta>0$ so that for small ρ and small fixed α

$$\inf_{\phi\in \bar A^a} S(\phi, T)\geqq \inf_{\phi\in \bar A} S(\phi, T)-h/4,$$

(A1.20)
$$\inf_{\phi\in \bar A^a} S_\alpha(\phi, T)\geqq \inf_{\phi\in \bar A^a} S(\phi, T)-h/4,$$
$$P_x\{\sup_{0\leqq t\leqq T}|x^\rho(t)-x^\rho_\alpha(t)|>a\}\leqq \exp(-M/\lambda_\rho).$$

By picking M large enough, the large deviations results for $x^\rho_\alpha(\cdot)$ (Theorem A1.1 for the $x^\rho_\alpha(\cdot)$ and action functional S_α) and (A1.20) yield (for small ρ and small fixed α)

$$P_x\{x^\rho(\cdot)\in A\}\leqq P_x\{x^\rho_\alpha(\cdot)\in \bar A^a\}+(\exp(-M/\lambda_\rho))$$
$$\leqq \exp(-[\inf_{\phi\in \bar A^a} S_\alpha(\phi, T)-h/4])+(\exp(-M/\lambda_\rho))$$
$$\leqq \exp(-[\inf_{\phi\in \bar A} S(\phi, T)-h]).$$

The proof of Theorem A1.1 is completed for Model I.

The proof for Model II is the same as that of Model I with the following three changes:

(a) When uniformity in $|\xi(0)| \leqq \gamma^{-\delta}\varepsilon^{-1}$ is required for Model I, we require $|\bar{\xi}^{\gamma}(0)| \leqq \gamma^{-\delta+1/2}\varepsilon^{-1}$ for Model II.

(b) When an estimate from Appendix II is used, we substitute the analogous estimate from Appendix IV.

(c) The upper bound on $\alpha_1(t)$ is even easier to get since we can use a standard differential. Q.E.D.

Appendix II. Estimates for noise Model I. The following estimates are used repeatedly in the large deviations proofs associated with the wide bandwidth Gaussian noise model I.

LEMMA A2.1. *Fix $a>0$, $T>0$, and $\delta>0$. For each $M<\infty$ there is a $\gamma_0>0$ such that for $\gamma \leqq \gamma_0$ and all T_0*

$$P\{\sup_{0\leqq t\leqq T/\gamma} [\xi((t+T_0)/\gamma)-\xi(T_0/\gamma)| \geqq a/\varepsilon\gamma^{\delta}\} \leqq \exp(-M/\varepsilon^2). \tag{A2.1}$$

If the probability in (A2.1) *is conditioned on $\xi(T_0/\gamma)$, then the estimates are uniform in $|\xi(T_0/\gamma)| \leqq \varepsilon^{-1}\gamma^{-\delta_1}$ for $\delta_1<\delta$.*

Proof. W.l.o.g., set $T_0=0$, then

$$\begin{aligned} &P\{\sup_{0\leqq t\leqq T/\gamma} |\xi(t)-\xi(0)| \geqq a/\varepsilon\gamma^{\delta}\} \\ &\quad \leqq \sum_{i=0}^{T/\gamma-1} P\{|\xi(i)| \geqq a/4\varepsilon\gamma^{\delta}\} + \sum_{i=0}^{T/\gamma-1} P\{\sup_{i\leqq s\leqq i+1} |\xi(s)-\xi(i)| \geqq a/4\varepsilon\gamma^{\delta}\}. \end{aligned} \tag{A2.2}$$

Via an integration by parts,

$$\begin{aligned} \xi(s)-\xi(i) &= [\exp As - I]\xi(i) + B[w(s)-w(i)] \\ &\quad + \int_i^s A\exp A(s-\tau)B[w(\tau)-w(i)]\,d\tau. \end{aligned} \tag{A2.3}$$

By the stability of A, there is $k<\infty$ such that

$$|\xi(s)-\xi(i)| \leqq k|\xi(i)| + k\sup_{i\leqq s\leqq i+1} |w(s)-w(i)|. \tag{A2.4}$$

Hence the estimate is reduced to sums of terms of the form (for appropriate values of a_1 and a_2)

$$P\{|\xi(i)| \geqq a_1/\varepsilon\gamma^{\delta}\}, \tag{A2.5}$$

$$P\{\sup_{i\leqq s\leqq i+1} |w(s)-w(i)| \geqq a_2/\varepsilon\gamma^{\delta}\}. \tag{A2.6}$$

As $\xi(i)$ is Gaussian, there are K_1 and K_2 such that

$$(\text{A2.5}) \leqq K_1\exp(-K_2/\varepsilon^2\gamma^{2\delta}).$$

To estimate (A2.6), we work with each component of $w(\cdot)$ separately. By the submartingale inequality, for $\lambda>0$,

$$P\{\sup_{i\leqq s\leqq i+1} \exp\lambda(w(s)-w(i)) \geqq \exp\lambda a_2/\varepsilon\gamma^{\delta}\} \leqq \exp(-\lambda a_2/\varepsilon\gamma^{\delta})\cdot\exp\lambda^2/2. \tag{A2.7}$$

By picking $\lambda = a_2/\varepsilon\gamma^{\delta}$, and repeating for $-w(\cdot)$, we see that there are K_3, K_4 such that

$$(\text{A2.6}) \leqq K_3\exp - K_4/\varepsilon^2\gamma^{2\delta}.$$

These estimates imply the first part of the Lemma, for small enough ρ_0.

In order to prove the uniformity assertion, note that the estimates on the Wiener process are unaffected by the values of $\xi(0)$, so all we need to do is get the appropriate bound on

$$P_{\xi(0)}\{|\xi(i)| \geqq \varepsilon^{-1}\gamma^{\delta}\}. \tag{A2.5'}$$

Since

$$\xi(i) = e^{Ai}\xi(0) + \int_0^i e^{A(i-t)}B\,dw(t), \tag{A2.8}$$

$$|\xi(i)| \leqq |\xi(0)| + \left|\int_0^i e^{A(i-t)}B\,dw(t)\right|. \tag{A2.9}$$

Thus there are a_3, a_4 such that

$$(\text{A2.5}') \leqq P_{\xi(0)}\{|\xi(0)| \geqq a_3/\varepsilon\gamma^{\delta}\} + P_{\xi(0)}\left\{\left|\int_0^i e^{A(i-t)}B\,dw(t)\right| \geqq a_4/\varepsilon\gamma^{\delta}\right\}.$$

The assumption $|\xi(0)| \leqq \varepsilon^{-1}\gamma^{-\delta_1}$ and the first part of the proof yield the uniformity. Q.E.D.

LEMMA A2.2. *Let $f(\cdot)$ be bounded (by, say, k) in norm and nonanticipative with respect to $w(\cdot)$. Then,*

$$P\left\{\sup_{0\leqq t\leqq T/\gamma}\left|\int_0^t f(s)\,dw(s)\right| \geqq a/\varepsilon\sqrt{\gamma}\right\} \leqq 2\exp(-a^2/2Tk^2\varepsilon^2). \tag{A2.10}$$

Proof. We can suppose that $w(\cdot)$ and $f(\cdot)$ are scalar valued, and (by symmetry of $w(\cdot)$) drop the absolute value bars. The proof follows by bounding the expectation of the exponential martingale as

$$E\exp\lambda\int_0^{T/\gamma} f(s)\,dw(s) \leqq \exp\lambda^2k^2T/2\gamma,$$

and choosing λ properly, as in the proof of Lemma A2.1. Q.E.D.

LEMMA A2.3. *Fix $a>0$, $T>0$. For each $M<\infty$, there is a $\gamma_0>0$ such that for $\gamma \leqq \gamma_0$,*

$$P\left\{\gamma\int_0^{T/\gamma}|\xi(s)|^2\,ds \geqq a/\varepsilon^2\right\} \leqq \exp(-M/\varepsilon^2). \tag{A2.11}$$

If the probability is conditioned on $\xi(0)$, then the result holds uniformly for $|\xi(0)| \leqq \varepsilon^{-1}\gamma^{-\delta}$ for any $\frac{1}{2} > \delta > 0$.

Proof. For simplicity, we let $\xi(\cdot)$ and $w(\cdot)$ be scalar valued. With the appropriate notational changes, the proof for the vector valued case is the same. The K_i denote constants. Define $Q_n = \int_{n-1}^n \exp A(n-\tau)B\,dw(\tau)$. We have, for $s \in [i, i+1]$

$$\xi(s) = \exp A(s-i)\xi(i) + \int_i^s \exp A(s-\tau)B\,dw(\tau),$$

$$\xi(i) = \sum_{-\infty}^{i} \exp A(i-n)Q_n, \tag{A2.12}$$

$$\int_i^{i+1} ds\left|\int_i^s \exp A(s-\tau)B\,dw(\tau)\right|^2 \leqq K_1\int_i^{i+1}|w(s)-w(i)|^2\,ds.$$

By an integration by parts

$$|Q_n|^2 \leqq K_2\int_{n-1}^n |w(s)-w(i)|^2\,ds + K_2|w(n)-w(n-1)|^2.$$

Now write (for $T/\gamma = \text{integer}$)

$$\int_0^{T/\gamma} |\xi(s)|^2\, ds = \sum_0^{T/\gamma-1} \int_i^{i+1} |\xi(s)|^2\, ds$$

$$\text{(A2.13)} \qquad \leqq K_3 \sum_0^{T/\gamma-1} |\xi(i)|^2 + K_3 \sum_0^{T/\gamma-1} \int_i^{i+1} |w(s)-w(i)|^2\, ds$$

$$\equiv S_1^\gamma + S_2^\gamma.$$

For some $b \in [0,1)$,

$$\text{(A2.14)} \qquad |\xi(i)|^2 \leqq K_4 \sum_{n=-\infty}^{i} b^{i-n} |Q_n|^2,$$

$$\sum_0^{T/\gamma-1} |\xi(i)|^2 \leqq K_5 \sum_0^{T/\gamma-1} Q_n^2 + K_5 \sum_{n=0}^{\infty} b^n Q_{-n}^2.$$

Relations (A2.12) to (A2.14) imply that

$$\text{(A2.15)} \qquad \int_0^{T/\gamma} |\xi(s)|^2\, ds \leqq K_6 \left\{ \sum_{i=1}^{T/\gamma-1} \left[\int_i^{i+1} |w(s)-w(i)|^2\, ds + |w(i+1)-w(i)|^2 \right] \right\}$$

$$+ K_6 \left\{ \sum_{i=-\infty}^{0} b^{-i} \left[\int_i^{i+1} |w(s)-w(i)|^2\, ds + |w(i+1)-w(i)|^2 \right] \right\}.$$

For Gaussian x, $E \exp \alpha x^2 = (1-2\alpha E x^2)^{-1}$ if $Ex = 0$ and $2\alpha E x^2 < 1$. Let us first consider the term

$$\text{(A2.16)} \qquad M^\gamma \equiv \sum_{i=1}^{T/\gamma-1} \int_i^{i+1} |w(s)-w(i)|^2\, ds + \sum_{i=-\infty}^{0} b^{-i} \int_i^{i+1} |w(s)-w(i)|^2\, ds.$$

For any $\lambda > 0$, if $2\lambda\gamma < 1$, then

$$P\{\gamma M^\gamma \geqq a/\varepsilon^2\} \leqq (\exp(-\lambda a/\varepsilon^2)) E \exp \lambda\gamma M^\gamma$$

$$\text{(A2.17)} \qquad = \exp(-\lambda a/\varepsilon^2) \prod_{i=0}^{T/\gamma-1} E \exp \lambda\gamma \int_i^{i+1} |w(s)-w(i)|^2\, ds$$

$$\cdot \prod_{i=-\infty}^{0} E \exp \lambda\gamma b^{-i} \int_i^{i+1} |w(s)-w(i)|^2\, ds.$$

By Jensen's inequality, and the fact that $E \exp \alpha |w(s)-w(i)|^2$ is increasing for $\alpha > 0$, we get

$$\text{(A2.18)} \qquad E \exp \lambda\gamma \int_i^{i+1} |w(s)-w(i)|^2\, ds \leqq E \int_i^{i+1} \exp \lambda\gamma |w(s)-w(i)|^2\, ds$$

$$\leqq E \exp \lambda\gamma |w(i+1)-w(i)|^2.$$

Hence,

$$\text{(A2.19)} \qquad P\{\gamma M^\gamma \geqq a/\varepsilon^2\} \leqq [\exp -\lambda a/\varepsilon^2](1-2\lambda\gamma)^{-T/\gamma-1} \prod_{i=-\infty}^{0} (1-2\lambda\gamma b^{-i})^{-1}.$$

Picking $\lambda = 2M/a$, there is $K_7 < \infty$ such that for small γ,

$$\text{(A2.20)} \qquad P\{\gamma M^\gamma \geqq a/\varepsilon^2\} \leqq K_7 \exp(-2M/\varepsilon^2).$$

The other terms in (A2.15) are easier to estimate, although the same general argument

is used and we omit the details. Hence the first part of the lemma is proved. To see the asserted uniformity in $\xi(0)$ of the conditioned probability, note that for some $K_1 < \infty$,

$$\int_0^{T/\gamma} |\xi(s)|^2\, ds \leqq K_1 \left[\sum_{i=1}^{T/\gamma - 1} \left(\int_i^{i+1} |w(s) - w(i)|^2\, ds + |w(i+1) - w(i)|^2 \right) + |\xi(0)|^2 \right]. \tag{A2.21}$$

This estimate and the previous argument complete the proof. Q.E.D.

LEMMA A2.4. *Under the hypotheses of Lemma* A2.3, *for fixed* M *and small* γ

$$P\left\{ \int_0^{T/\gamma} |\xi(s)|\, ds \geqq a/\varepsilon\gamma^{3/2} \right\} \leqq \exp(-M/\varepsilon^2).$$

If the probability is conditioned on $\xi(0)$, *then the estimate is uniform in* $|\xi(0)| \leqq \varepsilon^{-1}\gamma^{-\delta}$ *for* $\frac{1}{2} > \delta > 0$.

Proof. For any $b > 0$ and $u \geqq 0$, $u \leqq b + u^2/b$. Let $b = a/2\varepsilon\sqrt{\gamma}T$. Then

$$|\xi(s)| \leqq \frac{a}{2\varepsilon\sqrt{\gamma}T} + 2|\xi(s)|^2 \frac{\varepsilon\sqrt{\gamma}T}{a}.$$

Thus, we need only show that

$$P\left\{ \int_0^{T/\gamma} |\xi(s)|^2\, ds \left(\frac{2\varepsilon\sqrt{\gamma}T}{a} \right) \geqq \frac{a}{\varepsilon\gamma^{3/2}} \right\} \leqq \exp(-M/\varepsilon^2)$$

for small ρ. But this follows from Lemma A2.3. The proof for the conditional probability is similar. Q.E.D.

LEMMA A2.5. *Let* $0 < v < 1/2$. *Then for fixed* $M < \infty$, $T < \infty$, *there are* $\gamma_0 > 0$, $\Delta_0 > 0$, *such that for* $\gamma \leqq \gamma_0$, *and* $\Delta \leqq \Delta_0$,

$$P\{\sqrt{\gamma}\varepsilon \sup_{i \leqq T/\Delta} \sup_{t \leqq \Delta} |w(i\Delta + t) - w(i\Delta)| \geqq \Delta^v\} \leqq \exp(-M/\varepsilon^2),$$

$$P\{\sqrt{\gamma}\varepsilon \sup_{i \leqq T/\Delta} \sup_{t \leqq \Delta} |\xi(i\Delta + t) - \xi(i\Delta)| \geqq \Delta^v\} \leqq \exp(-M/\varepsilon^2).$$

If the last probability is conditioned on $\xi(0)$, *and* $0 < \delta < \frac{1}{2}$, *then the estimate is uniform for* $|\xi(0)| \leqq \varepsilon^{-1}\gamma^{-\delta}$.

The proof follows the lines of the previous lemmas and is omitted.

Appendix III. An estimate for the phase locked loop problem. We wish to show that the last (the high frequency) term in (5.3) can be dropped when calculating the large deviations estimates. Here, we do this for a simpler (and scalar) problem, to simplify the notation. We prove the result for noise Model I. An analogous result can be obtained for noise Model II.

Define $\tilde{x}^\rho(\cdot)$ by

$$\begin{aligned} \dot{\tilde{x}}^\rho(t) &= b(\tilde{x}^\rho(t)) + \varepsilon\sigma(\tilde{x}^\rho(t))\xi^\gamma(t) + \varepsilon(\sin \omega_0 t/\eta_\gamma)\sigma_1(\tilde{x}^\rho(t))\xi^\gamma(t), \\ \tilde{x}^\rho(0) &= x \end{aligned} \tag{A3.1}$$

where $\eta_\gamma/\gamma \to 0$, and $\sigma_1(\cdot)$ satisfies the same conditions as does $\sigma(\cdot)$. We first prove the following auxiliary lemma.

LEMMA A3.1. *For each* $M < \infty$, $T > 0$, *and* $a > 0$ *there is a* $\rho_0 > 0$ *such that for* $\rho < \rho_0$,

$$P\{ \sup_{0 \leqq t \leqq T} |I^\rho(t)| \geqq a\} \leqq \exp -M/\varepsilon^2, \tag{A3.2}$$

where

$$I^\rho(t) = \int_0^t \varepsilon(\sin \omega_0 s/\eta_\gamma)\sigma_1(\tilde{x}^\rho(s))\xi^\gamma(s)\, ds.$$

The same estimate holds uniformly in $|\xi(0)| \leqq \varepsilon^{-1}\gamma^{-\delta}$ *for* $\delta \in (0, \frac{1}{2})$, *and in* x *in any compact set, when the probability is replaced by the probability conditioned on* $\xi(0)$.

Proof. By repeating the proof for $-I^\rho$, we may drop the absolute value bars in (A3.2). By a change of variable $s/\gamma \to s$, we obtain

$$I^\rho(t) = \varepsilon\sqrt{\gamma}\int_0^{t/\gamma} (\sin \gamma\omega_0 s/\eta_\gamma)\sigma_1(\tilde{x}^\rho(\gamma s))\xi(s)\, ds.$$

By an integration by parts with

$$da = \sin \gamma\omega_0 s/\eta_\gamma, \qquad b = \sigma_1(\tilde{x}^\rho(\gamma s))\xi(s)\, ds$$

we obtain

$$I^\rho(t) = \frac{\varepsilon\sqrt{\gamma}\eta_\gamma}{\omega_0}\int_0^{t/\gamma} (\cos \gamma\omega_0 s/\eta_\gamma)\sigma_1(\tilde{x}^\rho(\gamma s))A\xi(s)\, ds + \text{terms of smaller order.} \tag{A3.3}$$

By repeating the integration by parts on the first term on the r.h.s. of (A3.3) and collecting like terms, we find $(N = [I + \eta_\gamma^2 A^2/\omega_0^2]^{-1} \cdot [\eta_\gamma I(\cos \gamma\omega_0 s/\eta_\gamma)\omega_0 - \eta_\gamma^2 A(\sin \gamma\omega_0 s/\eta_\gamma)/\omega_0^2])$

$$\begin{aligned} I^\rho(t) = {} & -\varepsilon\sqrt{\gamma}\sigma_1(\tilde{x}^\rho(\gamma s))N\xi(s)|_0^{t/\gamma} \\ & + \varepsilon\sqrt{\gamma}\int_0^{t/\gamma} \sigma_1(x^\rho(\gamma s))NB\, ds(s) \\ & + \varepsilon\gamma^{3/2}\int_0^{t/\gamma} b'(\tilde{x}^\rho(\gamma s))\sigma_{1x}(\tilde{x}^\rho(\gamma s))N\xi(s)\, ds \\ & + \varepsilon^2\gamma\int_0^{t/\gamma} \xi'(s)(\sigma(\tilde{x}^\rho(\gamma s)) \\ & \qquad + (\sin \gamma\omega_0 s/\eta_\gamma)\sigma_1(\tilde{x}^\rho(\gamma s)))'\sigma_{1x}(\tilde{x}^\rho(\gamma s))N\xi(s)\, ds. \end{aligned}$$

As $|N| \to 0$ when $\gamma \to 0$, Lemmas A2.1, A2.2, A2.3, and A2.4 imply the estimates and the required uniformity.

LEMMA A3.2. *Theorem A1.1 holds for the system defined by* (A3.1), *where the S-functional is that for* (A3.1) *with the high frequency term* (*the last one on the right of* (A3.1)) *dropped.*

Proof. The proof is essentially the same as that of Theorem A1.1. Simply add the "high frequency" term in (A3.1) to the β_1^0 in the proof of Theorem A1.1. This is justified by Lemma A3.2. There is also an additional term in the β_2^ρ in Theorem A1.1, due to the "high frequency term." The term appears due to the "high frequency" contribution to the terms which arise during the partial integration. But these terms are of the same type as (A3.1) and can also be collected in the β_1^β. The other details are the same. Q.E.D.

Remark. In (A3.1), we retain only one high frequency term, but the result for (5.3) is the same; i.e., Theorem 4.1 or Theorem A1.1 (equivalently) hold, where the S-functional is that for the system without the high frequency terms; namely for (5.4.).

Appendix IV. Estimates for noise Model II. Estimates analogous to those obtained in Appendix II can also be obtained for the noise Model II. Using the appropriate analogues of Lemmas A2.1 to A2.5, the proofs are the same as for the noise Model I case. In order to obtain the correct forms of the lemmas, simply replace $dw(s)$ by $d\bar{J}^\gamma(s)$ and $\xi(s)$ by $\bar{\xi}^\gamma(s)/\sqrt{\gamma}$ in the statements of the Lemmas A2.1 to A2.5. We only give details where they are significantly different from those used in Appendix II. We do the proofs in the scalar cases for notational convenience. In every lemma, the method of proving the required uniformity is the same as in Appendix II, and the details are omitted. For simplicity, all the details are for scalar case only.

LEMMA A4.1. *Fix $a>0$, $T>0$, and $\delta>0$. For each $M<\infty$ there is $\gamma_0>0$ such that for $\gamma\leqq\gamma_0$ and all T_0*

$$P\{\sup_{t\leqq T/\gamma}|\bar{\xi}^\gamma((t+T_0/\gamma))-\bar{\xi}^\gamma(T_0/\gamma)|\geqq a/\varepsilon\gamma^{\delta-1/2}\}\leqq\exp(-M/\varepsilon^2). \tag{A4.1}$$

If the probability in (A4.1) *is conditioned on $\xi^\gamma(T_0/\gamma)$, then the estimates are uniform in $|\bar{\xi}^\gamma(T_0/\gamma)|\leqq\varepsilon^{-1}\gamma^{-\delta_1+1/2}$ for $\delta_1<\delta$.*

Proof. Following the argument used in Lemma A2.1, all that is needed are estimates on (for arbitrary $a_i>0$)

$$P\{|\bar{\xi}^\gamma(i)|\geqq a_1/\varepsilon\gamma^{\delta-1/2}\}, \tag{A4.2}$$

$$P\{\sup_{i\leqq s\leqq i+1}|\bar{J}^\gamma(s)-\bar{J}^\gamma(i)|>a_2/\varepsilon\gamma^{\delta-1/2}\}. \tag{A4.3}$$

Write $\bar{\xi}^\gamma(i)=\int_{-\infty}^{i}(\exp(-A(s-i)))\,d\bar{J}^\gamma(s)$. Then for $\lambda>0$,

$$\log E\exp\lambda\bar{\xi}^\gamma(i)=\mu_\gamma\gamma\int_0^\infty E[\exp(\lambda\psi^\gamma\exp As)-1]\,ds.$$

Hence as $\gamma\to0$

$$\frac{1}{\gamma}\log E\exp\lambda\bar{\xi}^\gamma(i)\to C_0\frac{\lambda^2}{2}\int_0^\infty\exp 2As\,ds.$$

Via the exponential Chebyshev's inequality, for $\lambda=a_1/C_0[\int_0^\infty\exp 2As\,ds]\varepsilon\gamma^{\delta+1/2}$ we find (for small γ, and similarly for $-\bar{\xi}^\gamma(i)$)

$$\begin{aligned}P\{\bar{\xi}^\gamma(i)\geqq a_1/\varepsilon\gamma^{\delta-1/2}\}&\leqq(\exp-a_1\lambda/\varepsilon\gamma^{\delta-1/2})E\exp\lambda\bar{\xi}^\gamma(i)\\&\leqq\exp\left[-a_1^2\Big/\left[2C_0\left[\int_0^\infty\exp As\,ds\right]\varepsilon^2\gamma^{2\delta}\right]\right].\end{aligned}$$

In order to estimate (A4.3), we use the fact that $\exp\lambda(\bar{J}^\gamma(s)-\bar{J}^\gamma(i))$ is a submartingale to get

$$\begin{aligned}&P\{\sup_{i\leqq s\leqq i+1}(\bar{J}^\gamma(s)-\bar{J}^\gamma(i))\geqq a_2/\varepsilon\gamma^{\delta-1/2}\}\\&\qquad\leqq(\exp(-\lambda a_2/\varepsilon\gamma^{\delta-1/2}))[(E\exp\lambda(\bar{J}^\gamma(i+1)-\bar{J}^\gamma(i))].\end{aligned}$$

Picking $\lambda=a_2/C_0\varepsilon\gamma^{\delta+1/2}$, and using the fact that

$$\lim_\rho[E\exp\lambda(\bar{J}^\gamma(i+1)-\bar{J}^\gamma(i))]=\lim_\rho\gamma\mu_\gamma[E\exp\lambda\psi^\gamma-1]=C_0\lambda^2/2,$$

yields the desired estimate.

LEMMA A4.2. *Let $f(\cdot)$ be bounded (by, say, k) in norm and nonanticipative with respect to $\bar{J}^\gamma(\cdot)$. Then,*

$$P\left\{\sup_{0\leq t\leq T/\gamma}\left|\int_0^t f(s)\,d\bar{J}^\gamma(s)\right|\geq a/\varepsilon\right\}\leq 2\exp-(a^2/2TC_0k^2\varepsilon^2). \tag{A4.4}$$

LEMMA A4.3. *Fix $a>0$, $T>0$. For each $M<\infty$, there is a $\rho_0>0$ such that for $\rho\leq\rho_0$*

$$P\left\{\int_0^{T/\gamma}|\bar{\xi}^\gamma(s)|^2\,ds\geq a/\varepsilon^2\right\}\leq\exp(-M/\varepsilon^2). \tag{A4.5}$$

If the probability is conditioned on $\bar{\xi}^\gamma(0)$ then the result holds uniformly for $|\bar{\xi}^\gamma(0)|\leq\varepsilon^{-1}\gamma^{-\delta+1/2}$ for any $\frac{1}{2}>\delta>0$.

Proof. By an expansion analogous to that used in Lemma A2.3, there is a K such that

$$\begin{aligned}\int_0^{T/\gamma}|\bar{\xi}^\gamma(s)|^2\,ds &\leq K\sum_1^{T/\gamma-1}\int_i^{i+1}|\bar{J}^\gamma(s)-\bar{J}^\gamma(i)|^2\,ds\\ &\quad+K\sum_1^{T/\gamma-1}|\bar{J}^\gamma(i+1)-\bar{J}^\gamma(i)|^2\\ &\quad+K\sum_{-\infty}^{0}b^{-i}\int_i^{i+1}|\bar{J}^\gamma(s)-\bar{J}^\gamma(i)|^2\,ds+K\sum_{-\infty}^{0}b^{-i}|\bar{J}^\gamma(i+1)-\bar{J}^\gamma(i)|^2\\ &\equiv S_1^\gamma+S_2^\gamma+S_3^\gamma+S_4^\gamma.\end{aligned}$$

We evaluate only

$$P\{S_1^\gamma\geq a/\varepsilon^2\}, \tag{A4.6a}$$

as the others are handled in a similar manner. Recall the restriction (2.3):

$$\gamma^{1+\alpha/2}/\varepsilon\to 0\quad\text{as }\rho\to 0, \tag{A4.6b}$$

and that $\mu_\gamma=O(\gamma^{-2-\alpha})$ for some $\alpha>0$. For notational convenience (and w.l.o.g.), we let $\mu_\gamma=\gamma^{-2-\alpha}$. Let N_i denote the number of jumps of $\bar{J}^\gamma(\cdot)$ on the interval $[i,i+1)$. Then

$$P\{N_i=n\}=\exp(-\gamma\mu_\gamma)(\gamma\mu_\gamma)^n/n! \tag{A4.7}$$

Fix $m>1$. The terms in (A4.7) decrease geometrically for $n\geq\gamma\mu_\gamma m$, hence for $n\geq\mu_\gamma m$ if $\gamma<1$. Thus, using Stirling's formula (and always using $\gamma<1$) there are K_i (not depending on γ) such that

$$\begin{aligned}P\{N_i\geq m\mu_\gamma\}&\leq K_1(\exp-\gamma\mu_\gamma)(\gamma\mu_\gamma)^{m\mu_\gamma}/(m\mu_\gamma)!\\ &\leq K_2\exp-\left[1+\frac{m}{\gamma}\left(\log\left(\frac{m}{\gamma}\right)-1\right)\right]\gamma\mu_\gamma.\end{aligned} \tag{A4.8}$$

Hence

$$P\{\sup_{i\leq T/\gamma}N_i\geq m\mu_\gamma\}\leq\frac{T}{\gamma}\exp-\left[1+\frac{m}{\gamma}\left(\log\left(\frac{m}{\gamma}\right)-1\right)\right]\gamma^{-1-\alpha}.$$

By (A4.56b), we have

$$\varepsilon^2\left[1+\frac{m}{\gamma}\left(\log\left(\frac{m}{\gamma}\right)-1\right)\right]\gamma^{-1-\alpha}\geq(\varepsilon^2m/\gamma^{2+\alpha})\left(\log\left(\frac{m}{\gamma}\right)-1\right)=\delta_\rho\left(\log\left(\frac{m}{\gamma}\right)-1\right)$$

where $\delta_\rho \to \infty$ as $\rho \to 0$. Then

$$P\{\sup_{i \leqq T/\gamma} N_i \geqq m\mu_\gamma\} \leqq T/\gamma \exp\left(-\left[\delta_\rho\left(\log\left(\frac{m}{\gamma}\right)-1\right)\Big/\varepsilon^2\right]\right) \leqq \exp(-\delta_\rho/\varepsilon^2)$$

for small ρ. Henceforth, in evaluating (A4.6a), we can assume that $\delta_\rho > 2M$ and that $N_i \leqq m\mu_\gamma$. We have for $\lambda > 0$,

$$\begin{aligned} E\exp\lambda S_1^\gamma &= \left(E\exp\lambda\int_0^1 |\bar{J}^\gamma(s)-\bar{J}^\gamma(0)|^2\,ds\right)^{T/\gamma} \\ &\leqq (E\exp\lambda|\bar{J}^\gamma(1)-\bar{J}^\gamma(0)|^2)^{T/\gamma}. \end{aligned} \tag{A4.9}$$

The inequality in (A4.9) follows from Jensen's inequality and the fact that if $\{z_1, z_2\}$ is a zero mean martingale, then by the submartingale inequality,

$$E\exp\lambda|z_1|^2 \leqq E\exp\lambda|z_2|^2.$$

Neglecting jumps above $m\mu_\gamma$ on each interval $[i, i+1)$, we get

$$E\exp\lambda|\bar{J}^\gamma(1)-\bar{J}^\gamma(0)|^2 \sim \sum_{n=0}^{m\mu_\gamma} E\exp\lambda\left(\sum_1^n \psi_i^\gamma\right)^2 (\exp(-\gamma\mu_\gamma))(\gamma\mu_\gamma)^n/n!,$$

where the $\{\psi_i^\gamma\}$ are mutually independent and each has the distribution of ψ^γ. Assume for the moment that the ψ_i^γ are normally distributed. Then $\sum_1^n \psi_i^\gamma$ is normally distributed in the mean zero and variance $n\nu_\gamma$, and for small enough λ we would have the bounds

$$\begin{aligned} \sum_0^{m\mu_\gamma}(1-2\lambda n\nu_\gamma)^{-1}(\exp(-\gamma\mu_\gamma))(\gamma\mu_\gamma)^n/n! &\leqq \sum_0^{m\mu_\gamma}(1+4\lambda n\nu_\gamma)(\exp(-\gamma\mu_\gamma))(\gamma\mu_\gamma)^n/n! \\ &\leqq 1+4\lambda\nu_\gamma\mu_\gamma\gamma. \end{aligned}$$

Thus for purposes of evaluating (A4.6), we may bound (A4.9) by

$$(1+4\lambda C_0\gamma)^{T/\gamma} \leqq \exp 4\lambda C_0 T.$$

Next, use the exponential Chebyshev inequality to obtain (neglecting a set whose probability is $\leqq \exp(-2M/\varepsilon^2)$)

$$P\{S_1^\gamma \geqq a/\varepsilon^2\} \leqq (\exp(-\lambda a/\varepsilon^2))\exp 4\lambda C_0 T$$

choosing $\lambda = 2M/a$ yields the desired estimate for small γ.

If the ψ_i are not normally distributed, the estimate remains the same, owing to the assumptions made on the upper bounds to the moments of the ψ_i^γ in (2.3). The same procedure is used to estimate S_2^γ, S_3^γ, and S_4^γ. Q.E.D.

LEMMA A4.4 *Under the hypotheses of Lemma* A4.3, *for any fixed M and small* ρ

$$P\left\{\int_0^{T/\gamma} |\bar{\xi}^\gamma(s)|\,ds \geqq a/\varepsilon\gamma\right\} \leqq \exp(-M/\varepsilon^2).$$

If the probability is conditioned on $\bar{\xi}^\gamma(0)$, *then the estimate is uniform in* $|\bar{\xi}^\gamma(0)| \leqq \varepsilon^{-1}\gamma^{-\delta+1/2}$, *for* $\frac{1}{2} > \delta > 1$.

LEMMA A4.5. *Let* $0 < \nu < \frac{1}{2}$. *Then for fixed* $M < \infty$ *and* $T < \infty$, *there are* $\rho_0 > 0$ *and* $\Delta_0 > 0$ *such that for* $\rho \leqq \rho_0$ *and* $\Delta \leqq \Delta_0$,

$$P\{\varepsilon \sup_{i \leqq T/\Delta} \sup_{t \leqq \Delta} |\bar{J}^\gamma(i\Delta+t)-\bar{J}^\gamma(i\Delta)| \geqq \Delta^\nu\} \leqq \exp(-M/\varepsilon),$$

$$P\{\varepsilon \sup_{i \leqq T/\Delta} \sup_{t \leqq \Delta} |\bar{\xi}^\gamma(i\Delta+t)-\bar{\xi}^\gamma(i\Delta)| \geqq \Delta^\nu\} \leqq \exp(-M/\varepsilon^2).$$

If the last probability is conditioned on $\bar{\xi}^\gamma(0)$, *and* $0 < \delta < \frac{1}{2}$, *then the estimate is uniform for* $|\bar{\xi}^\gamma(0)| \leqq \varepsilon^{-1}\gamma^{-\delta+1/2}$.

REFERENCES

[1] M. I. FREIDLIN AND A. D. VENTSEL, *Large Deviations*, Springer, Berlin, 1984.

[2] R. AZENCOTT, *Grandes déviations et applications*, Lecture Notes in Mathematics 774, Springer, Berlin, 1980.

[3] B. Z. BOBROVSKY AND Z. SCHUSS, *A singular perturbation method for the computation of the mean first passage time in a non-linear filter*, SIAM J. Appl. Math., 42 (1982), pp. 174–187.

[4] H. KUSHNER, *A cautionary note on the use of singular perturbations*, Stochastics, 6 (1982), pp. 117–120.

[5] ———, *Robustness and approximation of escape times and large deviations estimates for systems with small noise effects*, SIAM J. Appl. Math., 44 (1984), pp. 160–182.

[6] A. D. VENTSEL, *Rough limit theorems on large deviations for Markov stochastic processes*, Theory Prob. Appl., 21 (1976), pp. 227–242, pp. 499–512.

[7] M. I. FREIDLIN, *The averaging principle and theorems on large deviations*, Russian Math. Surveys, 33 (July–Dec., 1978), pp. 117–176.

[8] S. R. S. VARADHAN, *Large Deviations and Applications*, CBMS Regional Conference Series in Applied Mathematics, 46, Society for Industrial and Applied Mathemetics, Philadelphia, 1984.

[9] R. AZENCOTT AND G. RUGET, *Melanges d'equations differentialles et grands écartes à la loi des grand nombres*, Z. Warsch., 38 (1977), pp. 1–54.

[10] A. J. VITERBI, *Principles of Coherent Communication*, McGraw-Hill, New York, 1966.

[11] W. C. LINDSEY AND M. K. SIMON, *Telecommunication Systems Engineering*, Prentice-Hall, Englewood Cliffs, NJ, 1973.

[12] L. A. ZADEH AND C. A. DESOER, *Linear System Theory: The State Space Approach*, McGraw-Hill, New York, 1963.

[13] D. LUDWIG, *Persistence of dynamical systems under random perturbations*, SIAM Rev., 17 (1975), pp. 605–640.

[14] Z. SCHUSS, *Theory and Applications of Stochastic Differential Equations*, John Wiley, New York, 1980.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 5, September 1986

007

INVARIANCE CONCEPTS IN INFINITE DIMENSIONS*

RUTH F. CURTAIN†

Abstract. The concepts of (A, B) and (C, A)-invariance are examined for infinite-dimensional linear systems. Examples such as reachability, A-reachability, inobservability and A-inobservability subspaces are given. These concepts are then used to solve various disturbance decoupling problems.

Key words. (A, B)-invariance, (C, A)-invariance, observability, reachability, disturbance decoupling, infinite-dimensional systems

1. Introduction. The concept of (A, B)-invariance has been used by Wonham in [11] to solve various decoupling and control problems. Use is made of the equivalence of (A, B)-invariance of a subspace V: $AV \subset V + \operatorname{Im} B$ and the holdability under the linear system $\dot{x} = Ax + Bu$. Unfortunately this does not hold in general in infinite-dimensions, as indicated by Schmidt and Stern in [6], where they investigated this concept in detail. Here we take a different approach and seek natural conditions under which (A, B)-invariance and holdability are equivalent: for example V is closed and in $D(A)$ and B has finite rank. A key concept in solving disturbance decoupling problems is the existence of a supremal (A, B)-invariant subspace contained in a closed subspace. Although this exists in infinite dimensions it is not in general invariant for trajectories of $\dot{x} = Ax + Bu$ and this is the property which is needed to solve disturbance decoupling problems. We are able to establish sufficient conditions for this property of trajectory invariance, but only by considering the dual concept of the trajectory invariance of the infimal (C, A)-invariant subspace. These dual concepts have also been exploited by researchers to solve various disturbance decoupling and regulator problems, for example by Schumacher in [7] and Willems and Commault in [10]. Surprisingly enough the duality in infinite dimensions is incomplete, but fortunately sufficient to allow us to solve several interesting disturbance decoupling problems.

The organization of this paper is as follows:

§ 2. A-invariance. § 3. (C, A)-invariance. § 4. (A, B)-invariance.
§ 5. Duality between (A, B)- and (C, A)-invariance.
§ 6. Supremal (A, B)-invariant subspaces and infimal (C, A)-invariant subspaces.
§ 7. Some disturbance decoupling problems.
§ 8. Examples.

Essentially the results on disturbance decoupling are that if the supremal (A, B)- (or infimal (C, A))-invariant subspace in question is trajectory invariant and we have finite-dimensional inputs and outputs and some operators are smooth, then we obtain the solvability conditions reminiscent of the finite-dimensional case in [10] and [11]. Some sufficient conditions for trajectory invariance of the supremal (A, B)- or infimal (C, A)-invariant subspaces are given in § 6 and they are applied to retarded systems using the M^2-formulation of [9] and also to the so-called spectral systems introduced by Curtain in [1] in § 8. These illustrate the fact that the existence of a trajectory invariant supremal (A, B)-invariant subspace is in general difficult to establish, especially for retarded systems. Various extensions have been completed recently, such as the disturbance decoupling problem by measurement feedback in [2] and disturbance

* Received by the editors November 15, 1983, and in revised form July 23, 1985.

† Institute of Mathematics, University of Groningen, 9700 AV Groningen, The Netherlands. This paper was written while the author was on study leave at the Department of Systems Engineering, Australian National University, Canberra, Australia.

decoupling problems with boundary control or boundary disturbances in [3]. It is also possible to solve disturbance decoupling problems with stability for certain classes of systems using a similar approach to that in [10] (see [1a]).

2. *A*-invariance. Throughout X will denote a real, separable Hilbert space and A the infinitesimal generator of a strongly continuous semigroup $T(t)$ on X. This implies that A is closed, linear and densely defined on X. We summarize the useful factors about A-invariance from Taylor and Lay [8].

DEFINITION 2.1. *A-invariance.* The linear subspace V of X is A-invariant if $A(V \cap D(A)) \subset V$. If V is closed, then X has the direct sum decomposition

$$X = V \oplus V^{\perp}$$

and relative to this A has the following block decomposition

$$A = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}.$$

The case of a diagonal decomposition with $A_{12} = 0$ corresponds to the following.

DEFINITION 2.2. *Reducing subspaces.* Let M_1, M_2 be 2 closed subspaces of X such that $X = M_1 \oplus M_2$. Then (M_1, M_2) are said to completely reduce A if M_i is A-invariant and $P_i D(A) \subset D(A)$, $i = 1, 2$, where P_1 is the continuous projection on M_1 along M_2 and similarly for P_2. Sufficient conditions for (M_1, M_2) to completely reduce A are that $P_i D(A) \subset D(A)$ and $P_i A x = A P_i x$ for all $x \in D(A)$ hold for *one* value of i.

We are interested in conditions under which the A-invariance of V implies that V is also $T(t)$-invariant. That this is not in general true was pointed out by Schmidt and Stern in [6], and sufficient conditions were established by Schumacher in [7, Chap. 4].

LEMMA 2.3. *Suppose that A is the infinitesimal generator of a semigroup $T(t)$ on X and that $X = S \oplus M$, where S and M are closed linear subspaces of X. If S is A-invariant and $S \subset D(A)$, then S is also $T(t)$ invariant for all $t \geqq 0$. Furthermore, the restriction of A to S is the infinitesimal generator of the semigroup $T(t)|S$. The factor space X/S is well defined and we can define the mappings $\bar{T}(t)$ and $\bar{A}$ by*

$$\bar{T}(t)(x \bmod S) = (T(t)x) \bmod S, \tag{2.1}$$

$$\bar{A}(x \bmod S) = (Ax) \bmod S \text{ for } x \in D(A). \tag{2.2}$$

Then $\bar{T}(t)$ is a semigroup on X/S whose generator is an extension of $\bar{A}$. We shall write $[T(t)]_S$ and $[A]_S$, for $\bar{T}(t)$ and its generator.

Another special case where A-invariance implies $T(t)$-invariance is when S and M are reducing subspaces for A.

LEMMA 2.4. *Suppose that (M_1, M_2) completely reduce A, then they also reduce $T(t)$ for $t \geqq 0$.*

Proof. Since (M_1, M_2) reduces A, we have that $P_1 D(A) \subset D(A)$ and $P_1 A x = A P_1 x$ for $x \in D(A)$. Then for $\lambda \notin \sigma(A)$, we have $P_1 R(\lambda, A)x = R(\lambda, A) P_1 x$. Now from [4] we have

$$R(\lambda, A)x = \int_0^{\infty} e^{-\lambda t} T(t)x \, dt$$

and so

$$0 = P_1 R(\lambda, A)x - R(\lambda, A) P_1 x = \int_0^{\infty} e^{-\lambda t} (P_1 T(t)x - T(t) P_1 x) \, dt;$$

and by the uniqueness of the Laplace transform, we obtain

$$P_1 T(t) = T(t) P_1$$

and M_1 reduces $T(t)$. Since $P_2 = I - P_1$, M_2 also reduces $T(t)$.

$T(t)$-invariance does however imply a type of A-invariance.

LEMMA 2.5. *If V is $T(t)$ invariant for $t \geqq 0$, then $A(V \cap D(A)) \subset \bar{V}$.*

Proof. $T(t)V \subset V$ for $t \geqq 0$, and for $v \in V \cap D(A)$. We may differentiate with respect to t

$$AT(t)V \subset \bar{V} \Rightarrow A(V \cap D(A)) \subset \bar{V}.$$

Lemmas 2.3 and 2.5 indicate that to obtain useful results we should only consider closed invariant subspaces, which we henceforth do.

3. (C, A)-invariance. We consider the following observed system on X

$$\begin{aligned} &\dot{x} = Ax, \qquad x(0) = x_0, \\ &y = Cx = CT(t)x_0 \end{aligned} \tag{3.1}$$

where $C \in \mathscr{L}(X, Y)$, $x_0 \in X$ and Y is a separable Hilbert space.

The finite-dimensional concept of (C, A)-invariance extends naturally to the following.

DEFINITION 3.1. *(C, A)-invariance.* A closed linear subspace S of X is (C, A)-invariant if

$$A(S \cap \operatorname{Ker} C \cap D(A)) \subset S \tag{3.2}$$

Anticipating that this will not be the same as the semigroup invariance we define $T(C, A)$-invariance.

DEFINITION 3.2. *$T(C, A)$-invariance.* A closed linear subspace S and X is $T(C, A)$-invariant if there exists a $G \in \mathscr{L}(Y, X)$ such that S is invariant under the semigroup $T_{A+GC}(t)$, which is generated by $A + GC$.

We shall also use a third related concept as follows.

DEFINITION 3.3. *Feedback (C, A)-invariance.* A closed linear subspace S of X is feedback (C, A)-invariant if there exists a $G \in \mathscr{L}(Y, X)$ such that

$$(A + GC)(S \cap D(A)) \subset S. \tag{3.3}$$

If S is $T(C, A)$-invariant, then we have the interpretation that the data processor on X/S defined by

$$\dot{w} = [A + GS]_S w - [G]_S y, \qquad w(0) \in S \tag{3.4}$$

with error dynamics

$$\dot{e} = [A + GC]_S e, \qquad e = w - [z]_S$$

is an estimator for $[z]_S$ in (3.1).

It is not enough that S be (C, A)-invariant, as then we could not guarantee that $[A + GC]_S$ is a generator of a semigroup on X/S. (See Lemma 2.3 and Willems and Commault [10]).

We are interested in conditions under which Definitions 3.1, 3.2, 3.3 coincide.

LEMMA 3.4. *$T(C, A)$-invariance $\Rightarrow$ feedback (C, A)-invariance $\Rightarrow$ (C, A)-invariance.*

Proof. Suppose that $s \in S \cap D(A)$. Then since S is closed,

$$\frac{d}{dt}(T_{A+GC}(t))S = (A+GC)T_{A+GC}(t)S \subset S$$

and setting $t=0$, we have $(A+GC)s \in S$ which yields (3.3) and this implies (3.2).

LEMMA 3.5. *Feedback (C, A)-invariance and $T(C, A)$-invariance are equivalent in either of the following cases*:

(a) *$S \subset D(A)$ and S is closed*;

(b) *S is reducing subspace for $A+GC$.*

Proof. Lemmas 2.3 and 2.4 since $D(A+GC) = D(A)$. Our main result is the following.

LEMMA 3.6. *Suppose that S and CS are closed subspaces, $S \subset D(A)$, then the following statements are equivalent*;

(i) *S is (C, A)-invariant.*

(ii) *There exists a bounded linear map $G: Y \to X$ such that $(A+GC)S \subset S$.*

Proof. (ii)$\Rightarrow$(i) is obvious.

(i)$\Rightarrow$(ii): Define $W = S \cap (S \cap \operatorname{Ker} C)^{\perp}$. Then W is closed and $W \cap \operatorname{Ker} C = \{0\}$, and

$$S = W \oplus (S \cap \operatorname{Ker} C). \tag{3.5}$$

Let $Y_0 = CW$ and note that since $W \cap \operatorname{Ker} C = \{0\}$, C is $1-1$ from W to Y_0. By our assumption, Y_0 is closed and thus C^{-1} is a bounded operator from Y_0 to W (Taylor and Lay [8]). We now decompose Y as follows

$$Y = Y_0 \oplus Y_0^{\perp} \tag{3.6}$$

and define

$$\begin{aligned} &G: Y \to X \text{ by} \\ &Gy = 0 \text{ if } y \in Y_0^{\perp}, \end{aligned} \tag{3.7}$$

$$\begin{aligned} Gy = -As \text{ if } y \in D(G) &= \{y \in Y_0 \colon S = C^{-1}y \in D(A)\} \\ &= Y_0 \text{ since } S \in D(A). \end{aligned} \tag{3.8}$$

We now show that that restriction of G to Y_0 is closed. Suppose that $\{y_i\} \subset Y_0$ and $y_i \to y$ in Y_0, $Gy_i \to x \in X$. Since $y_i \in Y_0$ there exists $s_i \in W$ such that $Gy_i = -As_i$ and $y_i = Cs_i$. Since C^{-1} is bounded from Y_0 to W, $s_i = C^{-1}y_i \to s = C^{-1}y$ as $i \to \infty$. Now A is closed and $s_i \to s$ together with $As_i \to x$ imply that $s \in D(A)$ and $As = -x$. Summarizing, we have shown that $Gy_i \to As$ and $y_i \to y = Cs$ as $i \to \infty$ and from (3.8) we have that $Gy_i \to Gy$ as $i \to \infty$, completing the proof that G is closed. Since $D(G) \supset Y_0$, G is bounded. Now if $s \in S \cap \operatorname{Ker} C$, we have $(A+GC)s = As \in S$ by (i) and if $s \in S$, $s \notin \operatorname{Ker} C$, then $s \in W$, $Cs \in D(G)$ and from (3.8) $(A+GC)s = As - As = 0$. Thus (ii) holds.

We remark that CS will be closed if C has finite rank and this is usually the case in applications, so we state a corollary for this special case:

COROLLARY 3.7. *If C has finite rank and S is closed and $\subset D(A)$, then the three concepts $T(C, A)$-invariance, feedback (C, A)-invariance and (C, A)-invariance are all equivalent.*

We now give examples of some infinite-dimensional $T(C, A)$-invariant subspace, which are useful in applications.

DEFINITION 3.8. *Inobservability subspace.* For the system (3.1) we define the inobservability subspace, $\langle \operatorname{Ker} C|T(t)\rangle$ by

$$\langle \operatorname{Ker} C|T(t)\rangle = \bigcap_{t\geqq 0} \operatorname{Ker} CT(t) = \operatorname{Ker}\left(\bigcap_{t\geqq 0} CT(t)\right). \tag{3.9}$$

$\langle \operatorname{Ker} C|T(t)\rangle$ is closed and if $\langle \operatorname{Ker} C|T(t)\rangle = \{0\}$, then the system is initially observable (cf. Curtain and Pritchard [4]). Useful properties of inobservability subspaces are the following.

LEMMA 3.9.

(a) $\langle \operatorname{Ker} C|T(t)\rangle$ *is the largest* $T(t)$*-invariant subspace which is contained in* $\operatorname{Ker} C$;

(b) $\langle \operatorname{Ker} C|T(t)\rangle$ *is output-injection-invariant*: $\langle \operatorname{Ker} C|T_{A+GC}(t)\rangle = \operatorname{Ker}\langle C|T(t)\rangle$;

(c) $\langle \operatorname{Ker} C|T(t)\rangle$ *is* $T(C, A)$*-invariant*;

(d) *If* $\operatorname{Ker} C$ *is* $T(t)$*-invariant, then* $\langle \operatorname{Ker} C|T(t)\rangle = \operatorname{Ker} C$;

(e) *If* $\operatorname{Ker} C \subset D(A)$, *then* $\langle \operatorname{Ker} C|T(t)\rangle = \bigcap_{n=0}^{\infty} CA^n$.

Proof.

(a) If $x \in \langle \operatorname{Ker} C|T(t)\rangle$ then $CT(0)x = Cx = 0$. So $x \in \operatorname{Ker} C$. Furthermore, $T(a)x \in \langle \operatorname{Ker} C|T(t)\rangle$ for all $a > 0$, since

$$CT(t)T(a)x = CT(t+a)x \quad \text{for } a+t \geqq 0.$$

Suppose that Q is another closed $T(t)$-invariant subspace with $Q \subset \operatorname{Ker} C$. If $q \in Q$, then $CT(t)q = 0$ for all $t \geqq 0$ and so $Q \subset \langle \operatorname{Ker} C|T(t)\rangle$.

(b) The semigroup $T_{A+GC}(t)$ generated by $A+GC$ is given by [4],

$$T_{A+GC}(t) = T(t) + \int_0^t T_{A+GC}(t-s)GCT(s)\, ds. \tag{3.10}$$

So if $x \in \langle \operatorname{Ker} C|T(t)\rangle$, $x \in \langle \operatorname{Ker} C|T(t)\rangle$ since C is bounded. The converse holds since $T(t)$ is generated by $(A+GC)-GC$.

(c) From (b) $\langle \operatorname{Ker} C|T_{A+GC}(t)\rangle = \langle \operatorname{Ker} C|T(t)\rangle$ and from (a), $\langle \operatorname{Ker} C|T(t)\rangle$ is $T(C, A)$-invariant and by Lemma 3.4, (C, A)-invariant.

(d), (e) obvious.

We find that the following related inobservability subspace is important for the applications.

DEFINITION 3.10. *A-inobservability subspace.* We define the A-inobservability subspace, $\langle \operatorname{Ker} C|A\rangle$, by

$$\langle \operatorname{Ker} C|A\rangle = \bigcap_{n=0}^{\infty} \operatorname{Ker} CA^n. \tag{3.11}$$

It is easy to show that $\langle \operatorname{Ker} C|A\rangle$ is the largest A-invariant subspace, which is contained in the $\operatorname{Ker} C \cap D(A)$.

Notice that $\langle \operatorname{Ker} C|A\rangle$ is not necessarily closed. If A is bounded, it is easy to show that $\langle \operatorname{Ker} C|A\rangle = \langle \operatorname{Ker} C|T(t)\rangle$, but in general they are different due to the lack of equivalence of A and $T(t)$-invariance. If $\operatorname{Ker} C \subset D(A)$, then

$$\langle \operatorname{Ker} C|A\rangle \supset \langle \operatorname{Ker} C|T(t)\rangle \cap \mathcal{D}(A). \tag{3.12}$$

4. (A, B)-invariance. We consider the following system on X

$$\dot{x} = Ax + Bu, \qquad x(0) = x_0 \tag{4.1}$$

and we define the solution to be mild solution

$$x(t) = T(t)x_0 + \int_0^t T(t-s)Bu(s)\,ds \tag{4.2}$$

which exists and is continuous for all $u \in L_2(0, t; U)$, and $B \in \mathscr{L}(U, X)$ where X is a real, separable Hilbert space.

We shall be concerned with the following concept for (4.1).

DEFINITION 4.1. *(A, B)-invariance.* A closed, linear subspace V of X is (A, B)-invariant if

$$A(V \cap D(A)) \subset \overline{V + \operatorname{Im} B} \tag{4.3}$$

where Im B denotes the range of B.

(A, B)-invariance in infinite dimensions was studied by Schmidt and Stern in [6], where they found that the concept is much more complicated than in finite dimensions. We use a modified version of their definition, because we need the closure of $V + \operatorname{Im} B$ to obtain the existence of a supremal (A, B)-invariant subspace and to obtain a duality with (C, A)-invariance in § 5. Even for our slightly stronger definition, however, (A, B)-invariance does not imply "holdability." This is the property that if the system (4.1) starts with $x(0) = x_0$ in V one can find a control u so that the mild solutions (4.2) always remain in V. In fact we need the following stronger property in our applications.

DEFINITION 4.2. *$T(A, B)$-invariance.* A closed, linear subspace V of X is $T(A, B)$-invariant if there exists an $F \in \mathscr{L}(X, U)$ such that V is invariant under the semigroup $T_{A+BF}(t)$, which is generated by $A + BF$. $T(A, B)$-invariance means that if $x_0 \in V$, then we can choose a feedback control $(u = Fx)$ so that the trajectory of (4.1) remains in V. As already mentioned, this is quite a strong property in infinite dimensions and it is convenient to introduce a weaker concept.

DEFINITION 4.3. *Feedback (A, B)-invariance.* A closed, linear subspace V of X is feedback (A, B)-invariant if there exists an $F \in \mathscr{L}(X, U)$ such that

$$(A + BF)(V \cap D(A)) \subset V. \tag{4.4}$$

In finite dimensions all three types of (A, B)-invariance are equivalent, but in infinite dimensions we have the following hierarchy.

LEMMA 4.4. *$T(A, B)$-invariance $\Rightarrow$ feedback (A, B)-invariance $\Rightarrow$ (A, B)-invariance.*

Proof. $T(A, B)$-invariance means that there exists an $F \in \mathscr{L}(X, U)$ such that

$$T_{A+BF}(t)V \subset V \quad \text{for all } t \geqq 0. \tag{4.5}$$

Differentiating (4.5) with respect to t yields

$$(A + BF)T_{A+BF}(t)V \cap D(A) = \bar{V} = V. \tag{4.6}$$

Letting $t = 0$ we obtain (4.4) and this implies (4.3).

An easy consequence of Lemma 2.3 is the following.

LEMMA 4.5. *If V is closed and $\subset D(A)$, then $T(A, B)$-invariance is equivalent to feedback (A, B)-invariance.*

As pointed out in Schmidt and Stern [6], (A, B)-invariance does not imply $T(A, B)$-invariance, in general, but we can prove equivalence under certain assumptions on V and B.

LEMMA 4.6. *If V is closed and $\subset D(A)$ and* Im B *and* $V+$ Im B *are closed subspaces, then (A, B), feedback (A, B) and $T(A, B)$-invariance are equivalent.*

Proof. In view of Lemmas 4.4 and 4.5 we only need prove that (A, B)-invariance implies feedback (A, B)-invariance. Under our assumptions on V and Im B we obtain the direct sum decomposition

$$V+\operatorname{Im} B = W\oplus \operatorname{Im} B \tag{4.7}$$

where

$$W = V\cap (V\cap \operatorname{Im} B)^{\perp}.$$

Let P be the projection of $W\oplus$ Im B onto Im B along W. We introduce further

$$U_0 = U/\operatorname{Ker} B \text{ and write } U = U_0\oplus \operatorname{Ker} B. \tag{4.8}$$

Now $AV\subset \bar{V}+$ Im $B = W\oplus$ Im B implies that for all $v\in V$ there exists a unique $u\in U_0$ so that

$$PAv = Bu. \tag{4.9}$$

We define $F: V\to U_0$ by

$$\begin{aligned} &Fv = -u \quad \text{if } v\in V,\\ &F = 0 \text{ on } V^{\perp}. \end{aligned} \tag{4.10}$$

Since $V\subset D(A)$ and A is closed we may conclude that A restricted to a closed subspace V is bounded (Taylor and Lay [8, p. 213]) and so PA is bounded. We now prove that F is closed by considering $v_i\to v$ in V and $u_i = -Fv_i\to y$ in U_0. Since PA is bounded, from (4.9) we have $PAv_i\to PAv$ and $Bu_i\to By$ as $i\to\infty$. So $PAv = By$ and $Fv = -y$ proving that F is closed. Since F is everywhere defined, F is bounded and V is invariant under $A+BF$.

We remark that Lemma 4.6 is not a strict dual of Lemma 3.6 and we have been unable to prove one. We do, however, have a dual to Corollary 3.7.

COROLLARY 4.7. *If B has finite rank and V is closed and $\subset D(A)$ then the three concepts $T(A, B)$-invariance, feedback (A, B)-invariance and (A, B)-invariance are equivalent.*

We remark in passing that if V is the span of finitely many eigenvectors of $A+BF$ for some bounded map F, then V satisfies Corollary 4.7 and these finite-dimensional (A, B)-invariant subspaces are used implicitly in the work of Schumacher in [7]. An infinite-dimensional example is the following.

DEFINITION 4.8. *Reachability subspace.* The reachability subspace $\langle T(t)|\operatorname{Im} B\rangle$ is defined by

$$\langle T(t)|\operatorname{Im} B\rangle = \overline{\bigcup_{t\geqq 0}\left\{\int_0^t T(t-s)Bu(s)\,ds\colon u\in L_2(0, t; U)\right\}}. \tag{4.11}$$

$\langle T(t)|\operatorname{Im} B\rangle$ is by definition closed and if $\langle T(t)|\operatorname{Im} B\rangle = X$, then (3.1) is approximately controllable. (See Curtain and Pritchard [4].) $\langle T(t)|\operatorname{Im} B\rangle$ is not in $D(A)$ in general, but we can prove the following.

LEMMA 4.9

(a) $\langle T(t)|\operatorname{Im} B\rangle$ *is the smallest closed, $T(t)$-invariant subspace which contains* Im B.

(b) $\langle T(t)|\operatorname{Im} B\rangle$ *is feedback invariant:* $\langle T_{A+BF}(t)|\operatorname{Im} B\rangle = \langle T(t)|\operatorname{Im} B\rangle$.

(c) $\langle T(t)|\operatorname{Im} B\rangle$ *is an (A, B)-invariant subspace.*

(d) *If* Im B *is $T(t)$-invariant, then* $\langle T(t)|\operatorname{Im} B\rangle = \overline{\operatorname{Im} B}$.

(e) $\overline{\langle T(t)|\operatorname{Im} B\rangle}\cap D(A) = \langle T(t)|\operatorname{Im} B\rangle$.

Proof.

(a) If $r \in \langle T(t)|\text{Im } B\rangle$ there exists a sequence $t_n \in R^+$, $u_n(\cdot) \in L_2([0, t]; U)$ so that

$$r = \lim_{n\to\infty} \int_0^{t_n} T(t_n - s)Bu_n(s)\, ds,$$

$$T(a)r = \lim_{n\to\infty} \int_0^{t_n} T(a + t_n - s)Bu_n(s)\, ds \quad \text{by the semigroup property} \tag{4.12}$$

$$= \lim_{n\to\infty} \int_0^{t_n} T(a + t_n - s)B\hat{u}_n(s)\, ds,$$

where

$$\hat{u}_n(s) = \begin{cases} u_n(s) & \text{on } [0, t_n], \\ 0 & \text{on } [t, t_n + a]. \end{cases}$$

Clearly,

$$T(a)r \in \langle T(t)|\text{Im } B\rangle.$$

Finally, for $u \in U$,

$$Bu = \lim_{t\to 0} 1/t \int_0^t T(t-s)Bu\, ds \in \rangle T(t)|\text{Im } B\rangle$$

since $\langle T(t)|\text{Im } B\rangle$ is closed. Suppose now that Q is another closed, $T(t)$-invariant subspace and $\text{Im } B \subset Q \subset \langle T(t)|\text{Im } B\rangle$. Consider r in $\langle T(t)|\text{Im } B\rangle$ given by (4.12). Then $Bu_n(s) \in Q$ since $Q \supset \text{Im } B$ and $T(t_n - s)Bu_n(s) \in Q$, since Q is $T(t)$-invariant, and since Q is closed $r \in Q$ and $Q = \langle T(t)|\text{Im } B\rangle$.

(b) The semigroup $T_{A+BF}(t)$ generated by $A + BF$ is given by [4]

$$T_{A+BF}(t) = T(t) + \int_0^t T(t-s)BFT_{A+BF}(s)\, ds \tag{4.13}$$

so

$$\int_0^t T_{A+BF}(t-s)Bu(s)\, ds$$

$$= \int_0^t T(t-s)Bu(s)\, ds + \int_0^t \int_s^t T(t-s)BFT_{A+BF}(\alpha - s)\, d\alpha\, Bu(s)\, ds$$

$$= \int_0^t T(t-s)Bu(s)\, ds + \int_0^t T(t-\alpha)B\left(\int_0^\alpha T_{A+BF}(\alpha - s)Bu(s)\, ds\right) d\alpha$$

$$\in \langle T(t)|\text{Im } B\rangle.$$

So $\langle T_{A+BF}(t)|\text{Im } B\rangle \subset \langle T(t)|\text{Im } B\rangle$ and a similar argument proves the reverse inclusion.

(c) $\langle T(t)|\text{Im } B\rangle$ is $T(t)$-invariant and $T_{A+BF}(t)$-invariant for $F = 0$ and so it is also (A, B)-invariant.

(d) Obvious.

(e) Since the simple functions are dense in $L_2(0, t; U)$, if $x \in \langle T(t)|\text{Im } B\rangle$ there exists a sequence $\{x_n\}$ converging to x, where

$$x_n = \int_0^{t_n} T(t_n - s)Bu_n\, ds = \int_0^{t_n} T(s)Bu_n\, ds$$

and from [4, p. 15], $x_n \in D(A) \cap \langle T(t)|\text{Im } B\rangle$.

In the applications it turns out that the following concept is more useful.

DEFINITION 4.10. *A-reachability subspace.* We define the A-reachability subspace of (4.1) to be

$$\langle A|\operatorname{Im} B\rangle = \operatorname{span}\{A^n B U^\infty;\ n = 0, 1, \cdots\} \tag{4.14}$$

where $U^\infty = \{u \in U: Bu \in D(A^\infty) = \bigcap_{n=1}^\infty D(A^n)\}$. $\langle A|\operatorname{Im} B\rangle$ is the smallest A-invariant subspace which contains $\operatorname{Im} B \cap D(A^\infty)$, but $\langle A|\operatorname{Im} B\rangle$ is not closed in general.

If A is bounded, then it is easy to see that $\langle A|\operatorname{Im} B\rangle = \langle T(t)|\operatorname{Im} B\rangle$ but that this does not hold in general is seen from the counterexample 3.17 of [4].

Often in the applications we assume that B has finite rank and in this case for $\langle A|\operatorname{Im} B\rangle$ to have meaning we shall require that $\operatorname{Im} B \subset D(A^\infty)$. If $\operatorname{Im} B \subset D(A^\infty)$, then we have the inclusion

$$\langle A|\operatorname{Im} B\rangle \subset \langle T(t)|\operatorname{Im} B\rangle. \tag{4.15}$$

5. Duality between (A, B)- and (C, A)-invariance. In finite dimensions there is a natural duality between (C, A)- and (A, B)-invariance concepts, which has been exploited by various researchers [7], [10]. In infinite dimensions the duality is not as sharp because of problems with the domain of the unbounded system operator. We need the following result, whose proof is straightforward and hence omitted.

LEMMA 5.1. *Suppose that X and Y are Hilbert spaces, $C \in \mathcal{L}(X, Y)$, V and S are closed linear subspaces of X, $T \in \mathcal{L}(X)$ and A is a closed, densely defined operator on X; then*

(a) $(S \cap V)^\perp = \overline{S^\perp + V^\perp}$,

(b) $(\operatorname{Ker} C)^\perp = \operatorname{Im} C^*$,

(c) $TS \subset V \Leftrightarrow T^* V^\perp \subset S^\perp$,

(d) $A(S \cap D(A)) \subset V \Rightarrow A^*(V^\perp \cap D(A^*)) \subset (S \cap D(A))^\perp$.

We have the following duality relationship between (A, B)- and (C, A)-invariance.

LEMMA 5.2.

(a) *V is $T(A, B)$-invariant $\Leftrightarrow$ $V^\perp$ is $T(B^*, A^*)$-invariant.*

(b) *If $V \subset D(A)$ is feedback (A, B)-invariant, then $V^\perp$ is feedback (B^*, A^*)-invariant.*

(c) *If $S \subset D(A)$ is feedback (C, A)-invariant, then $S^\perp$ is feedback-(A^*, C^*)-invariant.*

(d) *If $S \subset D(A)$ is (C, A)-invariant, then $S^\perp$ is (A^*, C^*)-invariant.*

(e) *If $V \subset D(A)$ is (A, B)-invariant and $\operatorname{Im} B$ and $V + \operatorname{Im} B$ are closed, then $V^\perp$ is (B^*, A^*)-invariant.*

Proof.

(a) This follows from (c) of Lemma 5.1 since $T_{A+BF}(t)$ is bounded and $A^* + F^* B^*$ is the generator of $(T_{A+BF}(t))^*$ in a Hilbert space.

(b) If V is feedback (A, B)-invariant, then by (e) Lemma 5.1 we have that

$$(A + BF)^*(V^\perp \cap D(A^*)) \subset (V \cap D(A))^\perp$$
$$\subset V^\perp \text{ under our assumption}$$

and $V^\perp$ is invariant under $A^* + F^* B^*$.

(c) Is proved similarly to (b).

(d) Applying (d) of Lemma 5.1, we see that if S is (C, A)-invariant, then

$$A^*(S^\perp \cap D(A^*)) \subset (S \cap D(A) \cap \operatorname{Ker} C)^\perp$$

and if $S \subset D(A)$, then by (a) of Lemma 5.1

$$A^*(S^\perp \cap D(A^*)) \subset \overline{S^\perp + \operatorname{Ker} C^\perp} = \overline{S^\perp + \overline{\operatorname{Im} C^*}} = \overline{S^\perp + \operatorname{Im} C^*}$$

by (b) of Lemma 5.1 and $S^\perp$ is (A^*, C^*)-invariant.

(e) Applying (d) of Lemma 5.1, we see that if V is (A, B)-invariant, then

$$A^*((V + Im\, B)^\perp \cap D(A^*)) \subset (V \cap D(A))^\perp = V^\perp \quad \text{if } V \subset D(A)$$

and

$$\begin{aligned} V + \operatorname{Im} B &= (V^\perp)^\perp + (\operatorname{Ker} B^*)^\perp \text{ since Im } B \text{ is closed} \\ &= (V^\perp \cap \operatorname{Ker} B^*)^\perp \text{ by Lemma 5.1(a),} \end{aligned}$$

since Im B and $V + \operatorname{Im} B$ are closed.

We see that there is a nice duality between $T(A, B)$- and $T(B^*, A^*)$-invariance, but not between (A, B) and (B^*, A^*)-invariance or between their feedback versions; one is forced to impose additional assumptions on the spaces and on B. This explains the lack of symmetry between Lemma 3.6 and Lemma 4.6. Of course for the special case that A is bounded one does have nice duality for all three invariance concepts. Finally we examine the duality between reachability and inobservability spaces.

LEMMA 5.3.

$$\langle \operatorname{Ker} C | T(t) \rangle = \langle T^*(t) | \operatorname{Im} C^* \rangle^\perp.$$

Proof. $x \in \langle T^*(t) | \operatorname{Im} C^* \rangle$ has the representation as the limit of terms like

$$x_n = \int_0^t T^*(t-s) C^* u_n(s)\, ds \quad \text{for some } u_n \in L_2(0, t; U).$$

So $y \in \langle T^*(t) | \operatorname{Im} C^* \rangle^\perp$ iff

$$\left\langle y, \int_0^t T^*(t \cdot s) C^* u(s)\, ds \right\rangle = 0 \quad \text{for all } u \in L_2(0, t; U)$$

iff

$$\int_0^t \langle CT(t-s)y, u(s) \rangle\, ds = 0 \quad \text{for all } u \in L_2(0, t; U)$$

and this holds iff $y \in \langle \operatorname{Ker} C | T(t) \rangle$.

6. Supremal (A, B)-invariant subspaces and infimal (C, A)-invariant subspaces. In finite-dimensions one can show that there always exists an infimal (C, A)-invariant subspace, which contains a given subspace. We address this question here for the infinite-dimensional case and introduce the following notation:

(6.1) $\underline{S}(C, A; \operatorname{Im} E)$ is the set of closed subspaces of X which are (C, A)-invariant and which contain Im E, where $E \in \mathscr{L}(Q, X)$ and Q is another Hilbert space.

We give conditions under which $\underline{S}(C, A; \operatorname{Im} E)$ possesses an infimal element, which we denote by $S^\infty(\operatorname{Im} E)$.

LEMMA 6.1. *Consider the following subspace algorithm*

$$S^n = \overline{\operatorname{Im} E + A(\operatorname{Ker} C \cap S^{n-1} \cap D(A))}, \qquad S_0 = \overline{\operatorname{Im} E}. \tag{6.2}$$

The algorithm (6.2) *is strictly increasing in n and converges in countably many steps to* S^∞. *If* S^∞ *is closed, then* S^∞ *is the smallest* (C, A)*-invariant subspace in* $\underline{S}(C, A; \operatorname{Im} E)$.

Proof.

(a) $S^n \supset S^{n-1}$ follows by a simple induction argument. So we have $X \supset S^n \supset S^{n-1} \supset \cdots \supset \operatorname{Im} E$ and since X is a separable Hilbert space, this must terminate in S^∞ after countably many steps. Since S^∞ is closed we have

$$\begin{aligned} S^\infty &= \overline{\operatorname{Im} E + A(\operatorname{Ker} C \cap S^\infty \cap D(A))} \\ &\supset \overline{\operatorname{Im} E} + \overline{A(\operatorname{Ker} C \cap S^\infty \cap D(A))} \\ &\supset A(\operatorname{Ker} C \cap S^\infty \cap D(A)). \end{aligned}$$

So S^∞ is (C, A)-invariant.

(b) Let V be any closed (C, A)-invariant subspace $\supset \operatorname{Im} E$. Then $V \supset S^n$ for all n by induction. For suppose that this holds for n, then

$$\begin{aligned} V &\supset \overline{\operatorname{Im} E + A(V \cap D(A) \cap \operatorname{Ker} C)} \\ &\supset \overline{\operatorname{Im} E + A(S^n \cap D(A) \cap \operatorname{Ker} C)} = S^{n+1}. \end{aligned}$$

In the limit $V \supset S^\infty$ and so S^∞ is the smallest element of $S(C, A; \operatorname{Im} E)$.

In general $S^\infty(\operatorname{Im} E)$ will not be $T(C, A)$-invariant and this property is essential in the applications. To understand this we consider the following special cases, where $\operatorname{Im} E = \operatorname{span}(e)$.

Case 1. $e \in D(A^{p+1})$; $CA^i e = 0$; $i = 0, \cdots, p-1$ and $CA^p e \neq 0$. Then we obtain $S^\infty = \operatorname{span}\{A^i e;\ i = 0, \cdots p\} \subset D(A)$ and from Corollary 3.7 we conclude that there exists a $G \in \mathscr{L}(Y, X)$ such that

$$(A + GC)S^\infty \subset S^\infty \quad \text{and} \quad S^\infty \text{ is } T_{A+GC}(t),$$

invariant, where $T_{A+GC}(t)$ is the semigroup generated by $A + GC$ given by (3.11).

Case 2. $e \in D(A^\infty) = \bigcap_{p=1}^\infty D(A^p)$ and $e \in \langle \operatorname{Ker} C | T(t) \rangle$. Then we obtain $S^\infty = \overline{\operatorname{span}}\{A^i e;\ i = 0, 1, \cdots, \infty\} = \overline{\langle A | e \rangle}$ and for all $G \in \mathscr{L}(Y, X)$, $(A + GC)(S^\infty \cap D(A)) = A(S^\infty \cap D(A)) \subset S^\infty$. The semigroup generated by $A + GC$ is given by (3.11) and so

$$T_{A+GC}(t)s = T(t)s \quad \text{for } s \in S^\infty.$$

Now S^∞ is not in general $T(t)$-invariant, but

$$S_1 = \langle T(t) | \operatorname{Im} e \rangle \text{ is} \quad \text{and} \quad T_{A+GC}|S_1 = T(t)|S_1$$

for all $G \in \mathscr{L}(Y, X)$. Thus S_1 is the infimal $T(C, A)$-invariant subspace containing e and it is invariant under $T_{A+GC}(t)$ for *all* $G \in \mathscr{L}(Y, X)$.

Case 3. $e \in \langle \operatorname{Ker} C | A \rangle$, but $e \notin \langle \operatorname{Ker} C | T(t) \rangle$. Again $S^\infty = \overline{\langle A | e \rangle}$ as in Case 2 and for all $G \in \mathscr{L}(Y, X)$, $(A + GC)(S^\infty \cap D(A)) \subset S^\infty$ but S^∞ need not be $T(C, A)$-invariant. Since $e \notin \langle \operatorname{Ker} C | T(t) \rangle$, we cannot proceed as in Case 2 to choose S_1 as S_1 will not be $T_{A+GC}(t)$-invariant in general. The only remaining alternative is to assume that $\langle A | e \rangle$ is $T(t)$-invariant, for then from (3.11) we may conclude that $T_{A+GC}(t)S^\infty = T(t)S^\infty$. Consequently, S^∞ is $T_{A+GC}(t)$-invariant for all $G \in \mathscr{L}(X, Y)$.

From these three special cases it is clear that the key to the general case is to consider the following partition of $\operatorname{Im} E \subset D(A^\infty)$. We shall use the following notation

$$\begin{aligned} &\mathscr{E}_1 = \overline{\operatorname{Im} E} \cap \langle \operatorname{Ker} C | T(t) \rangle;\ \mathscr{E}_2 = \mathscr{E}_1^\perp \cap \overline{\langle \operatorname{Ker} C | A \rangle \cap \operatorname{Im} E}, \\ &\mathscr{E}_3 = \overline{\operatorname{Im} E} \cap \overline{(\operatorname{Im} E \cap \langle \operatorname{Ker} C | A \rangle)}^\perp. \end{aligned} \tag{6.3}$$

We now state sufficient conditions for the existence of an infimal $T(C, A)$-invariant subspace containing $\operatorname{Im} E$.

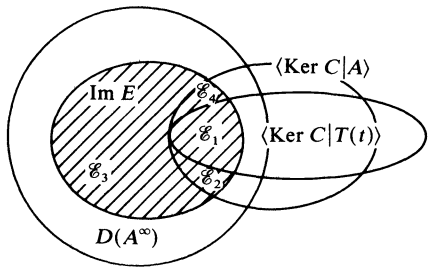


FIG. 6.1

LEMMA 6.2. *Suppose that E has finite rank,* Im $E \subset D(A^\infty)$,

$$\overline{\langle A|e\rangle} = \langle T(t)|e\rangle \quad \text{for } e \in \mathscr{E}_2 \tag{6.4}$$

and the algorithm (6.2) *with* $\mathscr{E}_3$ *in place of* Im E *terminates in finitely many steps, then there exists an infimal* $T(C, A)$*-invariant subspace containing* Im E, *which we denote by* $S^*(\text{Im } E)$, *and*

$$\overline{S^*(\text{Im } E) \cap D(A)} = S^*(\text{Im } E). \tag{6.5}$$

Proof.

(a) Since the algorithm (6.2) with $\mathscr{E}_3$ in place of Im E terminates in finite many steps, we obtain a finite-dimensional (C, A)-invariant subspace, S_f, which is therefore closed and is in the domain of A.

So by Corollary 3.7 there exists a $G \in \mathscr{L}(Y, X)$ such that S_f is $A+GC$ and $T_{A+GC}(t)$-invariant.

(b) For $e \in \mathscr{E}_2$ we are in Case 3 and writing $S_2 = \overline{\langle A|\mathscr{E}_2\rangle} = \langle T(t)|\mathscr{E}_2\rangle$, we see that $(A+GC)(S_2 \cap D(A)) = A(S_2 \cap D(A)) \subset S_2$ and $T_{A+GC}(t)S_2 = T(t)S_2 \subset S_2$. Thus S_2 is $T_{A+GC}(t)$-invariant.

(c) For $e \in \mathscr{E}_1$, $\overline{\langle A|e\rangle}$ may not necessarily equal $\langle T(t)|e\rangle$, but in any case, we choose $S_1 = \langle T(t)|\mathscr{E}_1\rangle$. As in Case 2, S_1 is $(A+GC)$-invariant and $T_{A+GC}(t)S_1 = T(t)S_1 \subset S_1$ and so S_1 is $T_{A+GC}(t)$-invariant. Choosing now

$$S^*(\text{Im } E) = S_f \oplus S_2 \oplus S_1 \tag{6.6}$$

we have a $T(C, A)$-invariant subspace containing Im E. An examination of the algorithm (6.2) and the special Cases 1-3 shows that the smallest (C, A)-invariant subspace containing Im E, $S^\infty(\text{Im } E)$ is given by

$$S^\infty(\text{Im } E) = S_f \oplus S_2 \oplus \overline{\langle A|\mathscr{E}_1\rangle}. \tag{6.7}$$

(d) *Now* $S^\infty(\text{Im } E) \subset S^*(\text{Im } E)$ and we wish to show that $S^*(\text{Im } E)$ is the smallest $T(C, A)$-invariant subspace. Suppose that S_0 is another $T(C, A)$-invariant subspace containing Im E. Since $S^\infty(\text{Im } E)$ is minimal, $S_0 \supset S^\infty(\text{Im } E)$ and

$$\begin{aligned} T_{A+G_0C}(t)S^\infty(\text{Im } E) &= T_{A+G_0C}(t)S_f \oplus T_{A+G_0C}(t)S_2 \oplus T_{A+G_0C}(t)\overline{\langle A|\mathscr{E}_1\rangle} \\ &= T_{A+G_0C}(t)S_f \oplus T(t)S_2 \oplus T(t)\overline{\langle A|\mathscr{E}_1\rangle}, \\ \bigcup_{t\geqq 0} T_{A+G_0C}(t)S^\infty(\text{Im } E) &= \bigcup_{t\geqq 0} T_{A+G_0C}(t)S_f \oplus S_2 \oplus S_1 \qquad \text{from (3.11)} \\ &\subset \bigcup_{t\geqq 0} T_{A+G_0C}(t)S_0 \subset S_0. \end{aligned}$$

But by (3.11) $\bigcup_{t\geqq 0} T_{A+G_0C}(t)S^*(\text{Im } E) = \bigcup_{t>0} T_{A+G_0C}(t)S^\infty(\text{Im } E) \subset S_0$ and in particular $S_0 \supset S^*(\text{Im } E)$ and $S^*(\text{Im } E)$ is indeed minimal.

(e) From (6.6) it follows that

$$S^*(\text{Im } E) = S_f \oplus \langle T(t) | \mathscr{E}_1 : \mathscr{E}_2 \rangle$$

and $S_f \subset D(A)$ and appealing to Lemma 4.9(e) proves (6.5).

In fact what we have shown in the proof is that $(S_f, S_1 \oplus S_2)$ reduces $A+GC$ in the sense of Definition 2.2.

We remark that Im $E \subset D(A^\infty)$ is stronger than is really needed; more precisely we need $\mathscr{E}_3 \subset D(A^{p+1})$, where $p = \dim S_f$ and $\mathscr{E}_1, \mathscr{E}_2 \subset D(A^\infty)$. This leads to the following useful corollary.

COROLLARY 6.3. *If E has finite rank, $\mathscr{E}_3 \subset D(A^{p+1})$, $\mathscr{E}_1 \subset D(A^\infty)$,*

$$\text{Im } E \cap \langle \text{Ker } C | T(t) \rangle = \text{Im } E \cap \langle \text{Ker } C | A \rangle \tag{6.8}$$

and the algorithm (6.2) *with $\mathscr{E}_3$ terminates in finitely many steps, then $S^*(\text{Im } E)$ exists and satisfies* (6.5).

In general $S^*(\text{Im } E)$ may not exist. This can be seen by examining Case 3 and (3.11): $\langle T_{A+GC}(t) | e \rangle$ depends on G. One could obtain a $T(C, A)$-invariant subspace containing $S^\infty(\text{Im } E)$ by choosing

$$S^\#(\text{Im } E) = S_f \oplus \langle T_{A+GC}(t) | \text{Im } E \cap \overline{\langle \text{Ker } C | A \rangle} \rangle$$

where S_f from the proof of Lemma 6.2 is $(A+GC)$-invariant. However, $S^\#(\text{Im } E)$ need not be infimal.

We now examine the dual problem of the existence of a supremal (A, B)-invariant subspace. We introduce the following notation.

(6.9) $\underline{V}(A, B; K)$ is the set of (A, B)-invariant subspaces contained in the closed subspace K.

Because of the lack of duality between (A, B) and (B^*, A^*)-invariance, we cannot dualize Lemma 6.1 to obtain the existence of a supremal element $V^\infty(K)$ of $\underline{V}(A, B; K)$. We consider the following algorithm

$$V^0 = K,\ V^n = K \cap (\lambda - A)^{-1}(V^{n-1} + \text{Im } B) \quad \text{for some } \lambda \in \rho(A). \tag{6.10}$$

LEMMA 6.4. *$\{V^n\}$ is strictly decreasing in n and has a limiting (A, B)-invariant subspace, $V^\infty(K)$, which is the largest (A, B)-invariant subspace contained in K, and is independent of λ; however, $V^\infty(K)$ need not be closed.*

Proof. Without loss of generality we can take $\lambda = 0$.

$V^n \supset V^{n+1}$ follows by induction. $V_1 \subset V_0 = K$ and suppose it holds for n. Then we have

$$V^n = K \cap A^{-1}(V^{n-1} + \text{Im } B) \supset K \cap A^{-1}(V^n + \text{Im } B) = V^{n+1}.$$

So $K \supset V_1 \supset V_2 \cdots \supset V^n \supset \{0\}$ and either the sequence terminates in finitely or countably many steps, since X is a separable Hilbert space. The limit V^∞ satisfies

$$V^\infty = K \cap A^{-1}(V^\infty + \text{Im } B). \tag{6.11}$$

So $V^\infty \subset K$ and $A(V^\infty \cap D(A)) \subset V^\infty + \text{Im } B$ and $V^\infty \subset D(A)$ and is (A, B)-invariant, but V^∞ need not be closed.

We show that V^∞ is the largest such subspace by supposing that W is another (A, B)-invariant subspace $\subset D(A)$

$$W \subset K \cap A^{-1}(W + \text{Im } B) \subset K = V_0.$$

If $W \subset V^n$, then from the above inclusion we obtain that $W \subset V^{n+1}$. So using induction

we have proved that $W \subset V^\infty$. So V^∞ is the largest (A, B)-invariant subspace $\supseteq D(A)$, but V^∞ need not be closed.

Unfortunately we cannot prove the existence of a closed supremal (A, B)-invariant subspace contained in K directly and the lack of duality between (B^*, A^*) and (A, B)-invariance means that we cannot appeal to Lemma 6.1. For the applications we also need $T(A, B)$-invariance and in this case we can appeal to the duality with $T(B^*, A^*)$-invariance implied by Lemma 5.2(a).

LEMMA 6.5. *Suppose that D has finite rank, $D \in L(X, R^k)$, say, and $\operatorname{Im} D^* \subset D(A^{*\infty}) = \bigcap_{n=1}^{\infty} D(A^{*n})$. Under the conditions that the algorithm* (6.2) *with* $\mathscr{E}_3 = \overline{(\langle \operatorname{Ker} B^* | A^* \rangle \cap \operatorname{Im} D^*)^\perp} \cap \operatorname{Im} D^*$ *terminates in finitely many steps and*

$$\overline{\langle A^* | d \rangle} = \langle T^*(t) | d \rangle \tag{6.12}$$

for all $d \in \operatorname{Im} D^* \cap \langle \operatorname{Ker} B^* | A^* \rangle$, but $d \perp \langle \operatorname{Ker} B^* | T^*(t) \rangle \cap \operatorname{Im} D^*$ there exists a supremal $T(A, B)$-invariant subspace contained in Ker D *which we denote by* $V^*(\operatorname{Ker} D)$.

Proof. We dualize Lemma 6.2, identifying

$$S^*(\operatorname{Im} D^*)^\perp = V^*(\operatorname{Ker} D). \tag{6.13}$$

An important special case in the applications is, in Lemma 6.2, when $\mathscr{E}_3$ has rank one, for then the algorithm (6.2) terminates in finitely many steps. This leads to a useful corollary for the existence of $V^*(\operatorname{Ker} D)$, where we make use of the fact that if $\operatorname{Im} D^* \subset D(A^{*\infty})$ and D has finite rank, then DA^i has a bounded extension for all i.

COROLLARY 6.6. *Suppose that D has finite rank, $D \in \mathscr{L}(X, R^k)$ and $DA^i \in \mathscr{L}(X; R^k)$ for all $i \geq 0$ and we partition D into three parts $D = (D_1 : D_2 : D_3)$, where*

$$\left.\begin{aligned} &D_1 T(t) B = 0 \quad \text{for all } t \geq 0; \\ &D_2 A^i B = 0,\ i \geq 0,\ D_2 T(t) B \neq 0,\ t \geq 0; \\ &D_3 = \langle \cdot, d_3 \rangle \quad \text{for some } d_3 \in X \text{ and} \\ &D_3 A^i B = 0,\ i = 0, \cdots p-1,\ D_3 A^p B \neq 0. \end{aligned}\right\} \tag{6.14}$$

Then there exists a supremal $T(A, B)$-invariant subspace contained in Ker D *if*

$$\langle T(t) | \operatorname{Im} D_2 \rangle = \overline{\langle A | \operatorname{Im} D_2 \rangle}. \tag{6.15}$$

Again we remark that we only need that $D_1 A^i$, $D_2 A^i \in \mathscr{L}(X, R^k)$ for all $i \geq 0$ and $D_3 A^i \in \mathscr{L}(X, R^k)$ for $i = 0, \cdots, p+1$.

7. Some disturbance decoupling problems. We consider the following disturbance decoupling problem (DDP)

$$\dot{x} = Ax + Bu + Eq, \tag{7.1}$$

$$z = Dx, \qquad u = Fx. \tag{7.2}$$

Where A, B are as before, q represents a disturbance, $q(\cdot) \in L_2(0, t; Q)$, $u(t) \in U$ represents the control and $z(t) \in Z$ represents the output to be decoupled. X, Q, Z and Y are real separable Hilbert spaces and $E \in \mathscr{L}(Q, X)$, $D \in \mathscr{L}(X, Z)$, $F \in \mathscr{L}(X, U)$. The DDP is to design a feedback control law $u = Fx$, such that

$$D \int_0^t T_{A+BF}(t-s) E q(s)\, ds = 0 \quad \text{for all } q \in L_2(0, t; Q). \tag{7.3}$$

From Definition 4.8 it is clear that (7.3) holds iff

$$\langle T_{A+BF}(t)|\operatorname{Im} E\rangle \subset \operatorname{Ker} D. \tag{7.4}$$

We now prove our first DDP theorem.

THEOREM 7.1. *If* $V^*(\ker D)$ *exists, then* DDP *is solvable iff*

$$V^*(\operatorname{Ker} D) \supset \operatorname{Im} E. \tag{7.5}$$

Proof.

(a) *Necessity.* Suppose that $F \in \mathscr{L}(X, U)$ is such that (7.4) holds. Now $\langle T_{A+BF}(t)|\operatorname{Im} E\rangle$ is $T(A, B)$-invariant and is contained in Ker D. Thus

$$V^*(\operatorname{Ker} D) \supset \langle T_{A+BF}(t)|\operatorname{Im} E\rangle \supset \operatorname{Im} E.$$

(b) *Sufficiency.* Suppose that (7.5) holds. Since $V^*(\operatorname{Ker} D)$ exists, there exists an $F \in \mathscr{L}(X, U)$ so that

$$(A+BF)(V^*(\operatorname{Ker} D) \cap D(A)) \subset V^*(\operatorname{Ker} D)$$

and by Lemma 4.9(d)

$$\langle T_{A+BF}(t)|V^*(\operatorname{Ker} D)\rangle = V^*(\operatorname{Ker} D).$$

Lemma 4.9 and (7.5) imply that

$$\langle T_{A+BF}(t)|\operatorname{Im} E\rangle \subset \rangle T_{A+BF}(t)|V^*(\operatorname{Ker} D)\rangle = V^*(\operatorname{Ker} D) \subset \operatorname{Ker} D$$

and (7.4) holds.

It is clear from the proof that the existence of $V^\infty(\operatorname{Ker} D)$ is not sufficient to solve DDP; one needs the $T(A, B)$-invariance. From Lemma 6.5 we see that for $V^*(\operatorname{Ker} D)$ to exist it is sufficient that B and D have finite rank (that is, finite-dimensional input and output) and D satisfy smoothness conditions. We do not need to impose any extra conditions on the disturbance other than (7.5).

One can also consider the disturbance decoupling problem when one allows a feedforward term in the control (DDPF) [10].

$$u = Fx + Rq, \qquad R \in \mathscr{L}(Q, U). \tag{7.6}$$

THEOREM 7.2. *If* $V^*(\operatorname{Ker} D)$ *exists, then* DDPF *has a solution iff*

$$V^*(\operatorname{Ker} D) + \operatorname{Im} B \supset \operatorname{Im} E. \tag{7.7}$$

Proof. From Theorem 7.1, DDPF has a solution iff there exists a $R \in \mathscr{L}(X, U)$, such that

$$\operatorname{Im}(BR+E) \subset V^*(\operatorname{Ker} D) \tag{7.8}$$

and clearly this implies (7.7).

Conversely, if (7.7) holds, then

$$V^*(\operatorname{Ker} D) \supset \operatorname{Im} B + \operatorname{Im} E \supset \operatorname{Im}(BR+E) \tag{7.9}$$

for any $R \in \mathscr{L}(Q, U)$ and in particular, we can always choose $R = 0$.

Now we consider the disturbance decoupled estimation problem DDEP (Willems and Commault [10]) for the system (7.1), (7.2), where we suppose that we cannot measure the state, but the following

$$y = Cx \tag{7.10}$$

where $C \in \mathscr{L}(X; Y)$ and the output space, Y, is a real, separable Hilbert space. The

DDEP is to construct a data processor for z on $W \cong Z$:

$$\begin{aligned} \dot{w} &= Fw + Hy + Ru, \\ \hat{z} &= Mw + Ny + Ju, \end{aligned} \tag{7.11}$$

where F is the infinitesimal generator of a strongly continuous semigroup on W and H, R, M, N and J are bounded linear operators and we require that the resulting estimation error $e = z - \hat{z}$ depends only on the initial conditions and not on q or u. For the DDEP we are unable to prove as sharp a result as for the DDP.

THEOREM 7.3. *If $S^*(\text{Im } E)$ exists and C has finite rank, then* DDEP *has a solution if*

$$S^*(\text{Im } E) \cap \text{Ker } C \subset \text{Ker } D \quad \textit{and} \quad J = 0 \tag{7.12}$$

and only if

$$S^{\infty}(\text{Im } E) \cap \text{Ker } C \subset \text{Ker } D \quad \textit{and} \quad J = 0. \tag{7.13}$$

Proof. (a) *Necessity.* Let $e = z - \hat{z}$ and $x^e = (x, w) \in X^e = X \oplus W$. Combining (7.1), (7.10, 11), we obtain the extended system

$$\begin{aligned} \Sigma^e: \dot{x}^e &= A^e x^e + E^e (u, q)^T, \\ y^e &= C^e x^e, \end{aligned} \tag{7.14}$$

where

$$A^e = \begin{pmatrix} A & 0 \\ HC & F \end{pmatrix}, \qquad E^e = \begin{pmatrix} B & E \\ R & O \end{pmatrix}$$

and the error is given by

$$e = D^e x^e + J^e (u, q) \tag{7.15}$$

where

$$D^e = (D - NC : -M); \qquad J^e = (-J, 0).$$

Solution of DDEP requires that $J^e = 0$ and

$$D^e \int_0^t T^e(t-s) E^e (u, q)(s)\, ds = 0 \quad \text{for } t \geqq 0 \tag{7.16}$$

where $T^e(t)$ is the strongly continuous semigroup generated by A^e.

Equivalently

$$\text{Im } E^e \subset \langle \text{Ker } D^e \mid T^e_{(t)} \rangle = S^e. \tag{7.17}$$

Defining

$$S_1 = \left\{ x \in X : \begin{pmatrix} x \\ 0 \end{pmatrix} \in S^e \right\}$$

it is easy to verify that S_1 is (C, A)-invariant. Taking intersections of (7.17) with X yields

$$\text{Im } E \subset S_1 \subset \text{Ker } (D - NC) \tag{7.18}$$

so $S_1 \in \underline{S}(C, A; \text{Im } E)$ and $S_1 \cap \text{Ker } C \subset \text{Ker } D$.

(b) *Sufficiency.* Suppose that $S \in \underline{S}(C, A; \operatorname{Im} E)$ and that S is $T(C, A)$-invariant with $S \cap \operatorname{Ker} C \subset \operatorname{Ker} D$. This holds for $S^*(\operatorname{Im} E)$, for example. So there exists an $L \in \mathscr{L}(Y, X)$ such that S is invariant under the semigroup, $T_L(t)$, generated by $A+LC$. We denote the generator of $T_L(t)$ by A_L.

We have the decomposition

$$X \cong X/S \oplus S \tag{7.19}$$

and since S is invariant with respect to A_L and $T_L(t)$ we have the block decompositions with respect to (7.19):

$$A_L = \begin{pmatrix} A_1 & 0 \\ A_2 & A_3 \end{pmatrix}, \qquad T_L(t) = \begin{pmatrix} T_1(t) & 0 \\ T_2(t) & T_3(t) \end{pmatrix}; \tag{7.20}$$

and the identities:

$$T_L(t)T_L(s) = T_L(t+s), \tag{7.21}$$

$$\frac{d}{dt}(T_L(t)x) = A_L T_L(t)x \quad \text{for } x \in D(A_L), \tag{7.22}$$

which yield

$$T_1(t)T_1(s)[x]_S = T_1(t+s)[x]_S \quad \text{for } [x]_S \in X/S, \tag{7.23}$$

$$\frac{d}{dt}(T_1(t)[x]_S) = A_1 T_1(t)[x]_S \quad \text{for } [x]_S \in D(A_1). \tag{7.24}$$

Thus, $T_1(t)$ defines a strongly continuous semigroup on X/S with generator A_1. We write these as $[T_L(t)]_S$ and $[A_L]_S$ and define the following observer on $W \cong X/S$

$$\begin{aligned} \dot{w} &= [A_L]_S w - [L]_S y + [B]_S u, \\ \hat{z} &= \bar{M} w + N y, \end{aligned} \tag{7.25}$$

where $[L]_S$ is defined by $[L]_S y = [Ly]_S$ and similarly for $[B]_S$ and we define $\bar{m}: W \to Z$ and $N: Y \to Z$ as follows. As in Lemma 3.6, we decompose $Y = Y_0 \oplus Y_0^{\perp}$, where $Y_0 = C(S \cap (S \cap \operatorname{Ker} C)^{\perp})$ is closed since C has finite rank. Then we define $N: Y \to Z$ by

$$\begin{aligned} Ny &= 0 \quad \text{if } y \in Y_0^{\perp}, \\ Ny &= D(C^{-1}y) \quad \text{if } y \in Y_0, \end{aligned} \tag{7.26}$$

and N is bounded, since D is bounded.

Note that $S \subset \operatorname{Ker}(D - NC)$ and so $\bar{M} \in \mathscr{L}(W, Z)$ is well defined by

$$\bar{M}[x]_S = (D - NC)[x]_S. \tag{7.27}$$

So (7.25) is a well defined observer and for $[x]_S \in D([A_L]_S)$, we have

$$\frac{d}{dt}[x]_S = [A_L]_S[x]_S - [L]_S y + [B]_S u \tag{7.28}$$

and

$$\dot{r} = [A_L]_S r \tag{7.29}$$

where $r=[x]_S-w$ and

$$\begin{aligned} e &= z-\hat{z} \\ &= Dx+\bar{M}w-NCx \\ &= [(D-NC)x]_S-\bar{M}w \qquad \text{by (7.26)} \\ &= \bar{M}([x]_S-w) \qquad \text{by (7.27)} \\ &= \bar{M}r \end{aligned}$$

and so (u, q) has no effect on the error e and the DDEP is solved.

We remark that the assumption that C have finite rank was only used to generate bounded operators $\bar{M}$ and N in the data processor in the sufficiency proof.

Sufficient conditions for $S^*(\text{Im } E)$ to exist are provided by Lemma 6.2, namely that C and E have finite rank and that E satisfy some smoothness assumptions. So for the DDEP we can only treat finite-dimensional disturbances, but infinite-dimensional outputs are allowed.

As remarked in [9], in some applications it may be desired that the observations be filtered before being used in $\hat{z}$, which means taking $N=J=0$. Then if $S^*(\text{Im } E)$ exists we can prove that the DDEP is solvable if

$$S^*(\text{Im } E)\subset \text{Ker } D. \tag{7.30}$$

8. Examples. From the results of § 7 we see that DDP (DDEP) can be solved if $V^*(\text{Ker } D)$ ($S^*(\text{Im } E)$) exists and Lemma 6.2 (Corollary 6.6) gives sufficient conditions for this. For distributed systems these conditions are easy to check and we illustrate this with a parabolic example.

Consider a heated rod which we suppose is heated around one point and due to some experimental setup is subject to disturbances in another region. We desire that the temperature at a certain measurement point be independent of the disturbances. The configuration can be schematized as below in Fig. 8.1. For the mathematical model we take

$$\frac{\partial x}{\partial t}=\frac{\partial^2 x}{\partial \xi^2}+b(\xi)u(t)+e(\xi)q(t), \tag{8.1}$$

$$x(0, t)=0=x(1, t), \tag{8.2}$$

$$z(t)=\int_0^1 d(\xi)x(t, \xi)\, d\xi \tag{8.3}$$

where we have chosen various shape-functions b, e and d to approximate our sensor and control actuators.

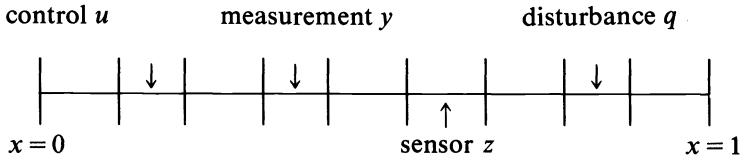


FIG. 8.1

This can be formulated as a system of the form (8.1), (8.2) on the Hilbert space $X = L_2(0, 1)$, where the system operator A is given by

$$A = \frac{d^2}{d\xi^2}; \qquad D(A) = \{h \in H^2 : h(0) = 0 = h(1)\}. \tag{8.4}$$

A is self adjoint and has eigenvalues $\{-n^2\pi^2; n = 1 \cdots \infty\}$ and eigenvectors $\{\phi_n(\xi) = \sqrt{2} \sin n\pi\xi; n = 1, 2, \cdots \infty\}$. The other operators are given by

$$B = b, \qquad E = e, \qquad D = \langle \cdot, d \rangle \tag{8.5}$$

where b, d and e are considered as elements of $L_2(0, 1)$. By Corollary 6.6, $V^*(\text{Ker } D)$ will exist if either of the following conditions hold

$$\langle T(t)b, d \rangle = 0, \qquad t \geqq 0, \tag{8.6}$$

$$\langle T(t) | \text{Im } D \rangle = \overline{\langle A | \text{Im } D \rangle}, \qquad \langle b, A^i d \rangle = 0, \; i \geqq 0, \tag{8.7}$$

$$\langle b, A^i d \rangle = 0, \quad i = 0, p-1, \quad \langle b, A^p d \rangle \neq 0. \tag{8.8}$$

Now if (A, b) is approximately controllable, (8.6) implies that $d = 0$, so (8.6) will not hold in general. Considering (8.7) we remark that $\langle T(t) | ImD \rangle = \overline{\langle A | \text{Im } D \rangle}$ will only hold if d is an eigenvector of A, so (8.7) will not hold in general. This leaves (8.8) as a possibility and the most likely situation is that $p = 0$. Then by Theorem 7.1, sufficient conditions for disturbance decoupling are

$$\langle b, d \rangle \neq 0, \quad \langle e, d \rangle = 0, \quad d \in D(A). \tag{8.9}$$

The feedback law is then given by

$$u = \langle \cdot ; f \rangle, \qquad f = \frac{\alpha d - Ad}{\langle b, d \rangle} \tag{8.10}$$

where α is an arbitrary constant.

Physically one can interpret the conditions (8.9) as requiring that the shape function d be smooth and confined to the interior of $(0, 1)$, that the shape functions e and d do not overlap, but that b and d do. One could give other sufficient conditions on b, d and e for the solvability of the disturbance decoupling problem, but they will not have such a natural physical interpretation.

If one is interested in disturbance decoupling and pole-assignment, then it is clear from (8.10) that at least two controls will be necessary.

Of course we cannot measure that full state $x(t, \cdot)$, but only a functional of it, for example

$$y(t) = \int_0^1 c(\xi) x(t, \xi) \, d\xi \tag{8.11}$$

and we can then consider the DDEP for (8.1)-(8.3), (8.11). From Theorem 7.3, we see that DDEP can be solved if $S^*(e)$ exists and

$$S^*(e) \cap \text{Ker } \langle \cdot, c \rangle \subset \text{Ker } \langle \cdot, d \rangle. \tag{8.12}$$

Generic conditions for the existence of $S^*(e)$ are

$$\langle c, e \rangle \neq 0; \; e \in D(A) \tag{8.13}$$

and then $S^*(e) = \text{span } \{e\}$ and (8.13) is sufficient for the solvability of DDEP; that is, we can construct an observer, as in the proof of Theorem 7.3, provided that the

measurements and disturbances are disjoint. We remark here that the observer is infinite-dimensional and that we have had to assume smooth and finite-dimensional disturbances to solve the DDEP. So the application of the DDEP is not quite as general as for the DDP.

We conclude with a simple example of a retarded system:

$$\dot{\xi}(t) = \alpha\xi(t) + \beta\xi(t-1) + \int_1^0 \gamma(\theta)\xi(t+\theta)\,d\theta + bu(t) + eq(t),$$
$$\xi(\theta) = \xi_0(\theta) \quad \text{on } [-1, 0], \tag{8.14}$$

$$z(t) = \delta\xi(t) + \int_{-1}^0 d(\theta)x(t+\theta)\,d\theta, \tag{8.15}$$

where α, β, b, e, δ are scalars and γ, ξ_0 and d are functions on $[-1, 0]$. Then (8.18) can be formulated as an abstract system on the product space $M^2 = R \times L^2(-1, 0)$, [9], where the system operator A on M^2 is defined by

$$A(h_0, h_1) = (\alpha h_1(0) + \beta h_1(-1) + \int_{-1}^0 \gamma(\theta)h_1(\theta)\,d\theta, Dh_1),$$
$$D(A) = W^{1,2}(-1, 0) = \{h \in L^2(-1, 0);\ Dh \in L^2(-1, 0)\}, \tag{8.16}$$

where Dh denotes the distributional derivative of h.

The remaining system operators in (7.1), (7.2) are

$$Bu = (bu, 0), \qquad Eq = (eq, 0),$$
$$Dh = \langle h, d^*\rangle_2 = \delta h_0 + \int_{-1}^0 d(\theta)h_1(\theta)\,d\theta, \tag{8.17}$$

where $\langle \cdot\,, \rangle_2$ denotes the inner product in M^2.

The conditions for DDP depend on the domain of A^*, which for retarded systems is very special.

$$D(A^*) = \{(h_0, h_1) \in M^2 | \exists w \in W^{1,2} \text{ with } w(-1) = 0 \text{ and } h_1(\theta) = w(\theta) + \beta h_0\}, \tag{8.18}$$

$$A^*(h_0, h_1) = ((\alpha + \beta)h_0 + \int_{-1}^0 Dw(\theta)\,d\theta,\ \gamma(\theta)h_0 - Dw(\theta)). \tag{8.19}$$

In our example, with one delay, $(h_0, h_1) \in D(A^*)$ implies that h_1 is in $W^{1,2}$, but in general it would be piecewise continuous [9]. The condition $d^* \in D(A^*)$ reduces to

$$d \in W^{1,2} \quad \text{and} \quad d(-1) = \beta\delta \tag{8.20}$$

and

$$A^*d^* = (\alpha\delta + d(0),\ \gamma(\theta)\delta - Dd(\theta)). \tag{8.21}$$

It is easy to see that $d^* \in D(A^{*p})$ implies that $D^p d \in W^{1,2}$ and that various conditions on $D^p d(-1)$ in terms of $D^{p-1}\gamma(-1)$ and $D^i d(0)$ be satisfied; $i = 0, \cdot, p-1$. Let us evaluate some terms DA^iV as in (6.13).

$$DB = \langle (b, 0)d^*\rangle_2 = b\delta, \tag{8.22}$$

$$DAB = \langle (b, 0), A^*d^*\rangle_2 = b\alpha\delta + bd(0), \tag{8.23}$$

$$DA^2B = \langle (b, 0), (A^*)^2d^*\rangle_2 = b(\alpha^2\delta + \delta d(0) + \gamma(0)\delta - Dd(0)). \tag{8.24}$$

If

$$DA^iB=0, \quad i=1\cdots p-1, \quad \text{and} \quad DA^pB\neq 0. \tag{8.25}$$

Then from Lemma 6.5 and Theorem 7.1 it follows that DDP is solvable iff

$$\langle (e,0), A^{*i}d^*\rangle_2=0, \qquad i=0,\cdot,p. \tag{8.26}$$

From (8.26) with $i=0$ it is clear that we must have $\delta=0$ and since $b, e\neq 0$ it is clear from (8.25) and (8.26) that DDP for system (8.14), (8.15) is *not* solvable. This was a very simple example, so now we suppose that the disturbances are of the form

$$eq(t)+\int_{-1}^{0}\delta(\theta)q(\theta)\,d\theta \tag{8.27}$$

where $\mu\in L_2(0,1)$. Then (8.26) is replaced by

$$\langle (e,\mu), A^{*i}d^*\rangle_2=0, \qquad i=0,\cdot,p \tag{8.28}$$

which yields for

$$i=0, \qquad e\delta+\int_{-1}^{0}\mu(\theta)\,d(\theta)\,d\theta=0, \tag{8.29}$$

$$i=1, \qquad e(\alpha\delta+d(0))+\int_{-1}^{0}\mu(\theta)(\delta\gamma(\theta)-Db(\theta))\,d\theta=0, \tag{8.30}$$

and so on. Examining the case $p=0$, $(b\delta\neq 0)$, we see that DDP for the disturbance (8.27) will be solvable iff (8.29) holds.

$$u(t)=\frac{\nu\delta-\alpha\delta-d(0)}{b\delta}x(t)+\frac{1}{b\delta}\int_{-1}^{0}x(t+\theta)[\nu d(\theta)-\delta\gamma(\theta)+Dd(\theta)]\,d\theta \tag{8.31}$$

is the decoupling control law, where ν is an arbitrary parameter.

For the case $p=1(\delta=0,\ d(0)\neq 0)$, DDP is solvable iff both (8.29) and (8.30) hold and these reduce to the following

$$\int_{-1}^{0}\mu(\theta)\,d(\theta)\,d\theta=0, \qquad ed(0)=\int_{-1}^{0}\mu(\theta)Dd(\theta)\,d\theta. \tag{8.32}$$

The decoupling feedback law in this case is

$$u(t)=\frac{1}{b}\left(\nu_2-\alpha+\frac{1}{d(0)}Dd(0)\right)x(t)+\frac{1}{bd(0)}\int_{-1}^{0}x(t+\theta)k(\theta)\,d\theta \tag{8.33}$$

where $k(\theta)=-d(\theta)\gamma(\theta)-D^2d(\theta)+\nu_1d(\theta)+\nu_2Dd(\theta)$ and ν_1, ν_2 are two arbitrary constants.

Acknowledgments. I would like to thank Hans Zwart for his careful reading of the original manuscript and for his valuable suggestions.

Note. After submitting this paper the author became aware of the related work by L. Pandolfi on DDP for retarded systems: L. Pandolfi, *Disturbance decoupling and invariant subspaces for delay systems*, Int. Report No. 7, 1984, Politechnico di Torino, Italy.

REFERENCES

[1] R. F. CURTAIN, *Spectral Systems*, Int. J. Control, 39 (1984), pp. 657-666.

[1a] R. F. CURTAIN, *Disturbance decoupling by measurement feedback with stability for infinite-dimensional systems*, Int. J. Control., to appear.

[2] ———, *(C, A, B)-pairs in infinite dimensions*, Systems Control Lett., 5 (1984), pp. 59-65.

[3] ———, *Disturbance decoupling for distributed systems by boundary control*, presented at the Conference on Control Theory for Distributed Parameter Systems and Applications, Vorau, Austria, July 1984.

[4] R. F. CURTAIN AND A. J. PRITCHARD, *Infinite dimensional linear systems theory*, Lecture Notes in Control and Information Sciences 8, Springer Verlag, Berlin, 1978.

[5] T. KATO, *Perturbation Theory for Linear Operators*, Springer Verlag, Berlin, 1966.

[6] E. J. P. GEORG SCHMIDT AND R. J. STERN, *Invariance theory for infinite dimensional linear control systems*, Appl. Math. Optim., 6 (1980), pp. 113-122.

[7] J. M. SCHUMACHER, *Dynamic Feedback in Finite and Infinite-Dimensional Linear Systems*, M. C. Tracts No. 143, Mathematisch Centrum, Amsterdam, 1982.

[8] A. E. TAYLOR AND D. C. LAY, *Introduction to Functional Analysis*, 2nd edition, John Wiley, New York, 1980.

[9] R. B. VINTER, *On the evolution of the state of linear differential delay equations in M^2, Properties of the generator*, J. Inst. Math. Appl., 21 (1978), pp. 13-23.

[10] J. C. WILLEMS AND C. COMMAULT, *Disturbance decoupling by measurement feedback with stability or pole placement*, this Journal, 19 (1981), pp. 490-504.

[11] W. M. WONHAM, *Linear Multivariable Control, a Geometric Approach*, 2nd edition, Springer-Verlag, New York, 1979.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 5, September 1986

008

DIFFUSIONS FOR GLOBAL OPTIMIZATION*

STUART GEMAN† AND CHII-RUEY HWANG‡

Abstract. We seek a global minimum of $U:[0,1]^n \to R$. The solution to $(d/dt)x_t = -\nabla U(x_t)$ will find local minima. The solution to $dx_t = -\nabla U(x_t)\,dt + \sqrt{2T}\,dw_t$, where w is standard (n-dimensional) Brownian motion and the boundaries are reflecting, will concentrate near the global minima of U, at least when "temperature" T is small: the equilibrium distribution for x_t is Gibbs with density $\pi_T(x)\alpha \exp\{-U(x)/T\}$. This suggests setting $T = T(t)\downarrow 0$ to find the global minima of U. We give conditions on $U(x)$ and $T(t)$ such that the solution to $dx_t = -\nabla U(x_t)\,dt + \sqrt{2T}\,dw_t$ converges weakly to a distribution concentrated on the global minima of U.

Key words. global optimization, simulated annealing, diffusion, reflecting boundaries

AMS(MOS) subject classifications. 60J60, 60J70

1. Introduction.

We can find a local minimum of a function U on R^n by starting at an arbitrary $x_0 \in R^n$ and solving the equation

$$\frac{dx_t}{dt} = -\nabla U(x_t).$$

A continuous path, x, seeking a *global* minimum will in general be forced to "climb hills" as well as follow down-hill gradients. One way of introducing hill-climbing, while preserving the tendency to descend along gradients, is to introduce random fluctuations into the path of x:

$$dx_t = -\nabla U(x_t)\,dt + \sqrt{2T}\,dw_t \tag{1.1}$$

where w is a standard Brownian motion and T, the "temperature," controls the magnitude of the random fluctuations. Under suitable conditions on U, x_t approaches (weakly) an equilibrium, which is a *Gibbs distribution* with density

$$\pi_T(x) = \frac{1}{Z_T}\exp\{-U(x)/T\} \quad \text{where } Z_T = \int_{R^n} \exp\{-U(x)/T\}\,dx.$$

As $T \to 0$, π_T concentrates on the global minima of U. Hence, in low temperature equilibrium we can expect to find x_t near a global minimum.

Unfortunately, the time required to approach equilibrium increases exponentially with $1/T$; solutions to (1.1) with small T will be very slow to find the important minima of U. This suggests that (1.1) be integrated with a gradually decreasing temperature, $T = T(t)\downarrow 0$. The hope is that the early and large random fluctuations will allow x_t to quickly escape from local minima, whereas the later (large t) behavior will be essentially a gradient descent into a prominent minimum of U.

The theorem presented here gives sufficient conditions on U and $T(t)$ for the weak convergence of x_t to a measure concentrating on the global minimum of U. We have simplified the mathematics by confining x to a rectangle in R^n (the diffusion is "reflected at the boundaries"). The rectangle is taken for convenience to be the unit

* Received by the editors March 4, 1985, and in revised form July 17, 1985.

† Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. This work was supported in part by the National Science Foundation under grants DMS-8352087 and DMS-8306507, and the U.S. Army Research Office under grant DAAG29-83-K-0116.

‡ Institute of Mathematics, Academia Sinica, Taipei, Taiwan, Republic of China.

cube. For illustration, let us assume that $U:[0,1]^n \to R$ has a unique global minimum at $x=\xi$. If U is sufficiently smooth, and properly-behaved at the boundaries (see § 2), and if $T(t)=c/\log(2+t)$ for c sufficiently large, then the solution to (1.1), with $T=T(t)$, converges to ξ:

$$P(|x_t-\xi|<\varepsilon)\to 1$$

for all $\varepsilon>0$ and all starting points.[1]

Our work was inspired by the "simulated annealing" recently proposed by Černý [2] and Kirkpatrick et al. [10]. Given a function U of n binary variables $x_1,\cdots,x_n$ they propose to find global minima of U by running the "Metropolis algorithm" [13] while gradually lowering the temperature. The Metropolis algorithm produces a Markov process with state space $\{0,1\}^n$. As in (1.1), there is a "temperature", T, and at fixed T the Metropolis algorithm also has the Gibbs distribution as equilibrium. The same heuristics, then, motivate gradually lowering $T=T(t)$. This is called simulated annealing since it copies the physical procedure, called annealing, of melting and then slowly cooling a physical substance (such as a crystal) in search of a low energy configuration. The latter typically corresponds to a high degree of spatial regularity, useful for some applications. Černý and Kirkpatrick apply their simulated annealing to certain combinatorial optimization problems, often with striking success.

Simulated annealing has also played a role in overcoming some of the computational problems that arise in image processing (Geman and Geman [4], Grenander [8], Marroquin [12]). In these applications, the procedure is modified to accommodate arbitrary discrete variables $x_1,\cdots,x_n$ with finite state spaces (rather than binary), and Geman and Geman have established weak convergence to the global minima of U, provided again that the temperature is lowered sufficiently slowly. Unfortunately, the extension of the Metropolis algorithm to *continuous* variables, $x_1,\cdots,x_n$, involves some awkward computational problems. Nevertheless, many of the variables that arise in image processing are most naturally modelled as continuous, such as pixel grey levels, line orientations, and the sizes and orientations of objects. This motivated both Grenander (in [8]) and us to look at a diffusion-process alternative. In future image processing experiments, we will be comparing the computational performance of the continuous-valued Metropolis scheme to the diffusion scheme presented here.

Some encouraging simulation results have been recently obtained by Aluffi-Pentini, Parisi, and Zirilli [1]. They study the performance of a modified version of (1.1), which includes repeated runs, and an interactive "annealing schedule" $T=T(t)$. The experiments involve 22 different test functions U. These are defined on R^n, with n ranging from one to fourteen, and have multiple local minima. Properly tuned, the algorithm finds a global minimum for each test function.

2. Statement of result. Given a real-valued function U on the unit cube

$$U:[0,1]^n \to R,$$

and an "annealing schedule" $T(t)\downarrow 0$, we define a diffusion x:

$$dx_t=-\nabla U(x_t)\,dt+\sqrt{2T(t)}\,dw_t$$

[1] B. Gidas [6] and H. Kushner [11] have recently improved on our result. Gidas gets a tight characterization of the minimum allowed c in the schedule $T(t)=c/\log(2+t)$, and removes the reflecting boundaries. Kushner generalizes to a richer class of diffusions, allowing state-dependent diffusion coefficients and a random drift. The latter makes the connection to "stochastic approximation" in which U, or its functionals, cannot be directly observed.

where $w_t \in R^n$ is standard Brownian motion. x is confined to $[0, 1]^n$ "by reflection," which will be made precise shortly. The theorem gives conditions on $T(t)$ and U which insure the convergence of x_t to the set of global minima of U, in a suitable (weak) sense. Conditions on $T(t)$ will be given later. As for U, the conditions include:

(A) There exists an extension of U to an open set $S \supseteq [0, 1]^n$, which is twice continuously differentiable, and whose gradient has zero normal component at all noncorner boundary elements of $[0, 1]^n$.[2]

There are many equivalent ways to make precise the notion of a reflected diffusion. We will proceed in a manner that best fits with the methods to be used later in the proof of the theorem. First, we extend U "periodically" to $\hat{U}$, defined on all of R^n. Let Z denote the integers, and for every $(i_1, \cdots, i_n) \in Z^n$ define

$$S_{i_1,\cdots,i_n} = \prod_{k=1}^{n} [i_k, i_k + 1]$$

and define $G_{i_1,\cdots,i_n}: [0, 1]^n \to S_{i_1,\cdots,i_n}$ by

$$(G_{i_1,\cdots,i_n}(x))_k = \begin{cases} i_k + x, & i_k \text{ even}, \\ i_k + 1 - x, & i_k \text{ odd}. \end{cases}$$

Finally, define $\hat{U}: R^n \to R$ by

$$x \in S_{i_1,\cdots,i_n} \Rightarrow \hat{U}(s) = U(G^{-1}_{i_1,\cdots,i_n}(x)).$$

If x is "on a boundary" (i.e. $x_k = l$ some $l \in Z$, $1 \leqq k \leqq n$), then x is an element of two or more cubes: for example $x \in S_{i_1,\cdots,i_n}$ and $x \in S_{j_1,\cdots,j_n}$ where $(i_1, \cdots, i_n) \neq (j_1, \cdots, j_n)$. But then

$$G^{-1}_{i_1,\cdots,i_n}(x) = G^{-1}_{j_1,\cdots,j_n}(x),$$

and hence $\hat{U}$ is well-defined.

The definition of the reflected process, x, is in terms of a "free" (ordinary diffusion) process $\hat{x}$:

$$d\hat{x}_t = -\nabla \hat{U}(\hat{x}_t)\, dt + \sqrt{2T(t)}\, dw_t.$$

Fix $t \geqq s \geqq 0$ and $x \in [0, 1]^n$. The conditional distribution on x_t given $x_s = x$ is the same as if we had set $\hat{x}_s = x$ and then defined $x_t = G^{-1}_{i_1,\cdots,i_n}(\hat{x}_t)$ whenever $\hat{x}_t \in S_{i_1,\cdots,i_n}$. In other words, we reflect $\hat{x}_t$ at the boundaries of the unit cube. More precisely, let $\hat{p}(s, x, t, y)$ be transition probability densities for the process $\hat{x}$ (density on x_t evaluated at y, given that $x_s = x$). Then x is the Markov process with the following transition probability densities:

$$p(s, x, t, y) = \sum_{i_1,\cdots,i_n} \hat{p}(s, x, t, G_{i_1,\cdots,i_n}(y))$$

$\forall x, y \in [0, 1]^n$, $t > s \geqq 0$. To check that these actually satisfy the Markov property, first observe that for any $(i_1, \cdots, i_n)$, $(j_1, \cdots, j_n) \in Z^n$, $x, y \in [0, 1]^n$, and $t > s \geqq 0$,

$$\hat{p}(s, G_{j_1,\cdots,j_n}(x), t, G_{i_1,\cdots,i_n}(y)) = \hat{p}(s, x, t, G_{k_1,\cdots,k_n}(y))$$

[2] We believe, but are not certain, that the theorem still holds when the normal component of ∇U does not vanish on the boundaries of $[0, 1]^n$.

where

$$k_p = \begin{cases} i_p - j_p & \text{if } j_p \text{ is even,} \\ j_p - i_p & \text{if } j_p \text{ is odd,} \end{cases}$$

for each $1 \leqq p \leqq n$. Thus, for any $x, y \in [0,1]^n$, $t > s \geqq 0$, and $t_0 \in (s, t)$,

$$\begin{aligned} p(s,x,t,y) &= \sum_{i_1,\cdots,i_n} \hat{p}(s,x,t,G_{i_1,\cdots,i_n}(y)) \\ &= \sum_{i_1,\cdots,i_n} \int_{R^n} \hat{p}(s,x,t_0,z)\hat{p}(t_0,z,t,G_{i_1,\cdots,i_n}(y))\,dz \\ &= \sum_{i_1,\cdots,i_n} \sum_{j_1,\cdots,j_n} \int_{[0,1]^n} \hat{p}(s,x,t_0,G_{j_1,\cdots,j_n}(z))\hat{p}(t_0,G_{j_1,\cdots,j_n}(z),t,G_{i_1,\cdots,i_n}(y))\,dz \\ &= \sum_{j_1,\cdots,j_n} \int_{[0,1]^n} \hat{p}(s,x,t_0,G_{j_1,\cdots,j_n}(z)) \sum_{i_1,\cdots,i_n} \hat{p}(t_0,G_{j_1,\cdots,j_n}(z),t,G_{i_1,\cdots,i_n}(y))\,dz \\ &= \sum_{j_1,\cdots,j_n} \int_{[0,1]^n} \hat{p}(s,x,t_0,G_{j_1,\cdots,j_n}(z)) \sum_{k_1,\cdots,k_n} \hat{p}(t_0,z,t,G_{k_1,\cdots,k_n}(y))\,dz \\ &= \sum_{j_1,\cdots,j_n} \int_{[0,1]^n} \hat{p}(s,x,t_0,G_{j_1,\cdots,j_n}(z))p(t_0,z,t,y)\,dz \\ &= \int_{[0,1]^n} p(s,x,t_0)p(t_0,z,t,y)\,dz. \end{aligned}$$

If the temperature were constant, then x would have a unique equilibrium distribution (as will be clear from the proof of the theorem):

$$\pi_T(B) = \int_{B \cap [0,1]^n} \frac{1}{Z_T} \exp\{-U(x)/T\}\,dx \tag{2.1}$$

where

$$Z_T = \int_{[0,1]^n} \exp\{-U(x)/T\}\,dx.$$

As $T \to 0$, π_T concentrates on the global minima of U. In fact, for well-behaved functions U, $\{\pi_T\}_{T>0}$ has a unique weak limit, call it π_0, and this satisfies

$$\pi_0(\{x\colon U(x) = \inf_y U(y)\}) = 1$$

(see Hwang [9]). If, for example, the global minimum of U is attained at a finite number of points, then $\pi_T \xrightarrow{w} \pi_0$, where π_0 concentrates on the global minima and has a simple characterization in terms of the Hessian of U. On the other hand, if the global minima of U form a set of positive Lebesgue measure, then again $\pi_T \xrightarrow{w} \pi_0$, but the latter is *uniform* on the set of global minima. In any case, we will assume the existence of a unique weak limit π_0:

(B) There exists π_0 such that $\pi_T \xrightarrow{w} \pi_0$ as $T \to 0$.

Most commonly, U will possess only one global minimum, in which case (B) is trivial. Of course, π_0 necessarily concentrates on the global minima.

The following theorem gives conditions for the weak convergence of x_t to π_0.

THEOREM. *Assume* (A) *and* (B) *and that* $T(t) = c/\log(2+t)$. *For all c sufficiently large,*

$$P(x_t \in \cdot \,|x_0 = x) \xrightarrow{w} \pi_0(\cdot) \quad \forall x \in [0,1]^n.$$

Remarks. (1) Actually, the result holds if $T(t) \geqq c/\log(2+t)$, c sufficiently large, and if (1) $T(t) \downarrow 0$; (2) $T(t)$ is continuously differentiable; and (3)

$$\frac{dT(t)/dt}{T(t)^3} e^{2\Delta/T(t)} \to 0$$

where $\Delta = \sup_{x,y\in[0,1]^n}(U(x) - U(y))$. The proof is the same.

(2) Almost sure convergence to the set minimizing U is, in general, impossible, as can be demonstrated already with $n = 1$ and a very simple function U. The reason can be put loosely as follows. If $T(t) \downarrow 0$ sufficiently slowly to guarantee escape from local minima, then repeated escapes from global minima are also guaranteed (albeit with increasing rareness).

(3) For most problems, the constant c necessary to guarantee convergence to global minima will most likely be too large to be practical. But if the discrete case is any guide, then our image processing experiments suggest that significant improvement is obtained over greedy algorithms (such as zero-temperature gradient descent) with a constant far too small to invoke the theorem. (See Geman and Geman [4] for further discussion. Hajek (personal communication) and Gidas [5] have actually identified the needed constant for the discrete case.)

3. Proof of the theorem. For any $x \in [0,1]^n$, $t > s \geqq 0$, $f \in C[0,1]^n$, and μ a probability measure on $[0,1]^n$, we give the following definitions:

(i) $\pi^s = \pi_{T(s)}$, π_T as in (2.1).

Notice that π^s has a density for all $0 \leqq s < \infty$. We will use the same symbol, π^s, to denote this density:

(ii) $$\pi^s(x) = \frac{\exp\{-U(x)/T(s)\}}{\int_{[0,1]^n} \exp\{-U(x)/T(s)\}\,dx},$$

(iii) $$\mu(f) = \int_{[0,1]^n} f(x)\,\mu(dx),$$

(iv) $$p(s, x, t, f) = \int_{[0,1]^n} f(y) p(s, x, t, y)\,dy,$$

(v) $$p(s, \mu, t, f) = \int_{[0,1]^n} \int_{[0,1]^n} f(y) p(s, x, t, y) \mu(dx)\,dy.$$

The proof of the theorem is based upon the following two lemmas.

LEMMA 1. $\forall f \in C[0,1]^n$, $s \geqq 0$,

$$\lim_{t\to\infty} \sup_{w\in[0,1]^n} |p(s, v, t, f) - p(s, w, t, f)| = 0.$$

LEMMA 2. $\forall f \in C[0,1]^n$,

$$\lim_{s\to\infty} \overline{\lim_{t\to\infty}} |p(s, \pi^s, t, f) - \pi^t(f)| = 0.$$

Assuming the validity of these, we establish the theorem as follows: fixing $x \in [0,1]^n$ and $f \in C[0,1]^n$,

$$\begin{aligned}
&\overline{\lim_{t\to\infty}} |p(0,x,t,f) - \pi_0(f)| \\
&\quad \leqq \overline{\lim_{s\to\infty}}\, \overline{\lim_{t\to\infty}} |p(s, p(0,x,s,\cdot), t, f) - p(s, \pi^s, t, f)| \\
&\qquad + \overline{\lim_{s\to\infty}}\, \overline{\lim_{t\to\infty}} |p(s, \pi^s, t, f) - \pi^t(f)| + \overline{\lim_{s\to\infty}}\, \overline{\lim_{t\to\infty}} |\pi^t(f) - \pi_0(f)| \\
&\qquad\qquad \text{(by Lemma 2 and } \pi^t = \pi_{T(t)} \xrightarrow{w} \pi_0) \\
&\quad = \overline{\lim_{s\to\infty}}\, \overline{\lim_{t\to\infty}} |p(s, p(0,x,s,\cdot), t, f) - p(s, \pi^s, t, f)| \\
&\quad = \overline{\lim_{s\to\infty}}\, \overline{\lim_{t\to\infty}} \left| \int_w \int_z [p(0,x,s,z) - \pi^s(z)] p(s,z,t,w) f(w)\, dz\, dw \right| \\
&\quad = \overline{\lim_{s\to\infty}}\, \overline{\lim_{t\to\infty}} \left| \int_z [p(0,x,s,z) - \pi^s(z)] p(s,z,t,f)\, dz \right| \\
&\quad \leqq \overline{\lim_{s\to\infty}}\, \overline{\lim_{t\to\infty}} \sup_{v,w} |p(s,v,t,f) - p(s,w,t,f)| = 0
\end{aligned}$$

by Lemma 1.

Proof of Lemma 1. $\forall t \geqq 0$ let

$$\delta_t = \inf_{x,y} p(t, x, t+1, y).$$

Then

$$\begin{aligned}
&\overline{\lim_{t\to\infty}} \sup_{v,w} |p(s,v,t,f) - p(s,w,t,f)| \\
&\quad = \overline{\lim_{t\to\infty}} \sup_{v,w} \left| \int p(s,v,s+1,z) p(s+1,z,t,f)\, dz \right. \\
&\qquad\qquad \left. - \int p(s,w,s+1,z) p(s+1,z,t,f)\, dz \right| \\
&\quad = \overline{\lim_{t\to\infty}} \sup_{v,w} \left| \int (p(s,v,s+1,z) - \delta_s) p(s+1,z,t,f)\, dz \right. \\
&\qquad\qquad \left. - \int (p(s,w,s+1,z) - \delta_s) p(s+1,z,t,f)\, dz \right| \\
&\quad \leqq \overline{\lim_{t\to\infty}} \sup_{v,w} |(1-\delta_s) \sup_z p(s+1,z,t,f) - (1-\delta_s) \inf_z p(s+1,z,t,f)| \\
&\quad = \overline{\lim_{t\to\infty}} (1-\delta_s) \sup_{v,w} |p(s+1,v,t,f) - p(s+1,w,t,f)| \\
&\quad \cdots \\
&\quad \leqq \overline{\lim_{t\to\infty}} \left\{ \prod_{k=0}^{[t-s]-1} (1-\delta_{s+k}) \right\} \\
&\qquad \cdot \sup_{v,w} |p(s+[t-s], v, t, f) - p(s+[t-s], w, t, f)| \\
&\quad \leqq 2\|f\|_\infty \overline{\lim_{t\to\infty}} \prod_{k=0}^{[t-s]-1} (1-\delta_{s+k}) = 2\|f\|_\infty \prod_{k=0}^{\infty} (1-\delta_{s+k})
\end{aligned}$$

(where $[x]$ is the greatest integer not exceeding x). Hence, for the proof of Lemma 1 it is sufficient to show that

$$\sum_{k=0}^{\infty} \delta_{s+k} = \infty \quad \forall s \geqq 0.$$

Let $\hat{\delta}_t = \inf_{x,y\in[0,1]^n} \hat{p}(t, x, t+1, y)$. Notice that $\hat{\delta}_t \leqq \delta_t$ for all $t \geqq 0$. We will show that

$$\sum_{k=0}^{\infty} \hat{\delta}_{s+k} = \infty \quad \forall s \geqq 0.$$

Define $\mathcal{H} = \{f : [t, t+1] \to R^n, f \text{ continuous}\}$, and let P_x and Q_x be the probability measures on $\mathcal{H}$ induced by

$$dZ_u = -\nabla \hat{U}(Z_u)\, du + \sqrt{2T(u)}\, dw_u, \quad Z_t = x, \quad u \in [t, t+1],$$

and

$$dZ_u = \sqrt{2T(u)}\, dw_u, \quad Z_t = x, \quad u \in [t, t+1],$$

respectively. Then $P_x \ll Q_x$ and

$$\frac{dP_x}{dQ_x}(Z(\circ)) = \exp\left\{\int_t^{t+1} \frac{1}{2T(u)} \langle -\nabla \hat{U}(Z(u)), dZ(u)\rangle - \frac{1}{2}\int_t^{t+1} \frac{1}{2T(u)} |\nabla \hat{U}(Z(u))|^2\, du\right\} \tag{3.1}$$

(see Stroock and Varadhan [14]). We will bound the exponent on the right-hand side. Apply Ito's formula, for the zero drift equations (i.e. under Q_x):

$$\frac{1}{2T(u)} \langle -\nabla \hat{U}(Z(u)), dZ(u)\rangle = \frac{1}{2}\sum_{i=1}^{n} \hat{U}_{x_i x_i}(Z(u))\, du - \frac{1}{2T(u)}\, d\hat{U}(Z(u))$$

$$\Rightarrow \int_t^{t+1} \frac{1}{2T(u)} \langle -\nabla \hat{U}(Z(u)), dZ(u)\rangle = \frac{1}{2}\sum_{i=1}^{n} \int_t^{t+1} \hat{U}_{x_i x_i}(Z(u))\, du - \int_t^{t+1} \frac{1}{2T(u)}\, d\hat{U}(Z(u)).$$

Under the assumptions on U, $\sup_{z\in R^n} |\hat{U}_{x_i x_i}(z)| \leqq C_1 < \infty$ for some C_1, and consequently

$$\left|\frac{1}{2}\sum_{i=1}^{n} \int_t^{t+1} \hat{U}_{x_i x_i}(Z(u))\, du\right| \leqq \frac{nC_1}{2}.$$

Using again the assumptions on U, together with the monotonicity and smoothness of $T(t)$,

$$\left|\int_t^{t+1} \frac{1}{2T(u)}\, d\hat{U}(Z(u))\right| = \left|\frac{\hat{U}(Z(t+1))}{2T(t+1)} - \frac{\hat{U}(Z(t))}{2T(t)} - \int_t^{t+1} \hat{U}(Z(u))\, d\left(\frac{1}{2T(u)}\right)\right| \leqq \frac{C_2}{T(t+1)}.$$

Hence

$$\left|\int_t^{t+1} \frac{1}{2T(u)}\langle -\nabla \hat{U}(Z(u)),\, dZ(u)\rangle\right| \leqq \frac{nC_1}{2} + \frac{C_2}{T(t+1)}.$$

As for the other term in the exponent of (3.1), we easily get a bound $C_3/T(t+1)$. Therefore, for some constant C_4,

$$\frac{dP_x}{dQ_x}(Z(\cdot)) \geqq \exp\{-C_4/T(t+1)\}.$$

Consequently, for any $\varepsilon > 0$, $x, y \in R^n$,

$$P_x(|Z(t+1)-y|<\varepsilon) \geqq e^{-C_4/T(t+1)} Q_x(|Z(t+1)-y|<\varepsilon).$$

Under Q_x, $\{Z_i(t+1)\}_{i=1}^n$ are independent normal with

$$Z_i(t+1) \sim N\left(x_i, \int_t^{t+1} 2T(u)\, du\right).$$

Taking $x, y \in [0,1]^n$,

$$\begin{aligned} P_x(|Z(t+1)-y|<\varepsilon) &\geqq e^{-C_4/T(t+1)} \int_{|z-y|<\varepsilon} \frac{1}{(2\pi \int_t^{t+1} 2T(u)\, du)^{n/2}} \\ &\quad \cdot \exp\left(-|z-x|^2 \Big/ \left\{4\int_t^{t+1} T(u)\, du\right\}\right) dz \\ &\geqq C_5\, e^{-C_4/T(t+1)} \int_{|z-y|<\varepsilon} \exp\left(-(\sqrt{n}+\varepsilon)^2 \Big/ \left\{4\int_t^{t+1} T(u)\, du\right\}\right) dz. \end{aligned}$$

Finally, then,

$$\begin{aligned} \hat{\delta}_t &= \inf_{x,y\in[0,1]^n} \hat{p}(t, x, t+1, y) \\ &= C_6 \inf_{x,y\in[0,1]^n} \lim_{\varepsilon\to 0} \frac{1}{\varepsilon^n} P_x(|z(t+1)-y|<\varepsilon) \\ &\geqq e^{-C_7/T(t+1)} \end{aligned}$$

for a sufficiently large constant C_7. It now follows that the condition

$$\sum_{k=0}^{\infty} \hat{\delta}_{s+k} = \infty \quad \forall s \geqq 0$$

is satisfied for $T(t) \geqq c/\log(2+t)$, provided c is sufficiently large.

Proof of Lemma 2. For $t > s \geqq 0$ define

$$N(s,t) = \int \pi^t(x)\left(\frac{p(s, \pi^2, t, x)}{\pi^t(x)} - 1\right)^2 dx.$$

We will show that

$$\lim_{s\to\infty} \overline{\lim_{t\to\infty}}\, N(s,t) = 0. \tag{3.2}$$

From this, Lemma 2 is obtained as follows: For any $f \in [0,1]^n$

$$\begin{aligned}
&\overline{\lim_{s\to\infty}}\,\overline{\lim_{t\to\infty}} \left| p(s, \pi^s, t, f) - \pi^t(f) \right| \\
&\quad = \overline{\lim_{s\to\infty}}\,\overline{\lim_{t\to\infty}} \left| \int (p(s, \pi^s, t, x) - \pi^t(x)) f(x)\, dx \right| \\
&\quad \leqq \|f\|_\infty \overline{\lim_{s\to\infty}}\,\overline{\lim_{t\to\infty}} \int \left| p(s, \pi^s, t, x) - \pi^t(x) \right| dx \\
&\quad = \|f\|_\infty \overline{\lim_{s\to\infty}}\,\overline{\lim_{t\to\infty}} \int \pi^t(x) \left| \frac{p(s, \pi^s, t, x)}{\pi^t(x)} - 1 \right| dx \\
&\quad \leqq \|f\|_\infty \overline{\lim_{s\to\infty}}\,\overline{\lim_{t\to\infty}} \sqrt{\int \pi^t(x) \left(\frac{p(s, \pi^s, t, x)}{\pi^t(x)} - 1 \right)^2 dx} \\
&\quad = 0.
\end{aligned}$$

The proof of (3.2) rests upon the following lemma.

LEMMA 3. *Let* $\Delta = \sup_{x,y\in[0,1]^n} (U(x) - U(y))$. *For all* $t > s \geqq 0$

$$\frac{\partial}{\partial t} N(s, t) \leqq \Delta \left(\frac{d}{dt} \left(\frac{1}{T(t)} \right) \right) (1 + N(s, t)) - 2T(t)\, e^{-2\Delta/T(t)} N(s, t).$$

Accept, for now, Lemma 3. We have with $T(t) = c/\log(2+t)$,

$$\frac{\partial}{\partial t} N(s, t) \leqq \frac{\Delta}{c} \left(\frac{1}{2+t} \right) - \left\{ \frac{2c}{\log(2+t)} \left(\frac{1}{2+t} \right)^{2\Delta/c} - \frac{\Delta}{c} \left(\frac{1}{2+t} \right) \right\} N(s, t). \tag{3.3}$$

From this, and the observation that $\lim_{t\downarrow s} N(s, t) = 0$, (3.2) is easily established, provided that c is sufficiently large. We will forgo these details; they only involve integrating (3.3), with the inequality replaced by equality.

All that remains is the proof of Lemma 3.

Proof of Lemma 3. First, observe that

$$N(s, t) = \int \frac{p(s, \pi^s, t, x)^2}{\pi^t(x)}\, dx - 1.$$

Hence

$$\begin{aligned}
N_t(s, t) &= \int \left(\frac{d}{dt} \frac{1}{\pi^t(x)} \right) p(s, \pi^s, t, x)^2\, dx \\
&\quad + 2 \int p_t(s, \pi^s, t, x) p(s, \pi^s, t, x) \left(\frac{1}{\pi^t(x)} \right) dx \\
&= A(s, t) + B(s, t)
\end{aligned}$$

where

$$A(s, t) = \int \left(\frac{d}{dt} \frac{1}{\pi^t(x)} \right) p(s, \pi^s, t, x)^2\, dx$$

and

$$B(s, t) = 2 \int p_t(s, \pi^s, t, x) p(s, \pi^s, t, x) \left(\frac{1}{\pi^t(x)} \right) dx.$$

Let $g(t) = 1/T(t)$. Then

$$\begin{aligned} A(s,t) &= -\int \left(\frac{1}{\pi^t(x)}\right)^2 \left(\frac{d}{dt}\pi^t(x)\right) p(s,\pi^s,t,x)^2\,dx \\ &= -\int \left(\frac{1}{\pi^t(x)}\right)^2 (-g_t(t)U(x)\pi^t(x) + g_t(t)\pi^t(x)\pi^t(U)) p(s,\pi^s,t,x)^2\,dx \\ &= g_t(t)\int \frac{1}{\pi^t(x)}(U(x)-\pi^t(U)) p(s,\pi^s,t,x)^2\,dx \qquad (3.4) \\ &\leqq \Delta g_t(t)\int \frac{p(s,\pi^s,t,x)^2}{\pi^t(x)}\,dx \\ &= \Delta g_t(t)(1+N(s,t)). \end{aligned}$$

The treatment of $B(s,t)$ is more involved. The first step will be to show that

$$B(s,t) = -2T(t)\int |\nabla[p(s,\pi^s,t,x)/\pi^t(x)]|^2 \pi^t(x)\,dx.$$

From this, we will then derive the bound

$$B(s,t) \leqq -2T(t)\, e^{-2\Delta/T(t)} N(s,t),$$

which, together with (3.4), completes the proof.

We rewrite $B(s,t)$ with the help of the forward equation for the original ($\hat{}$) process: for $t > s \geqq 0$,

$$\hat{p}_t(s,y,t,x) = \sum_{k=1}^{n} \{T(t)\hat{p}_{x_kx_k}(s,y,t,x) + \hat{U}_{x_k}(x)\hat{p}_{x_k}(s,y,t,x) + \hat{U}_{x_kx_k}(x)\hat{p}(s,y,t,x)\}.$$

Integration over y, with respect to π^s, gives

$$\begin{aligned} \hat{p}_t(s,\pi^s,t,x) = \sum_{k=1}^{n} \{&T(t)\hat{p}_{x_kx_k}(s,\pi^s,t,x) \\ &+ \hat{U}_{x_k}(x)\hat{p}_{x_k}(s,\pi^s,t,x) + \hat{U}_{x_kx_k}(x)\hat{p}(s,\pi^s,t,x)\}. \end{aligned} \qquad (3.5)$$

We wish to convert (3.5) into a similar equation for p. This conversion is based upon the following identities, which are justified by the assumed smoothness of U, and the resulting smoothness of $\hat{p}$ (see, for example, [3]). For each integer i define

$$P(i) = \begin{cases} 1 & \text{if } i \text{ is even,} \\ -1 & \text{if } i \text{ is odd.} \end{cases}$$

Recalling that

$$x, y \in [0,1]^n \Rightarrow p(s,y,t,x) = \sum_{i_1,\cdots,i_n} \hat{p}(s,y,t,G_{i_1,\cdots,i_n}(x)),$$

we have, for each $1 \leqq k \leqq n$ and each $x, y \in [0,1]^n$:

$$\begin{aligned} &\hat{U}_{x_k}(G_{i_1,\cdots,i_n}(x)) = P(i_k)U_{x_k}(x), \\ &\hat{U}_{x_kx_k}(G_{i_1,\cdots,i_n}(x)) = U_{x_kx_k}(x), \\ &p(s,\pi^s,t,x) = \sum_{i_1,\cdots,i_n} \hat{p}(s,\pi^s,t,G_{i_1,\cdots,i_n}(x)), \end{aligned}$$

$$p_t(s, \pi^s, t, x) = \sum_{i_1,\cdots,i_n} \hat{p}_t(s, \pi^s, t, G_{i_1,\cdots,i_n}(x)),$$

$$p_{x_k}(s, \pi^s, t, x) = \sum_{i_1,\cdots,i_n} p(i_k)\hat{p}_{x_k}(s, \pi^s, t, G_{i_1,\cdots,i_n}(x)),$$

$$p_{x_kx_k}(s, \pi^s, t, x) = \sum_{i_1,\cdots,i_n} \hat{p}_{x_kx_k}(s, \pi^s, t, G_{i_1,\cdots,i_n}(x)).$$

Now take $x, y \in [0, 1]^n$ and replace x by $G_{i_1,\cdots,i_n}(s)$ in (3.5):

$$\begin{aligned}\hat{p}_t(s, \pi^s, t, G_{i_1,\cdots,i_n}(x)) = \sum_{k=1}^{n} \{&T(t)\hat{p}_{x_kx_k}(s, \pi^s, t, G_{i_1,\cdots,i_n}(x)) \\ &+ U_{x_k}(x)P(i_k)\hat{p}_{x_k}(s, \pi^s, t, G_{i_1,\cdots,i_n}(x)) \\ &+ U_{x_kx_k}(x)\hat{p}(s, \pi^s, t, G_{i_1,\cdots,i_n}(x))\}.\end{aligned}$$

Summation over $i_1, \cdots, i_n$ yields:

$$\begin{aligned}p_t(s, \pi^s, t, x) &= \sum_{k=1}^{n} \{T(t)p_{x_kx_k}(s, \pi^s, t, x) + U_{x_k}(x)p_{x_k}(s, \pi^s, t, x) + U_{x_kx_k}(x)p(s, \pi^s, t, x)\} \\ &= T(t) \sum_{k=1}^{n} \frac{\partial}{\partial x_k}\left\{\pi^t(x)\frac{\partial}{\partial x_k}[p(s, \pi^s, t, x)/\pi^t(x)]\right\}. \qquad (3.6)\end{aligned}$$

The associated boundary conditions are

$$p_{x_k}(s, \pi^s, t, x) = 0 \quad \text{whenever } x_k = 0 \text{ or } 1.$$

To see how these arise, take, for example, $x_k = 0$: letting $x = (x_1, x_2, \cdots, x_{k-1}, 0, x_{k+1}, \cdots, x_n)$, and letting e_k be the unit vector along the kth coordinate,

$$\begin{aligned}&p_{x_k}(s, \pi^s, t, x) \\ &= \lim_{\varepsilon \downarrow 0} \sum_{i_1,\cdots,i_n} P(i_k)\hat{p}_{x_k}(s, \pi^s, t, G_{i_1,\cdots,i_n}(x_1, \cdots, x_{k-1}, \varepsilon, x_{k+1}, \cdots, x_n)) \\ &= \lim_{\varepsilon \downarrow 0} \sum_{i_j : j \neq k} \sum_{p=-\infty}^{\infty} \\ &\quad \cdot \{\hat{p}_{x_k}(s, \pi^s, t, G_{i_1,\cdots,i_n}(x_1, \cdots, x_{k-1}, 0, x_{k+1}, \cdots, x_n) + (2p + \varepsilon)e_k) \\ &\qquad - \hat{p}_{x_k}(s, \pi^s, t, G_{i_1,\cdots,i_n}(x_1, \cdots, x_{k-1}, 0, x_{k+1}, \cdots, x_n) + (2p - \varepsilon)e_k)\} \\ &= 0.\end{aligned}$$

Now combine the boundary conditions on p with our boundary assumptions on ∇U (set out in (A)):

$$\frac{\partial}{\partial x_k}[p(t, \pi^s, t, x)/\pi^t(x)] = 0$$

for all x such that $x_k = 0$ or 1. Multiplying the equation in (3.6) by $p(s, \pi^s, t, x)/\pi^t(x)$

and integrating x over $[0,1]^n$ gives:

$$\frac{1}{2}B(s,t) = \int p_t(s,\pi^s,t,x)p(s,\pi^s,t,x)\left(\frac{1}{\pi^t(x)}\right)dx$$

$$= T(t)\sum_{k=1}^{n}\int_{x_j:j\neq k}\left\{\int_{x_k}(p(s,\pi^s,t,x)/\pi^t(x))\right.$$

$$\left.\cdot\frac{\partial}{\partial x_k}\left\{\pi^t(x)\frac{\partial}{\partial x_k}[p(s,\pi^s,t,x)/\pi^t(x)]\right\}dx_k\right\}$$

$$\cdot\, dx_1\cdots dx_{k-1}\,dx_{k+1}\cdots dx_n$$

$$= (\text{integrating over } x_k \text{ by parts})$$

$$-T(t)\sum_{k=1}^{n}\int\left\{\frac{\partial}{\partial x_k}[p(s,\pi^s,t,x)/\pi^t(x)]\right\}^2\pi^t(x)\,dx$$

$$= -T(t)\int|\nabla[p(s,\pi^s,t,x)/\pi^t(x)]|^2\pi^t(x)\,dx.$$

It remains to show that

$$\int|\nabla[p(s,\pi^s,t,x)/\pi^t(x)]|^2\pi^t(x)\,dx \geqq e^{-2\Delta/T(t)}N(s,t). \tag{3.7}$$

This final step is a consequence of the following proposition.

PROPOSITION. *If $\theta:[0,1]^n\to R$ is continuously differentiable, and if*

$$\int\theta(x)\,dx = 0,$$

then

$$\int\theta(x)^2\,dx \leqq \int|\nabla\theta(x)|^2\,dx.$$

For a more general version of this, see Gilbarg and Trudinger [7, p. 157].

Finally, fix s and t, and let

$$\phi(x) = \frac{p(s,\pi^s,t,x)}{\pi^t(x)} - 1$$

and $\alpha = \int\phi(x)\,dx$. Since $\int\phi(x)\pi^t(x)\,dx = 0$,

$$N(s,t) = \int\phi(x)^2\pi^t(x)\,dx \leqq \int(\phi(x)-\alpha)^2\pi^t(x)\,dx.$$

Notice that $e^{-\Delta/T(t)} \leqq \pi^t(x) \leqq e^{\Delta/T(t)}$, and therefore

$$N(s,t) \leqq e^{\Delta/T(t)}\int(\phi(x)-\alpha)^2\,dx$$

$$\leqq e^{\Delta/T(t)}\int|\nabla\phi(x)|^2\,dx \quad (\text{by the proposition})$$

$$\leqq e^{2\Delta/T(t)}\int|\nabla\phi(x)|^2\pi^t(x)\,dx,$$

which is the same as (3.7).

REFERENCES

[1] F. ALUFFI-PENTINI, V. PARISI AND F. ZIRILLI, *Global optimization and stochastic differential equations*, J. Optim. Theory Applications, 1985, to appear.

[2] C. ČERNÝ, *A thermodynamical approach to the travelling salesman problem: an efficient simulation algorithm*, preprint, Institute of Physics and Biophysics, Comenius Univ., Bratislava, 1982.

[3] E. B. DYNKIN, *Markov Processes*—II, Springer-Verlag, New York, 1965.

[4] S. GEMAN AND D. GEMAN, *Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images*, IEEE-PAMI, 6 (1984), pp. 721-741.

[5] B. GIDAS, *Non-stationary Markov chains and convergence of the annealing algorithm*, J. Statistical Physics, 39 (1985), pp. 73-131.

[6] ———, *The Langevin equation as a global minimization algorithm*, to appear in Disordered Systems and Biological Organizations, E. Bienenstock, F. Flogeman and G. Weisbuch, eds., Springer-Verlag, Berlin.

[7] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second-Order*, Springer-Verlag, Berlin, 1977.

[8] U. GRENENDER, *Tutorial in Pattern Theory*, Div. Applied Mathematics, Brown Univ., Providence, RI, 1984.

[9] C.-R. HWANG, *Laplace's method revisited: weak convergence of probability measures*, Ann. Probab., 8 (1980), pp. 1177-1182.

[10] S. KIRKPATRICK, C. D. GELATT, JR. AND M. P. VECCHI, *Optimization by simulated annealing*, IBM Thomas J. Watson Research Center, Yorktown Heights, NY, 1982.

[11] H. J. KUSHNER, *Asymptotic global behavior for stochastic approximations and diffusions with slowly decreasing noise effects: global minimization via Monte Carlo*, Lefschetz Center for Dynamical Systems report 85-7, Div. Applied Mathematics, Brown Univ., Providence, RI, 1985.

[12] J. L. MARROQUIN, *Surface reconstruction preserving discontinuities*, preprint, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA, 1984.

[13] M. METROPOLIS, A. W. ROSENBLUTH, M. N. ROSENBLUTH, A. H. TELLER AND E. TELLER, *Equations of state calculations by fast computing machines*, J. Chem. Phys., 21 (1953), pp. 1087-1091.

[14] D. W. STROOCK AND S. R. S. VARADHAN, *Multidimensional Diffusion Processes*, Springer-Verlag, New York, 1979.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 5, September 1986

009

NECESSARY AND SUFFICIENT CONDITIONS FOR ISOLATED LOCAL MINIMA OF NONSMOOTH FUNCTIONS*

MARCIN STUDNIARSKI†

Abstract. We consider the mathematical programming problem: find $\inf\{f(x)|x\in C\}$ where f is an arbitrary extended-real-valued function, and C a subset of a finite dimensional space. We give necessary and sufficient optimality conditions for this problem, generalizing previous results of A. Auslender (this Journal, 22 (1984), pp. 239-254).

Key words. nonsmooth optimization, higher order optimality conditions, isolated local minima, lower and upper Dini directional derivatives

1. Introduction. This research was inspired by the recent results of Auslender [2, § 2], who derived necessary and sufficient conditions for isolated local minima with order 1 and 2 for the mathematical programming problem

$$\text{(P)} \qquad \inf\{f(x)|x\in C\},$$

assuming that $f:\mathbb{R}^n\to\mathbb{R}$ is a locally Lipschitzian function, and C a closed subset of $\mathbb{R}^n$. Our aim is to demonstrate that those conditions can be reformulated and generalized so as to be valid for any extended-real-valued function $f:\mathbb{R}^n\to\bar{\mathbb{R}}$ and to encompass isolated local minima with order greater than two.

Throughout the paper we assume that C is a subset of $\mathbb{R}^n$ (not necessarily closed), $\bar{x}$ is a point of C and $f:\mathbb{R}^n\to\mathbb{R}$ is a function such that $|f(\bar{x})|<\infty$. We denote by $\mathcal{N}(x)$ the collection of all neighbourhoods of the point $x\in\mathbb{R}^n$.

Let k be any positive integer. Extending Definition 2.1 of [2], we shall say that $\bar{x}$ is an *isolated local minimum with order k of problem* (P) if there exist $\beta>0$ and $U\in\mathcal{N}(\bar{x})$ such that

$$f(x)>f(\bar{x})+\beta|x-\bar{x}|^k \quad \text{for all } x\in U\cap C,\quad x\neq\bar{x} \tag{1.1}$$

(here $|\cdot|$ stands for the Euclidean norm).

We shall state our optimality conditions for problem (P) by means of lower and upper Dini directional derivatives of f at $\bar{x}$, and of higher order counterparts of these notions. Before formulating the definitions, let us note the following.

Remark 1.1. For any function $\varphi:]0,+\infty[\times\mathbb{R}^n\to\bar{\mathbb{R}}$, the function

$$\psi(x)=\liminf_{\substack{t\downarrow 0\\ v\to x}}\varphi(t,v)=\sup_{V\in\mathcal{N}(x)}\inf_{\substack{\alpha>0\\ 0<t<\alpha\\ v\in V}}\varphi(t,v) \tag{1.2}$$

is lower semicontinuous on $\mathbb{R}^n$.

Let us now introduce the following notation (for each $h\in\mathbb{R}^n$ and $k=1,2,\cdots$):

$$\begin{aligned}\underline{d}^k f(\bar{x};h)&=\liminf_{\substack{t\downarrow 0\\ v\to h}} t^{-k}(f(\bar{x}+tv)-f(\bar{x})),\\ \bar{d}^k f(\bar{x};h)&=\limsup_{\substack{t\downarrow 0\\ v\to h}} t^{-k}(f(\bar{x}+tv)-f(\bar{x})).\end{aligned} \tag{1.3}$$

If $k=1$, we obtain the lower and the upper Dini directional derivatives [3], [4] (called also contingent derivatives [1], or semiderivatives [5]). One can refer to the above-

* Received by the editors December 12, 1984, and in revised form August 2, 1985.
† Institute of Mathematics, University of Łódź, 90-238 Łódź, Poland.

mentioned literature for more information on this subject as well as for further references. We shall write $\underline{d}f$ and $\bar{d}f$ instead of $\underline{d}^1 f$ and $\bar{d}^1 f$; this notation coincides with that used in [5].

Remark 1.2 [1, p. 286]. If f is locally Lipschitzian, then

$$\underline{d}f(\bar{x}; h) = \liminf_{t \downarrow 0} t^{-1}(f(\bar{x}+th) - f(\bar{x}));$$

of course, a similar formula is valid for $\bar{d}f$.

Let $\delta(\cdot|C)$ denote the indicator function of C:

$$\delta(x|C) = \begin{cases} 0 & \text{if } x \in C, \\ +\infty & \text{if } x \notin C, \end{cases}$$

and let $f_C = f + \delta(\cdot|C)$. Then

$$\underline{d}^k f_C(\bar{x}; h) = \liminf_{\substack{t \downarrow 0 \\ v \to h}} [t^{-k}(f(\bar{x}+tv) - f(\bar{x})) + \delta(\bar{x}+tv|C)]. \tag{1.4}$$

We now introduce lower kth order directional derivatives of f at $\bar{x}$ (in the direction $h \in \mathbb{R}^n$). Let us denote $h^k = (h, \cdots, h) \in (\mathbb{R}^n)^k$ and define

$$f'_-(\bar{x}; h) = \underline{d}f(\bar{x}; h),$$

$$f^{(k)}_-(\bar{x}; h^k) = k! \liminf_{\substack{t \downarrow 0 \\ v \to h}} t^{-k}[f(\bar{x}+tv) - f(\bar{x}) - \sum_{j=1}^{k-1} t^j f^{(j)}_-(\bar{x}; v^j)/j!], \tag{1.5}$$

$$k = 2, 3, \cdots.$$

It follows from Remark 1.1 that the functions $\underline{d}^k f(\bar{x}; \cdot)$, $\underline{d}^k f_C(\bar{x}; \cdot)$ and $h \mapsto f^{(k)}_-(\bar{x}; h^k)$ $(k = 1, 2, \cdots)$ are lower semicontinuous. Moreover, we have the following.

Remark 1.3. If $k > 1$ and

$$f^{(j)}_-(\bar{x}; h^j) \geqq 0 \quad \text{for all } h \in \mathbb{R}^n \quad \text{and } j = 1, \cdots, k-1, \tag{1.6}$$

then

$$f^{(k)}_-(\bar{x}; h^k) \leqq k! \underline{d}^k f(\bar{x}; h) \quad \text{for all } h \in \mathbb{R}^n. \tag{1.7}$$

If we assume (1.6) to be equalities, then also equality holds in (1.7).

Using Taylor's formula (in such a form as in [6, Thm. 21]), we easily obtain, by induction with respect to k, the following

PROPOSITION 1.1. *If f is $(k-1)$-times $(k > 1)$ Fréchet differentiable on $\mathbb{R}^n$ and if the kth derivative $f^{(k)}(\bar{x})$ of f at $\bar{x}$ exists, then*

$$f^{(k)}(\bar{x})h^k = f^{(k)}_-(\bar{x}; h^k) \quad \textit{for all } h \in \mathbb{R}^n.$$

2. Optimality conditions for problem (P). In this section we present different types of optimality conditions for (P). We first consider conditions which characterize isolated local minima with order k of (P) and are stated by means of limits (1.4). They are difficult to apply as they involve the indicator function of C. Therefore, we shall next give other conditions, stated in terms of limits (1.3), which are easier to verify. However, in this form, the necessary conditions are essentially weaker than the sufficient ones; this is demonstrated by simple examples. Finally, we shall give, for the case $C = \mathbb{R}^n$, sufficient optimality conditions using directional derivatives (1.5). A detailed comparison between our results and those of Auslender [2] will be made in § 3.

We denote by $T(C;\bar{x})$ the tangent cone to C at $\bar{x}$ [5] (called also contingent cone [1], [3]). It may be defined in some equivalent manners. We recall three of them:

(2.1) $T(C;\bar{x})=\{h\in\mathbb{R}^n|\forall\alpha>0,\forall V\in\mathcal{N}(h),\ C\cap(\bar{x}+]0,\alpha[V)\neq\emptyset\}$,

(2.2) $h\in T(C;\bar{x})$ if and only if there exist sequences $t_m\downarrow 0$ and $h_m\to h$ such that $\bar{x}+t_mh_m\in C$ for all m,

(2.3) $h\in T(C;\bar{x})$ if and only if there exists a sequence $\{x_m\}\subset C$ converging to $\bar{x}$ and a sequence $\{\lambda_m\}\subset]0,+\infty[$ such that $h=\lim_{m\to\infty}\lambda_m(x_m-\bar{x})$.

Since $T(C;\bar{x})$ is a closed cone and $\underline{d}f(\bar{x};\cdot)$ is positively homogeneous and lower semicontinuous, the set

$$K(\bar{x})=T(C;\bar{x})\cap\{h\in\mathbb{R}^n|\underline{d}f(\bar{x};h)\leqq 0\} \tag{2.4}$$

is also a closed cone (containing 0).

THEOREM 2.1. (i) *If $k>1$, then the following three conditions are equivalent*:

(a) *$\bar{x}$ is an isolated local minimum with order k of problem* (P);

(b) *for all $h\in\mathbb{R}^n\backslash\{0\}$, we have*

$$\underline{d}^kf_C(\bar{x};h)>0; \tag{2.5}$$

(c) *inequality* (2.5) *holds for all $h\in K(\bar{x})\backslash\{0\}$.*

(ii) *If $k=1$, then analogous equivalences are true with condition* (c) *replaced by the following one*:

(c′) *inequality* (2.5) *holds for all $h\in T(C;\bar{x})\backslash\{0\}$.*

Proof. (i); (a)$\Rightarrow$(b): Suppose that (1.1) holds and that the desired conclusion is false, that is, there exists $y\in\mathbb{R}^n\backslash\{0\}$ satisfying $\underline{d}^kf_C(\bar{x};y)\leqq 0$. Hence, by the definition of "lim inf" (see (1.2)), it follows that, for each $\alpha>0$ and $V\in\mathcal{N}(y)$,

$$\inf_{\substack{0<t<\alpha\\ v\in V}}[t^{-k}(f(\bar{x}+tv)-f(\bar{x}))+\delta(\bar{x}+tv|C)]\leqq 0. \tag{2.6}$$

In particular, we may choose α and V so as to satisfy the following conditions:

$$V\subset\{x\in\mathbb{R}^n||x-y|<|y|/2\}, \tag{2.7}$$

$$\bar{x}+]0,\alpha[V\subset U \tag{2.8}$$

(where U is the neighbourhood occurring in (1.1)). Take any $\varepsilon>0$. By (2.6), there exist $s\in]0,\alpha[$ and $w\in V$ such that

$$\bar{x}+sw\in C\quad\text{and}\quad s^{-k}(f(\bar{x}+sw)-f(\bar{x}))\leqq\varepsilon. \tag{2.9}$$

In view of (2.7), we have $|w-y|<|y|/2$, hence $|w|>|y|/2$. By (2.8), $\bar{x}+sw\in U\cap C$. Thus, making use of (2.9) and (1.1), we obtain

$$\varepsilon\geqq s^{-k}(f(\bar{x}+sw)-f(\bar{x}))>s^{-k}\beta|sw|^k>\beta(|y|/2)^k,$$

which leads to a contradiction since ε is arbitrary whereas $\beta(|y|/2)^k>0$ does not depend on ε.

(b)$\Rightarrow$(c) is trivial.

(c)$\Rightarrow$(a). Since the set $S(\bar{x})=\{h\in K(\bar{x})||h|=1\}$ is compact and $\underline{d}^kf_C(\bar{x};\cdot)$ is lower semicontinuous, we deduce from (c) that there exists $\gamma=\min\{\underline{d}^kf_C(\bar{x};h)|h\in S(\bar{x})\}>0$. (*If* $S(\bar{x})$ is empty, the proof is valid for any $\gamma>0$.) Suppose that (a) is false, then (1.1) is false with $\beta=\gamma/2$. Thus, we can choose a sequence $\{x_m\}\subset C$ such that $x_m\to\bar{x}$, $x_m\neq\bar{x}$

for all m, and

$$f(x_m) \leqq f(\bar{x}) + (\gamma/2)|x_m - \bar{x}|^k \quad \text{for all } m. \tag{2.10}$$

Let $t_m = |x_m - \bar{x}|$, $v_m = (x_m - \bar{x})/|x_m - \bar{x}|$. We have $t_m \downarrow 0$ and we may assume, by taking a subsequence, that $\{v_m\}$ converges to a vector y such that $|y| = 1$. By (2.3), we have also $y \in T(C; \bar{x})$. Moreover, in view of (2.10),

$$\begin{aligned} \underline{d}f(\bar{x}; y) &\leqq \liminf_{m\to\infty} t_m^{-1}(f(\bar{x} + t_m v_m) - f(\bar{x})) \\ &= \liminf_{m\to\infty} t_m^{-1}(f(x_m) - f(\bar{x})) \\ &\leqq \lim_{m\to\infty} (\gamma/2) t_m^{k-1} = 0, \end{aligned} \tag{2.11}$$

and so $y \in S(\bar{x})$ (see (2.4)), which implies $\underline{d}^k f_C(\bar{x}; y) \geqq \gamma$. Hence, by (1.4), we obtain that there exist $\alpha > 0$ and $V \in \mathcal{N}(y)$ such that

$$t^{-k}(f(\bar{x} + tv) - f(\bar{x})) > \gamma/2 \tag{2.12}$$

for all $t \in]0, \alpha[$ and $v \in V$ satisfying $\bar{x} + tv \in C$. For all sufficiently large m, we have $t_m \in]0, \alpha[$ and $v_m \in V$. Since $\bar{x} + t_m v_m = x_m \in C$, we obtain for those m, by (2.12), $t_m^{-k}(f(x_m) - f(\bar{x})) > \gamma/2$, which contradicts (2.10).

(ii) The only change with respect to the case (i) is now the following: when proving $(c') \Rightarrow (a)$, we cannot verify (2.11) (since it contradicts (2.5) for $k = 1$ and $C = \mathbb{R}^n$), and so $K(\bar{x})$ has to be replaced by $T(C; \bar{x})$. □

From Theorem 2.1 and from the obvious inequality $\underline{d}^k f_C(\bar{x}; \cdot) \geqq \underline{d}^k f(\bar{x}; \cdot)$, we deduce the following.

COROLLARY 2.1. (i) *If* $\underline{d}f(\bar{x}; h) > 0$ *for all* $h \in T(C; \bar{x})\backslash\{0\}$, *then* $\bar{x}$ *is an isolated local minimum with order* 1 *of problem* (P).

(ii) *If* $k > 1$ *and* $\underline{d}^k f(\bar{x}; h) > 0$ *for all* $h \in K(\bar{x})\backslash\{0\}$, *then* $\bar{x}$ *is an isolated local minimum with order* k *of problem* (P).

We now establish necessary optimality conditions in a similar form.

THEOREM 2.2. *If* $\bar{x}$ *is an isolated local minimum with order* $k \geqq 1$ *of problem* (P), *then* $\bar{d}^k f(\bar{x}; h) > 0$ *for all* $h \in T(C; \bar{x})\backslash\{0\}$.

Proof. Suppose that the conclusion is false, hence there exists $y \in T(C; \bar{x})\backslash\{0\}$ satisfying $\bar{d}^k f(\bar{x}; y) \leqq 0$. This means, in view of (1.3), that, for any $\varepsilon > 0$, there exist $\alpha > 0$ and $V \in \mathcal{N}(y)$ such that

$$t^{-k}(f(\bar{x} + tv) - f(\bar{x})) \leqq \varepsilon \quad \text{for all } t \in]0, \alpha[\quad \text{and } v \in V. \tag{2.13}$$

We may assume that α and V satisfy conditions (2.7) and (2.8) (if necessary, we can take them smaller and (2.13) remains true). Since $y \in T(C; \bar{x})$, we have $(\bar{x} +]0, \alpha[\, V) \cap C \neq \emptyset$ by (2.1). From this fact and from (2.13) we infer that there exist $s \in]0, \alpha[$ and $w \in V$ satisfying (2.9). The remaining part of the proof is the same as in Theorem 2.1, (a)$\Rightarrow$(b). □

In the following two examples we assume that $k \geqq 1$, $n = 2$, $C = \mathbb{R} \times \{0\}$, and $\bar{x} = (0, 0)$.

Example 2.1. Let f be given by

$$f(x_1, x_2) = \begin{cases} 0 & \text{if } x_2 = 0, \\ |x_1|^k & \text{if } x_2 \neq 0. \end{cases}$$

We have $\bar{d}^k f(\bar{x}; (h_1, 0)) = |h_1|^k > 0$ for all $h_1 \neq 0$, and so f satisfies the necessary optimality condition given in Theorem 2.2. Nevertheless, $\bar{x}$ is not an isolated local minimum with order k of problem (P).

Example 2.2. Let f be given by

$$f(x_1, x_2) = \begin{cases} |x_1|^k & \text{if } x_2 = 0, \\ 0 & \text{if } x_2 \neq 0. \end{cases}$$

Then $\bar{x}$ is an isolated local minimum with order k of problem (P), but sufficient conditions of Corollary 2.1 are not satisfied since $\underline{d}f(\bar{x}; \cdot)$ and $\underline{d}^k f(\bar{x}; \cdot)$ are identically zero.

Let us now consider the unconstrained case $C = \mathbb{R}^n$. The following corollary is a direct consequence of Corollary 2.1 (ii) and Remark 1.3.

COROLLARY 2.2. *Suppose that* $k > 1$, *that conditions* (1.6) *are fulfilled, and that* $f_-^{(k)}(\bar{x}; h^k) > 0$ *for all* $h \neq 0$ *satisfying* $\underline{d}f(\bar{x}; h) = 0$. *Then* $\bar{x}$ *is an isolated local minimum with order* k *of problem* (P) *where* $C = \mathbb{R}^n$.

It follows from Proposition 1.1 that Corollary 2.2 generalizes the classical higher order sufficient conditions for a local minimum of a smooth function (let us note that, in the smooth case, the condition $\underline{d}f(\bar{x}; h) = 0$ is a consequence of (1.6), and so may be omitted).

3. Comparison with Auslender's results. Auslender [2, Prop. 2.1] gives necessary and sufficient optimality conditions for (P) (with locally Lipschitzian f), using the following expression:

$$f^*(\bar{x}; h) = 2 \liminf_{\substack{h \\ u \to 0}} [|u|^{-2}(f(\bar{x}+u) - f(\bar{x})) + \delta(\bar{x}+u|C)] \tag{3.1}$$

where, for any $\varphi : \mathbb{R}^n \setminus \{0\} \to \bar{\mathbb{R}}$ and $h \neq 0$,

$$\liminf_{\substack{h \\ u \to 0}} \varphi(u) = \sup_{\varepsilon > 0} \left(\inf \left\{ \varphi(u) \middle| \left| \frac{u}{|u|} - \frac{h}{|h|} \right| \leqq \varepsilon, 0 < |u| \leqq \varepsilon \right\} \right). \tag{3.2}$$

Let us first compare the limits (1.2) and (3.2).

LEMMA 3.1.

$$\liminf_{\substack{h \\ u \to 0}} \varphi(u) = \liminf_{\substack{t \downarrow 0 \\ v \to h}} \varphi(tv).$$

Proof. Observe that

$$\liminf_{\substack{t \downarrow 0 \\ v \to h}} \varphi(tv) = \sup_{\eta > 0} (\inf \{ \varphi(tv) | 0 < t \leqq \eta, |v - h| \leqq \eta \}).$$

Thus, it suffices to verify the following two conditions:

(3.3) for any $\varepsilon > 0$, there exists $\eta > 0$ such that $|tv/|tv| - h/|h|| \leqq \varepsilon$ and $0 < |tv| \leqq \varepsilon$ for all t and v satisfying $0 < t \leqq \eta$, $|v - h| \leqq \eta$;

(3.4) for any $\eta > 0$, there exists $\varepsilon > 0$ such that each vector u satisfying $|u/|u| - h/|h|| \leqq \varepsilon$ and $0 < |u| \leqq \varepsilon$ can be represented in the form $u = tv$ where $0 < t \leqq \eta$ and $|v - h| \leqq \eta$.

In order to prove (3.3), take any $\varepsilon > 0$ (we may assume $\varepsilon \leqq 1$) and define $\eta = \min(\varepsilon|h|/2, 2\varepsilon/3|h|)$. If $0 < t \leqq \eta$ and $|v - h| \leqq \eta$, then $|v - h| \leqq |h|/2$, hence $0 < |h|/2 \leqq$

$|v| \leqq (3/2)|h|$. Consequently, $0 < |tv| \leqq \varepsilon$ and

$$\left|\frac{tv}{|tv|} - \frac{h}{|h|}\right| = \frac{1}{|v||h|}|(|h| - |v|)v + |v|(v-h)|$$

$$\leqq \frac{1}{|h|}(\|h| - |v\| + |v-h|) \leqq \frac{2}{|h|}|v-h| \leqq \varepsilon.$$

Let us now verify (3.4). For any $\eta > 0$, we define $\varepsilon = \min(\eta|h|, \eta/|h|)$. If u satisfies the assumptions of (3.4), we put $t = |u|/|h|$, $v = (|h|/|u|)u$, and obtain $0 < t \leqq \eta$, $|v-h| = |h||u/|u| - h/|h|| \leqq \eta$. □

We are now able to establish a relation between the limits (1.4) (for $k=2$) and (3.1). Note that we need not assume f to be locally Lipschitzian.

PROPOSITION 3.1. *For all $h \neq 0$, we have*

$$f^*(\bar{x}; h) = \frac{2}{|h|^2}\, \underline{d}^2 f_C(\bar{x}; h).$$

Proof. Making use of (1.4), (3.1) and Lemma 3.1, we find that

$$f^*(\bar{x}; h) = 2 \liminf_{\substack{t \downarrow 0 \\ v \to h}} [|tv|^{-2}(f(\bar{x}+tv) - f(\bar{x})) + \delta(\bar{x}+tv|C)]$$

$$= 2\,(\lim_{\substack{t \downarrow 0 \\ v \to h}} |v|^{-2})\, \underline{d}^2 f_C(\bar{x}; h) = \frac{2}{|h|^2}\, \underline{d}^2 f_C(\bar{x}; h). \qquad \square$$

Remark 3.1. In a similar way, using (1.5) (for $k=2$), Remark 1.2 and Lemma 3.1, we can show that the lower second-order directional derivative introduced in [2, (2.2)] is equal to $h \mapsto (1/|h|^2) f_-^{(2)}(\bar{x}; h^2)$ if f is locally Lipschitzian.

4. Conclusions. Let us now consider the case of locally Lipschitzian f. It follows from Remark 1.2 and Proposition 3.1 that all the statements of [2, Prop. 2.1] are particular cases of our results presented in § 2. More precisely, it is easy to see that:

(a) the necessary condition for an isolated local minimum with order 1 is a consequence of Theorem 2.2 (where $k=1$),

(b) the necessary condition [2, (2.6)] for an isolated local minimum with order 2 follows from Theorem 2.1, (a)⇒(b) (where $k=2$),

(c) the sufficient condition for an isolated local minimum with order 1 is another formulation of Corollary 2.1(i),

(d) the sufficient condition [2, (2.8)] for an isolated local minimum with order 2 follows from Theorem 2.1, (c)⇒(a) (where $k=2$).

Similarly, it follows from Remark 3.1 that Corollary 2.2 of [2] is a consequence of our Corollary 2.2 (where $k=2$).

REFERENCES

[1] J. P. AUBIN AND A. CELLINA, *Differential Inclusions*, Springer-Verlag, Berlin, 1984.

[2] A. AUSLENDER, *Stability in mathematical programming with nondifferentiable data*, this Journal, 22 (1984), pp. 239-254.

[3] A. D. IOFFE, *Calculus of Dini subdifferentials of functions and contingent coderivatives of set-valued maps*, Nonlinear Anal., 8 (1984), pp. 517-539.

[4] S. MIRICĂ, *The contingent and the paratingent as generalized derivatives for vector-valued and set-valued mappings*, Nonlinear Anal., 6 (1982), pp. 1335-1368.

[5] J. P. PENOT, *Calcul sous-différentiel et optimisation*, J. Funct. Anal., 27 (1978), pp. 248-276.

[6] L. SCHWARTZ, *Analyse mathématique* I, Hermann, Paris, 1967.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 5, September 1986

010

SOLVING THE LINEAR COMPLEMENTARITY PROBLEM IN CIRCUIT SIMULATION*

J. T. J. VAN EIJNDHOVEN†

Abstract. In the simulation of electronic circuits piecewise linear modelling yields a global circuit description which in principle can be used to solve for a circuit response in a finite number of steps. During the solution process a sequence of linear complementarity problems (LCP) has to be solved within the piecewise linear system description. The purpose of this paper is to present and discuss some new methods to solve this LCP for certain matrix classes.

To start with, two types of LCP solution algorithms are briefly described: pivoting algorithms and the modulus algorithm. It is shown that these algorithms have certain disadvantages if applied to the problem as stated above. Those problems can be overcome by the new methods to be presented. The first one is a modified version of an iterative algorithm of O. L. Mangasarian. The second one is a so-called simplicial method, based on a new integer labelling and an efficient labelling algorithm. Convergence conditions are given, as is a bound for the error in the approximate solution.

In both new algorithms full advantage can be taken of sparse matrix techniques. The labelling algorithm turns out to converge for a large class of matrices, comparable with standard pivoting methods.

Key words. linear complementarity, piecewise linear, large scale, simulation, integer labelling

AMS(MOS) subject classifications. 65-C20, 65-D20, 65-F50, 90-C06, 90-C33

1. Introduction. In the electronic industry, the use of simulation programs is widely accepted. Especially in the design of large scale integrated circuits these programs have become indispensable. The testing of the correctness of a circuit prior to the actual integration by building a model with discrete components, will deliver increasingly poorer results with growing circuit complexity. The high frequencies and the short distances in the extremely small and complex structures, being recognized in recent devices, make a good electronic equivalent of an integrated circuit with discrete components impossible.

In a computer simulation all these effects can be incorporated. Hence a better approximation can be achieved. However, the design of such a program is enormously complex. The program must be able to perform logic simulation of digital circuits, timing simulation of these devices as well as the computation of transient responses of analog circuits. Furthermore, the communication between these different simulation levels through a common database is a very difficult problem. Finally, the resulting set of nonlinear equations is in general solved by applying the Newton-Raphson iteration scheme, which implies problems of numerical stability.

Recently, a new method was proposed for the simulation of electronic circuits. This method uses piecewise-linear approximations for all nonlinear functions [v. Bokhoven, '80]. It has the following advantages:

- The method is extremely well suited for macro modelling, which is an important feature in the simulation of LSI circuits.
- Mixed analog/digital simulation is automatically incorporated. The structure of the equations is identical for each device, no matter whether it is a macromodel of a logic gate or a detailed Gummel-Poon transistor model.

* Received by the editors April 30, 1982, and in final revised form August 1, 1985.

† Eindhoven University of Technology, Department of Electrical Engineering, P.O. Box 513, Eindhoven, the Netherlands. This work was sponsored by the Netherlands Organization for the Advancement of Pure Research (ZWO).

- A complete system description consists of a single matrix, with a standard block structure yielding a very compact datastructure.
- Analytic nonlinear functions (like exp, ln, sqrt) no longer need to be evaluated, so a speed-up of the simulation can be expected.
- A network can easily be built up from smaller blocks, which allows for a hierarchical structure. This is indispensable in the design of LSI circuits. It also renders the possibility to construct a library of standard functions and networks.
- Numerical problems with discontinuous signals are almost absent, while conventional programs, using some version of Newton–Raphson iteration, show very serious convergence problems and large computation times if signals contain fast transients.

Some new problems are introduced, however. The most serious of those problems is that during a transient simulation at each time step a linear complementarity problem (LCP) has to be solved. This LCP is given as follows: For given vector q and matrix M, find vectors v and i such that

$$v = M \cdot i + q, \quad v, i \in \mathbb{R}^n_+, \quad v^t \cdot i = 0$$

where:

- M is a large sparse matrix,
- M is generally not restricted to a certain class of matrices (see § 2),
- the LCP has one or more solutions.

The main purpose of this paper is to present new algorithms particularly suitable to solve the LCP as stated above. In the next section we shall give the piecewise linear system description and derive the associate LCP.

2. The linear complementarity problem. A piecewise linear system (see Fig. 1) will be described by the following equations:

$$\begin{pmatrix} y \\ v \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} x \\ i \end{pmatrix} + \begin{pmatrix} g \\ h \end{pmatrix} \tag{2.1}$$

$$v, i \in \mathbb{R}^n_+, \qquad v^t \cdot i = 0.$$

In general this set of equations describes different linear relations between x and y on maximally 2^n different polytopes. This can be shown as follows: Define 2^n different n-vectors S^k, $1 \leqq k \leqq 2^n$, such that

$$S^k_j \in \{0, 1\}, \qquad 1 \leqq j \leqq n.$$

Let the n-vectors i^k and v^k be such that:

$$i^k_j = i_j, \qquad v^k_j = v_j \quad \text{if } S^k_j = 0,$$
$$i^k_j = v_j, \qquad v^k_j = i_j \quad \text{if } S^k_j = 1.$$

Transform the equations (2.1) by partial inversion into

$$\begin{pmatrix} y \\ v^k \end{pmatrix} = \begin{pmatrix} A^k & B^k \\ C^k & D^k \end{pmatrix} \begin{pmatrix} x \\ i^k \end{pmatrix} + \begin{pmatrix} g^k \\ h^k \end{pmatrix} \tag{2.2}$$

and assume $V^k = \{x \mid C^k x + h^k \in \mathbb{R}^n_+\}$.

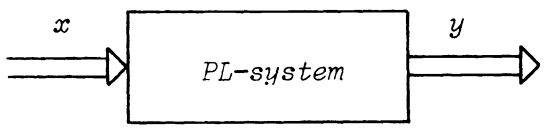


FIG. 1. *A piecewise linear system.*

Now $v^k = C^k \cdot x + h^k$, $i^k = 0$ is a solution of (2.1) for a given $x \in V^k$. Thus $y = A^k \cdot x + g^k$ is the linear mapping in the region V^k. Obviously, there are at most 2^n different regions V^k and any of these regions may have its own linear mapping. In general some V^k may be empty, may overlap, or some x may not belong to any V^k (no solution exists).

For given x, (2.1) can be solved by determining

$$q = Cx + h, \qquad M = D$$

and solving the LCP:

$$\begin{aligned} &v = M \cdot i + q, \\ &v, i \in \mathbb{R}^n_+, \qquad v^t \cdot i = 0. \end{aligned} \tag{2.3}$$

Then y can easily be computed when i is found. However, solving the LCP may be difficult. For example multiple solutions may exist (circuits with memory, for instance flip-flops and thyristors).

For the characterization of the LCP some classes of matrices have been defined by various authors (see for instance [Karamardian, '72]). Four of them are repeated here:

- *Positive definite matrices* (PD).
  M is of class PD if and only if
  $\forall x \in \mathbb{R}^n, \quad x \neq 0: x^t M x > 0.$
- *P-matrices* (P).
  M is of class P if and only if
  $\forall x \in \mathbb{R}^n, \quad x \neq 0, \quad \exists k: x_k \cdot (Mx)_k > 0.$
- *Strictly copositive matrices* (SCP).
  M is of class SCP if and only if
  $\forall x \in \mathbb{R}^n_+, \quad x \neq 0: x^t M x > 0$
- *Strictly semimonotone matrices* (SSM) (*completely $-Q$ matrices*).
  M is of class SSM if and only if
  $\forall x \in \mathbb{R}^n_+, \quad x \neq 0, \quad \exists k: x_k \cdot (Mx)_k > 0.$

These definitions imply that PD $\subset$ P $\subset$ SSM and PD $\subset$ SCP $\subset$ SSM.

In relation to the LCP it is known that $M \in$ SSM implies that there exists at least one solution for each q. There exists a unique solution for each q if and only if $M \in$ P.

An example of a PL-system with hysteresis (multiple solutions) and $M \in$ SCP is the following system:

$$\left(\begin{array}{c} y \\ \hline v_1 \\ v_2 \end{array}\right) = \left(\begin{array}{c|cc} 0 & 1 & 0 \\ \hline 0 & 1 & 2 \\ \frac{1}{2} & 1 & 1 \end{array}\right) \left(\begin{array}{c} x \\ \hline i_1 \\ i_2 \end{array}\right) + \left(\begin{array}{c} 0 \\ \hline -1 \\ -1 \end{array}\right).$$

This system has an input-output relation as shown in Fig. 2.

Solving the LCP, in the context of piecewise linear simulation, was investigated earlier by van Bokhoven (see [van Bokhoven, '81]). He mainly considered pivoting algorithms and the modulus algorithm.

At the moment there are many different versions of pivoting algorithms, (see for instance [Lemke, '70], [Eaves, '71], [Cottle, '68], [van der Heyden '80]. They are based on the original algorithm of Lemke [Lemke, '65], and differ mainly in the choice of the pivots and the class of matrices for which convergence is assured.

In a simulation process, the matrices do not necessarily belong to any of such classes. The algorithm may be unable to choose a nonzero pivot and terminates

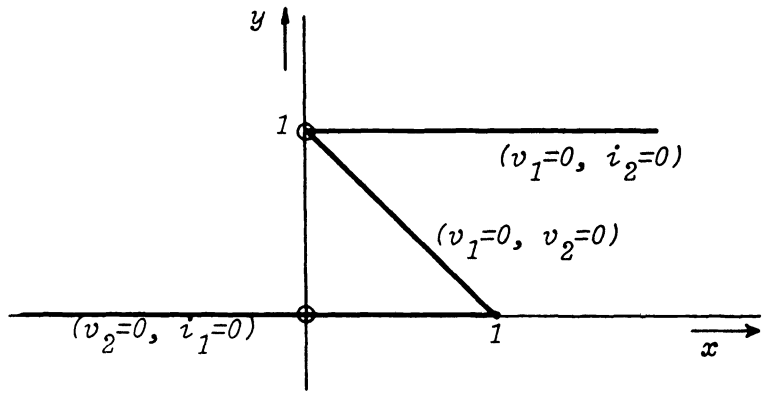


FIG. 2. *An example of a system with hysteresis.*

unsuccessfully. Furthermore, M will be a large sparse matrix, and a complex datastructure is required for the pivoting operations. In addition sparsity may be lost as a consequence of pivoting [van Eijndhoven, '82] [van Eijndhoven, '84] [Saigal, '81] [Fujisawa, '72].

The modulus algorithm [van Bokhoven, '81] has the advantage that for a given conditioning of the problem the time-complexity is a polynomial function of the dimension n. The method is based on contraction mapping and is more appropriate for sparse matrix processing. Unfortunately however, convergence can only be assured for positive-definite matrices.

Because both the pivoting method and the modulus algorithm did not have the desired properties, other methods have been searched for. The results are presented in the next two sections. In § 3 an iterative algorithm is presented to minimize a quadratic function, related to the LCP. Basically the algorithm is the same as [Mangasarian, '77], but by using a different quadratic function we can process non-symmetric matrices as well. The new method solves the LCP for all class P matrices.

In § 4 a simplex method is presented. The space $\mathbb{R}^n_+$ is triangulated and a new type of integer labelling is given. An algorithm of v.d. Laan and Talman [v.d. Laan, '80] is used to search for a completely labelled simplex, starting from an arbitrary chosen initial point. Convergence is proved for all matrices of class SSM, independent of the gridsize. A bound for the error of the approximate solution is derived.

3. A minimization algorithm. In this section we will solve the LCP by transforming it into a quadratic programming problem, which is subsequently solved by systematic overrelaxation. This approach was also followed in [Cryer, '71] and [Mangasarian, '77] but we apply this method to a different quadratic function. This will lead to a different class of matrices for which convergence can be guaranteed as we shall see later on.

Consider the LCP:

find vectors $v, i \in \mathbb{R}^n_+$ such that

$$v = Mi + q, \qquad v^t i = 0. \tag{3.1}$$

This problem is equivalent to:

find vectors $v, i \in \mathbb{R}^n_+$ such that

$$f(i, v) = \alpha v^t \cdot i + \tfrac{1}{2}(Mi + q - v)^t (Mi + q - v) = 0 \tag{3.2}$$

for some $\alpha > 0$.

The equivalence of (3.1) and (3.2) is obvious: $f(i^*, v^*) = 0$ for all pairs (i^*, v^*) that solve (3.1), and $f(i, v) > 0$ for all other pairs $(i, v) \in \mathbb{R}^n_+ \times \mathbb{R}^n_+$.

To find the point (i^*, v^*) (which minimizes f) one can use an algorithm as found in [Mangasarian, '77] and we shall show that $M \in P$ is a sufficient condition to solve (3.2) by this algorithm.

Let $u = \binom{i}{v} \in \mathbb{R}^{2n}_+$ and let $f(u) = f(i, v)$ according to (3.2). Denote the gradient $g(u)$ and the hessian H_α of $f(u)$ by:

$$g(u) = g(i, v) = \alpha \begin{pmatrix} v \\ i \end{pmatrix} + \begin{pmatrix} M^t(Mi + q - v) \\ -(Mi + q - v) \end{pmatrix}, \tag{3.3}$$

$$H_\alpha = (h_{ij}) = \alpha \begin{pmatrix} 0 & I \\ I & 0 \end{pmatrix} + \begin{pmatrix} M^tM & -M^t \\ -M & I \end{pmatrix}. \tag{3.4}$$

A possible algorithm to minimize $f(u)$ can be:

Define for $k = 1, 2, \cdots$ the vectors $u^k \in \mathbb{R}^{2n}_+$ and for each k, for $l = 0, 1, 2, \cdots, 2n$ the vectors $u^{k,l} \in \mathbb{R}^{2n}_+$ such that

$u^0 \in \mathbb{R}^{2n}_+$ is a given starting point of the algorithm

$$u_j^{k,l} = \begin{cases} u_j^k & \text{for } 1 \leqq j \leqq l, \\ u_j^{k-1} & \text{for } l+1 \leqq j \leqq 2n, \end{cases}$$

with

$$u_j^k = \max\,(0,\, u_j^{k-1} - \omega g_j(u^{k,j-1})/h_{jj}) \tag{3.5}$$

where ω is a relaxation parameter with $0 < \omega < 2$, and h_{jj} is positive due to the definition of f.

This algorithm can be found in [Cryer, '71] and is only a special case of a more general class of algorithms found in [Mangasarian, '77]. In the next three theorems it will be proved that the sequence will converge for $M \in$ SSM and that the limit points determine a solution of the LCP if $M \in$ P.

THEOREM 3.1. *H_α is symmetrical and $\forall \alpha > 0$: $M \in$ SSM $\rightarrow H_\alpha \in$ SCP.*

Proof.

(1) The symmetry of H_α is obvious.

(2) $$\begin{pmatrix} i \\ v \end{pmatrix}^t H_\alpha \begin{pmatrix} i \\ v \end{pmatrix} = 2\alpha i^t v + (Mi - v)^t(Mi - v)$$

$$M \in \text{SSM} \rightarrow \forall 0 \neq i \geqq 0 \exists_k\colon i_k(Mi)_k > 0$$

$$\rightarrow \forall 0 \neq \begin{pmatrix} i \\ v \end{pmatrix} \geqq 0\colon v = M \cdot i,\ i^t v = 0 \text{ has no solution}$$

$$\rightarrow \forall 0 \neq \begin{pmatrix} i \\ v \end{pmatrix} \geqq 0,\ \alpha > 0\colon \begin{pmatrix} i \\ v \end{pmatrix}^t H_\alpha \begin{pmatrix} i \\ v \end{pmatrix} > 0$$

$$\rightarrow \forall \alpha > 0\colon H_\alpha \in \text{SCP}. \qquad \square$$

THEOREM 3.2. *The sequence (u^k) has at least one limit point u^* if $M \in$ SSM. Each limit point u^* satisfies*

$$u^* \geqq 0, \qquad g_\alpha(u^*) \geqq 0, \qquad u^{*t} \cdot g_\alpha(u^*) = 0. \tag{3.6}$$

Proof. Because $M \in$ SSM implies $H_\alpha \in$ SCP, the more general algorithms with convergence proofs found in [Mangasarian, '77] apply here. Basically the existence

of limit points follows from the nonincreasing property of the sequence $(f(u^k))$ and $H_\alpha \in \mathrm{SCP}$, which implies that the sequence (u^k) is bounded an *a* compact set. The inequalities for u^* easily follow from the definition of the algorithm. □

THEOREM 3.3. *If* $M \in \mathrm{P}$ *then* $f(u^*)=0$, *with* u^* *a limit point of* (3.5), *and hence a solution of* (3.1) *is found.*

Proof. From (3.5) (see also [Cryer, '71]) it follows that:

$$\forall k, 1 \leqq k \leqq 2n \quad u_k^* \geqq 0, \quad g_k(u^*) \geqq 0, \quad u_k^* \cdot g_k(u^*) = 0.$$

Now returning to a notation in i and v, computing $g_k(u^*) \cdot g_{k+n}(u^*) \geqq 0$ $(1 \leqq k \leqq n)$ and substituting $u_k^* \cdot g_k(u^*) = 0 \wedge u_{k+n}^* \cdot g_{k+n}(u^*) = 0$ yields:

$$(3.7) \qquad \forall k, 1 \leqq k \leqq n \qquad \alpha^2 i_k^* \cdot v_k^* + (Mi^* + q - v^*)_k [M^t(Mi^* + q - v^*)]_k \leqq 0.$$

Now let M in class P (which implies M^t in class P) and assume $\|Mi^* + q - v^*\| \neq 0$.

Then, due to the definition of class P

$$\exists j, 1 \leqq j \leqq n \qquad (Mi^* + q - v^*)_j \cdot [M^t(Mi^* + q - v^*)]_j > 0$$

which contradicts (3.7) because $i_j^* \cdot v_j^* \geqq 0$. Therefore $\|Mi^* + q - v^*\| = 0$ and (from (3.7)) $v^{*t} \cdot i^* = 0$. So $f(i^*, v^*) = 0$ and (i^*, v^*) is a solution of (3.1). □

If $M \in \mathrm{P}$ it is known that there exists only one solution to the LCP [Karamardian, '72], therefore only one limit point u^* exists and thus (see Theorem 3.2)

$$\lim_{k \to \infty} u^k = u^* \text{ is the unique solution of (3.1).}$$

About the algorithm we would like to remark:

1. The new method solves (3.1) for all class P matrices, whereas the algorithms of Mangasarian apply to symmetrical matrices only. For our application though, the restriction to class P remains disappointing.

2. Like the problem, the algorithm is symmetrical in v and i.

3. The sequence (u^k) will converge for a much larger class of matrices. However, for $M \notin \mathrm{P}$ the algorithm may find a vector u^* for which $f(u^*) > 0$ and then u^* is not a solution of the LCP.

4. One must choose $\alpha > 0$ and $0 < \omega < 2$. Fast convergence is obtained in practice with $\omega = 1.5$ and α in the order of magnitude of the elements of M.

4. A simplex algorithm. In this section we shall use a simplex method to find a solution of the LCP. A choice was made for an algorithm developed by v.d. Laan and Talman because they presented a closed theory in combination with a set of related efficient algorithms, triangulations and labelling functions (see [v.d. Laan, '80] and [Talman, '80]).

However, these algorithms were mainly intended to solve equations of the form $x = f(x)$. Therefore another integer labelling is developed, especially suited for the LCP, which satisfies all conditions of a so-called "proper" labelling. It can be proved that the algorithm will find an approximate solution from each starting point in $\mathbb{R}_+^n$ and with each grid-size if the matrix of the LCP is of class SSM. We will restrict ourselves to integer labelling because vector labelling requires the repeated solving of large sets of linear equations, which is just the type of operations we would like to avoid as stated in § 2. It is noted that the infinite norm will be used, i.e.

$$\|x\| = \|x\|_\infty = \max_i |x_i|, \qquad \|M\| = \|M\|_\infty = \max_i \sum_{j=1}^{n} |m_{ij}|.$$

4.1. Triangulation. Let $I_n = \{1, 2, \cdots, n\}$ and for each $k \in I_n$ the vector $e(k)$ be defined such that

$$e_i(k) = \begin{cases} 1 & \text{if } i = k, \\ 0 & \text{if } i \neq k,\ i \in I_n. \end{cases}$$

Let $C_\lambda = \{x \in \mathbb{R}^n_+ \mid \|x\| \leqq \lambda\}$ and the boundaries C^j_λ of C_λ

$$C^j_\lambda = \begin{cases} \{x \in C_\lambda \mid \|x\| = \lambda\} & \text{if } j = 0, \\ \{x \in C_\lambda \mid x_j = 0\} & \text{if } j \in I_n. \end{cases}$$

C_λ is triangulated by the collection of simplices $\sigma(y^1, P)$ with vertices $y^1, y^2, \cdots, y^{n+1}$ in C_λ given by:

- each coordinate of y^1 is an integer multiple of δ(=gridsize) where $\lambda = k\delta$ for some positive integer k,
- $P = (p_1, p_2, \cdots, p_n)$ is a permutation of the elements of I_n,
- $y^{i+1} = y^i + \delta \cdot e(p_i)$, $i \in I_n$.

This collection of simplices yields the standard K-triangulation.

4.2. Labelling. Define a labelling function l on C_λ such that

$$l(x) \in \{0, 1, \cdots, n\} \quad \text{for all } x \in C_\lambda.$$

A simplex is called *completely labelled* if the $(n+1)$ vertices carry all the $(n+1)$ different labels.

A labelling function is called *proper* if

$$l(x) \neq j \text{ for all } x \in C^j_\lambda, \qquad j \in I_n \cup \{0\}.$$

Let $f(x) = Mx + q$; the LCP can then be rephrased as follows:

(4.1) find a vector $x \in \mathbb{R}^n_+$ such that $f(x) \in \mathbb{R}^n_+$ and $x^t \cdot f(x) = 0$.

For solving the LCP by using a simplicial algorithm the labelling function must be such that

- The labelling is proper. This guarantees that the algorithm, given in § 4.3, does not leave C_λ and will find a completely labelled simplex.
- A completely labelled simplex provides an approximation of the solution of the LCP (4.1).

We prove that the labelling defined below satisfies these conditions if $M \in \text{SSM}$.

$$l(x) = \begin{cases} 0 & \text{if } \forall i \in I_n\colon x_i \cdot f_i(x) \leqq 0, \\ j & \text{if } j = \min\{k \in I_n \mid x_k \cdot f_k(x) = \max_{i \in I_n} x_i \cdot f_i(x) > 0\}. \end{cases} \tag{4.2}$$

THEOREM 4.1. *A completely labelled simplex yields an approximation of the* LCP.

Proof. Let x^j the vertex of the completely labelled simplex with label j and

$$f^j = f(x^j), \qquad j \in I_n \cup \{0\}$$

then

$$\|x^j - x^0\| = \delta, \qquad j \in I_n$$

and

$$|f^j_j - f^0_j| \leqq \|f^j - f^0\| \leqq \delta \cdot \|M\| \to f^0_j > -\delta \cdot \|M\|.$$

Furthermore by definition (4.2)

$$x_j^0 = 0 \quad \text{for all } j \text{ with } f_j^0 > 0.$$

Thus x^0, f^0approximates the solution of the LCP, and the approximation becomes arbitrarily accurate with decreasing δ. □

Furthermore, if $M \in \mathrm{P}$ it is known that there exists a unique solution $\tilde{x}, \tilde{f}$ of the LCP and the distance between x^0 and $\tilde{x}$ can be bounded.

THEOREM 4.2. *If $M \in \mathrm{P}$ then $\|x^0 - \tilde{x}\| < \delta \cdot C$ with C a condition number of the matrix M and δ the gridsize.*

Proof. If $x^0 = \tilde{x}$ we are ready. Otherwise for some $\beta \in \mathbb{R}^n$, $\beta \neq 0$, $\tilde{x} = x^0 + \beta$ and:

$$\tilde{x} = x^0 + \beta \in \mathbb{R}^n_+,$$
$$\tilde{f} = f^0 + M \cdot \beta \in \mathbb{R}^n_+,$$
$$\tilde{x}^t \cdot \tilde{f} = 0.$$

Next we use

$$M \in \mathrm{P} \to \exists \varepsilon_M > 0 \forall \beta \in \mathbb{R}^n, \beta \neq 0 \exists k \in I_n : \frac{\beta_k \cdot (M\beta)_k}{\|\beta\|^2} \geqq \varepsilon_M. \tag{4.3}$$

The proof of this implication is analogous to the proof of Lemma 1, given in the Appendix.

Now let the pair β, k satisfy (4.3). Then

$$\left.\begin{aligned} &\beta_k \cdot (M\beta)_k > 0 \\ &\tilde{x}_k \geqq 0 \wedge \tilde{f}_k \geqq 0 \\ &x_k^0 \cdot f_k^0 \leqq 0 \text{ (due to the labelling)} \end{aligned}\right\} \to \beta_k > 0 \wedge (M\beta)_k > 0.$$

This implies that $\tilde{x}_k > 0$, thus $\tilde{f}_k = 0$ and $(M\beta)_k = -f_k^0 > 0$ where $-f_k^0 < \delta \cdot \|M\|$ due to labelling.

Thus finally

$$\|\beta\| \leqq \frac{1}{\varepsilon_M} \frac{\beta_k}{\|\beta\|} (M\beta)_k < \frac{1}{\varepsilon_M} \cdot \delta \cdot \|M\|$$

or

$$\|\beta\| < \delta \cdot C \text{ with condition number } C = \frac{\|M\|}{\varepsilon_M}. \qquad \square$$

Next we shall prove that the labelling (4.2) is proper for matrices $M \in \mathrm{SSM}$. To this purpose we shall apply again Lemma 1 which is proved in the Appendix.

LEMMA 1.

$$M \in \mathrm{SSM} \to \exists \varepsilon_M > 0 \forall x \in \mathbb{R}^n_+, x \neq 0 \exists k \in I_n : \frac{x_k \cdot (Mx)_k}{\|x\|^2} \geqq \varepsilon_M. \tag{4.4}$$

THEOREM 4.3. *$M \in \mathrm{SSM} \to \exists \lambda_M > 0 \forall \lambda \geqq \lambda_M [l(x)$ as defined in (4.2) is proper on $C_\lambda] \to$*

$$x_i = 0 \to l(x) \neq i, \qquad i \in I_n, \tag{4.5a}$$

$$\|x\| = \lambda \geqq \lambda_M \to l(x) \neq 0. \tag{4.5b}$$

Proof. (4.5a) follows immediately from (4.2) independent of the matrix. So only (4.5b) needs to be proved.

Let ε_M according to (4.4) and choose $\lambda_M > 0$ such that

$$\min_{i \in I_n} (\varepsilon_M \cdot \lambda_M + q_i) > 0.$$

Then

$$x \in \mathbb{R}^n_+ \wedge \|x\| = \lambda \geqq \lambda_M \to \exists_{i \in I_n} x_i \cdot f_i = x_i (Mx+q)_i$$
$$\geqq \varepsilon_M \|x\|^2 + x_i \cdot q_i \geqq x_i \cdot (\varepsilon_M \cdot \lambda_M + q_i) > 0 \to l(x) \neq 0.$$

We shall see that ε_M and λ_M need not be computed for the algorithm. Their existence is already sufficient for the convergence of the algorithm. □

4.3. The algorithm. Now we shall use the "variable dimension restart algorithm" of v. d. Laan and Talman in order to find a completely labelled simplex (see for instance [v.d. Laan, '80a] or [v.d. Laan, '80b]). The algorithm will be given in a slightly different notation and without proof. It has the following properties:

- No special points are to be defined outside C_λ to bound the domain or to start the algorithm.
- The algorithm generates a sequence of adjacent simplices of variable dimension.
- The algorithm is started with a zero-dimensional simplex: the starting point, an arbitrary chosen gridpoint in $\mathbb{R}^n_+$

Let

$$u(i) = \begin{cases} -e(i) & \text{if } i \in I_n, \\ (1, 1, \cdots, 1)^t & \text{if } i = 0. \end{cases}$$

Define a t-dimensional simplex $\sigma(y^1, P(T))$, $0 \leqq t \leqq n$ with its vertices $y^1, y^2, \cdots, y^{t+1}$ by:

- y^1 is a gridpoint of the triangulation,
- $P(T) = (p_1, p_2, \cdots, p_t)$ is a permutation of the elements of T with $|T| = t$, $T \subset \{0, 1, \cdots, n\}$,
- $y^{i+1} = y^i + \delta \cdot u(p_i)$, $1 \leqq i \leqq t$.

Now the algorithm generates for varying T a sequence of simplices such that:

1. Each t-dimensional simplex has at least t different labels.

2. Each t-simplex $\sigma(y^1, \cdots, y^{t+1})$ satisfies $y^1 = v + \delta \cdot \sum_{j=0}^n R_j u(j)$ where v is the starting point and R is an $(n+1)$ vector with $R_j = 0$ if $j \notin T$, $R_j \geqq 0$ if $j \in T$ $(0 \leqq j \leqq n)$.

The algorithm is given in a flow diagram in Fig. 3. In each cycle of the algorithm a t-dimensional simplex σ is "flipped" to its successor, which is determined by having a face in common containing all different labels present in σ. So if σ has t different labels, σ has exactly two such faces, one in common with its predecessor and one wth its successor. If σ has $t+1$ different labels $(t < n)$ its successor is a $(t+1)$-dimensional simplex with σ as one face. The dimension of σ can be decreased to prevent one component of R becoming negative. During these steps σ will never leave C_λ due to the (with the labels of σ related) positiveness of R together with the proper labelling. In this way each simplex has a unique predecessor and successor in C_λ except for the starting point (a zero dimensional simplex) which only has a successor and cannot be entered again. Because there is only a finite number of simplices in C_λ the algorithm must end, which only occurs by finding a completely labelled simplex, thus yielding an approximation of the solution of the LCP. For the various proofs see [v. d. Laan, '80a] or [v. d. Laan, '80b].

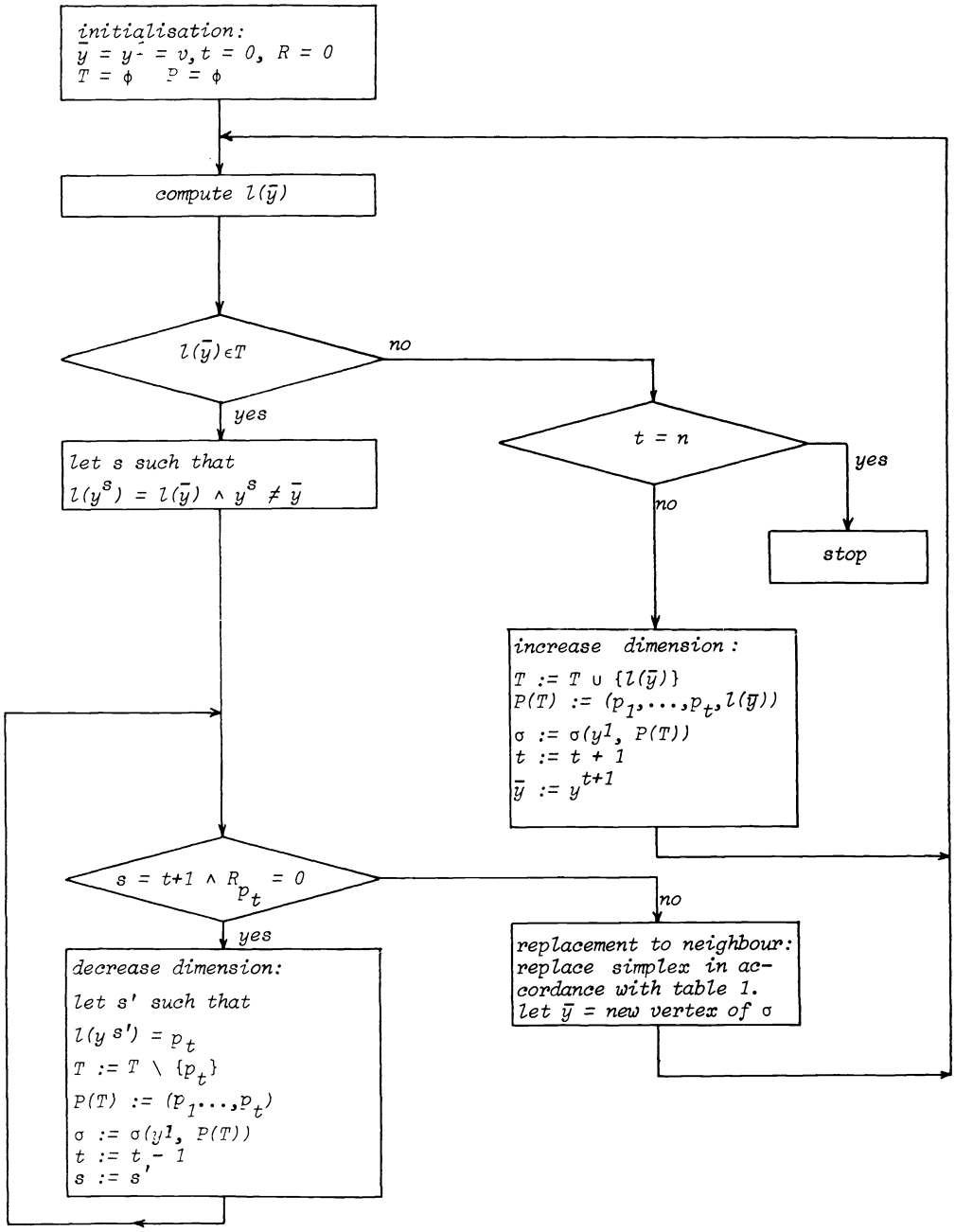


FIG. 3. *Flow diagram of simplex algorithm.*

The flow of the algorithm for $n = 2$ is demonstrated in Fig. 4. Note the proper labelling.

About this method to solve the LCP the following remarks must be applied.

- Full advantage can be taken of sparse matrix techniques, because only matrix-vector products have to be computed.
- Like the pivoting algorithms, the simplex algorithms are combinatorial algorithms and they converge for a relatively large class of matrices.

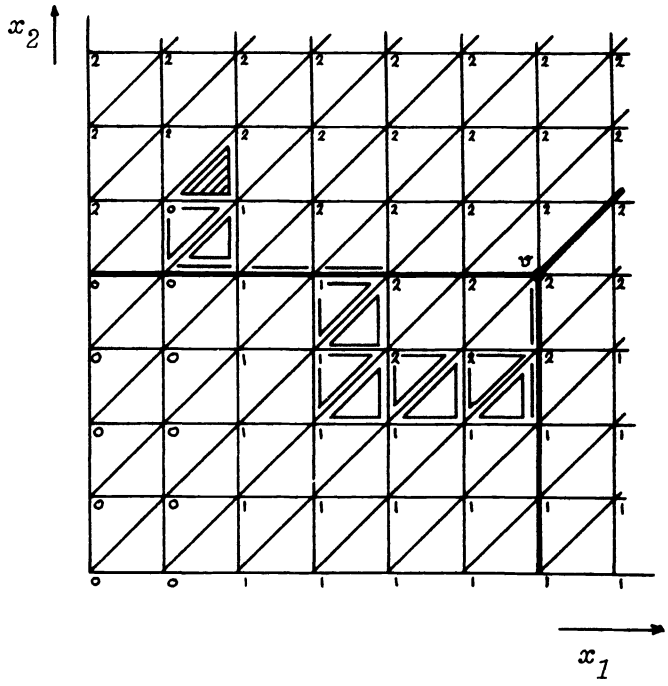


FIG. 4. *Flow of the algorithm for* $n = 2$.

- The algorithm can also be used for matrices that do not belong to SSM. The algorithm will either stop with a solution or generate a chain of simplices towards infinity and thus will not end.
- The algorithm will find a solution if $M \in$ SSM with each starting point and with each grid size. Therefore the approximate solution can be used as initial point in a next run with a smaller grid size.
- Where we want to solve a sequence of LCP's, we can choose the starting point to be the solution to the previous problem.

TABLE 1
Table for the replacement of the simplex. s is the number of the vertex to be replaced.

	y^1 becomes	$P(T)$ becomes	R
$s = 1$	$y^1 + \delta \cdot u(p_1)$	$(p_2, \cdots, p_t, p_1)$	$R_{p_1} := R_{p_1} + 1$
$2 \leqq s \leqq t$	no changes	$(p_1, \cdots, p_{s-2}, p_s, p_{s-1}, p_{s+1}, \cdots, p_t)$	no changes
$s = t+1$	$y^1 - \delta \cdot u(p_t)$	$(p_t, p_1, \cdots, p_{t-1})$	$R_{p_t} := R_{p_t} - 1$

5. Conclusions. A search is made of suitable algorithms to solve the LCP, characterized by a large sparse matrix, not necessarily restricted to some specific class. A minimization method is developed that can process matrices with a good sparse matrix behaviour. Convergence is proved for all class P matrices.

Pivoting algorithms are combinatorial methods which converge for a relatively large class of matrices. However, these algorithms lead to complex sparse matrix processing.

A good alternative is found in a simplex method. Good sparse matrix behaviour is obtained and convergence is proved for a large class of matrices (SSM), independent of starting point and grid size. Furthermore, any matrix can be processed but convergence is not assured. This method could be a good alternative for the pivoting algorithms when the dimension of the problem increases or pivoting does not find a solution of the LCP.

Appendix.

LEMMA 1.

$$M \in \mathrm{SSM} \to \exists \varepsilon > 0 \forall x \in \mathbb{R}^n_+, x \neq 0 \exists k, 1 \leqq k \leqq n \quad \frac{x_k \cdot (Mx)_k}{\|x\|^2} \geqq \varepsilon.$$

Proof. Let $U = \{x \in \mathbb{R}^n_+ \mid \|x\| = 1\}$ and for all $x \in U$: $h(x) = \max_i x_i \cdot (Mx)_i$, $1 \leqq i \leqq n$. U is a compact set and $h(x)$ is continuous and thus has a minimum ε and U [Griffiths, '70 pp. 427]. Then from the definition of SSM, ε will be positive. Thus:

$$\forall x \in \mathbb{R}^n_+, \|x\| = 1 \exists k, 1 \leqq k \leqq n \quad x_k \cdot (Mx)_k \geqq \varepsilon.$$

Now substitute $y \in \mathbb{R}^n_+$, $y \neq 0$, $y = ax$, $a = \|y\|$ which leads to:

$$\forall y \in \mathbb{R}^n_+, y \neq 0 \exists k, 1 \leqq k \leqq n \quad \frac{y_k \cdot (My)_k}{\|y\|^2} \geqq \varepsilon. \qquad \square$$

Acknowledgment. The author is indebted to the referees for clarifying the text and mentioning related work, thus improving the paper.

REFERENCES

[van Bokhoven, '80] W. M. G. VAN BOKHOVEN, *Macromodelling and simulation of mixed analog-digital networks by a piecewise-linear system approach*, Proc. International Conference on Circuits and Computers, 1980, pp. 361-365.

[van Bokhoven, '81] ———, *Piecewise linear modelling and analysis*, Kluwer, Deventer, The Netherlands, 1981, or, Ph.D. dissertation, Eindhoven University of Technology, Eindhoven, The Netherlands.

[Cottle, '68] R. W. COTTLE AND G. B. DANTZIG, *Complementary pivot theory of mathematical programming*, Lin. Algebra and Appl., 1 (1968), pp. 103-125.

[Cryer, '71] C. W. CRYER, *The solution of a quadratic programming problem using systematic overrelaxation*, this Journal, 9 (1971), pp. 385-392.

[Eaves, '71] C. B. EAVES, *The linear complementarity problem*, Management Sci., 17 (1971), pp. 612-634.

[van Eijndhoven, '82] J. T. J. VAN EIJNDHOVEN AND J. A. G. JESS, *The solution of large piecewise linear systems*, Proc. IEEE, International Symposium on Circuits and Systems, Rome, Italy, May 10-12, 1982, pp. 597-600.

[van Eijndhoven, '84] J. T. J. VAN EIJNDHOVEN, *A piecewise linear simulator for large scale integrated circuits*, Ph.D. dissertation, December 1984, Eindhoven University of Technology, Eindhoven, The Netherlands.

[Fujisawa, '72] T. FUJISAWA, E. S. KUH AND T. OHTSUKI, *A sparse matrix method for analysis of piecewise linear resistive networks*, IEEE Trans. Circuit Theory, CT-19 (1972), pp. 571-584.

[Griffiths, '70] H. B. GRIFFITHS AND P. J. HILTON, *A Comprehensive Textbook of Classical Mathematics*, Van Nostrand-Reinhold, London, New York, 1970.

[van der Heyden, '80] L. VAN DER HEYDEN, *A variable dimension algorithm for the linear complementarity problem*, Math. Programming, 19 (1980), pp. 328-346.

[Karamardian, '72] S. KARAMARDIAN, *The complementarity problem*, Math. Programming, 2 (1972) pp. 107-129.

[v. d. Laan '80a] G. VAN DER LAAN, *Simplicial fixed point algorithms*, Mathematical Centre, Amsterdam, 1980.

[v. d. Laan '80b] G. VAN DER LAAN AND A. J. J. TALMAN, *Convergence and properties of recent variable dimension algorithms*, in Numerical Solution of Highly Nonlinear Problems, W. Forster, ed., North-Holland, Amsterdam, New York, 1980.

[Lemke, '65] C. E. LEMKE, *Bimatrix equilibrium points and mathematical programming*, Management Sci., 11 (1965), pp. 681-689.

[Lemke, '70] ———, *Complementarity problems*, Nonlinear Programming, Proceeding of a Symposium, Rosen, Mangasarian and Ritter, eds., Academic Press, New York, London, 1970.

[Mangasarian, '77] O. L. MANGASARIAN, *Solution of symmetric linear complementarity problems by interactive methods*, J. Optim., Theory Appl., 22 (1977), pp. 465-485.

[Saigal, '81] R. SAIGAL, *A homotopy for solving large, sparse and structured fixed point problems*, Northwestern University, Evanston, IL, January 1981.

[Talman, '80] A. J. J. TALMAN, *Variable dimension fixed point algorithms and triangulations*, Mathematical Centre, Amsterdam, 1980.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 5, September 1986

THE MATCHING OF NONLINEAR MODELS VIA DYNAMIC STATE FEEDBACK*

MARIA DOMENICA DI BENEDETTO† AND ALBERTO ISIDORI†

Abstract. It is well known that, for linear systems, the model matching problem is equivalent to a disturbance decoupling problem with disturbance measurement. The solution of both problems can be expressed in terms of properties of invariant subspaces of the system and of the model.

In this paper, it is shown that, for nonlinear systems, under appropriate hypotheses, analogous results can be obtained. The solvability of a model matching problem can be expressed in terms of properties of suitable invariant distributions. It is also shown that, as for linear systems, those properties are related to the so-called "structures at infinity" of the system and of the model.

Key words. nonlinear control systems, model matching, dynamic state feedback

AMS(MOS) subject classifications. 93, also 93B, 93C

1. Introduction. In this paper we investigate the problem of designing, for a nonlinear system, a compensating control such that the resulting input-output behavior exactly matches that of a prespecified nonlinear model.

The problem of compensating a linear multivariable system in order to match a prescribed linear model has been solved in different forms by several authors in the last decades. Moore and Silverman [1] gave a solution based on the use of Silverman's structure algorithm [2]. Morse [3] used geometric concepts and proposed a solution which involves the construction of a suitable controllability subspace. Later, Morse [4] and Emre and Hautus [5] pointed out the equivalence between model matching and disturbance decoupling. Recently, Malabre [6] showed that the conditions found by Morse in [3] may be expressed as the equality of suitable "structures at infinity."

More recently, the problem of compensating a nonlinear system in order to obtain a linear input-output behavior has been investigated in [7], [8]. In particular, a solution to the problem of matching a prescribed linear model has been found [8], which somehow extends the one proposed by Silverman.

In a recent paper [9], the problem of matching a nonlinear model is set, by definition, as a disturbance decoupling problem. The corresponding solution may thus be found in terms of earlier results on nonlinear decoupling (see e.g. [11]). A similar approach is followed in [10].

In this paper we investigate the nonlinear version of Morse's approach. For this purpose, we make extensive use of the differential geometric concepts introduced in [11]. We also discuss the nonlinear version of Malabre's results, using the recent Nijmeijer and Schumacher's definition for the "structure at infinity" of a nonlinear system [12].

2. Problem statement. Consider a fixed nonlinear *plant P*, described by equations of the form

$$\dot{x} = f(x) + g(x)u, \tag{2.1a}$$

$$y = h(x), \tag{2.1b}$$

* Received by the editors March 4, 1985, and in revised form July 26, 1985. Portions of this paper appear in The 23rd IEEE Conference on Decision and Control, December 12-14, 1984, Las Vegas, Nevada, pp. 416-420.

† Department of Systems and Computer Science, University of Rome "La Sapienza", Rome, Italy.

with state $x \in X \subset \mathbb{R}^n$, input $u \in \mathbb{R}^m$ and output $y \in \mathbb{R}^p$. f and the m columns $g_1, \cdots, g_m$ of the matrix g are real analytic vector fields on $\mathbb{R}^n$ and h is a real analytic function.

In addition, suppose a *model* M is given, described by equations of the form

$$\dot{x}_M = f_M(x_M) + g_M(x_M)u_M, \tag{2.2a}$$

$$y_M = h_M(x_M), \tag{2.2b}$$

with state $x_M \in X_M \subset \mathbb{R}^{n_M}$, input $u_M \in \mathbb{R}^{m_M}$ and output $y_M \in \mathbb{R}^p$, and real analytic f_M, g_M, h_M. The problem of interest is to find a compensator Q for the plant P so that the resulting closed loop system displays the same input-output behavior as the model M.

The compensator Q used to control P consists of a dynamical system with inputs x and u_M and output u, described by equations of the form:

$$\dot{z} = a(z, x) + b(z, x)u_M, \tag{2.3a}$$

$$u = c(z, x) + d(z, x)u_M, \tag{2.3b}$$

with state $z \in Z \subset \mathbb{R}^\nu$ and real analytic a, b, c, d. The composition $P \circ Q$ of (2.1) and (2.3) is clearly a new dynamical system with the same structure as (2.1).

In the case of linear systems, the objective of model matching synthesis is to design a compensator such as to impose the coincidence between the transfer function of the model and that of the compensated plant. In the case of nonlinear systems, where the input-output behavior is usually described in terms of Volterra series expansions, the object of the model matching is to impose the coincidence of the associated Volterra kernels.

To be precise, let us recall that the output $y(t)$ of any nonlinear system of the form (2.1) may be expanded as

$$y(t) = w_0(t, x) + \sum_{i=1}^{m} \int_0^t w_i(t, \tau_1, x)u_i(\tau_1)\, d\tau_1 + \sum_{i_1, i_2 = 1}^{m} \int_0^t \int_0^{\tau_1} w_{i_1 i_2}(t, \tau_1, \tau_2, x)u_{i_1}(\tau_1)u_{i_2}(\tau_2)\, d\tau_1\, d\tau_2 + \cdots$$

where x is the initial state at time $t = 0$. Let $w^M_{j_1 \cdots j_i}(t, \tau_1, \cdots, \tau_i, x_M)$ denote the $(j_1 \cdots j_i)$th kernel of model M and $w^{P \circ Q}_{j_1 \cdots j_i}(t, \tau_1, \cdots, \tau_i, (x, z))$ the $(j_1 \cdots j_i)$th kernel of the compensated plant $P \circ Q$. Since $w^M_{j_1 \cdots j_i}$ depends on the initial state x_M of M and $w^{P \circ Q}_{j_1 \cdots j_i}$ on the initial state (x, z) of $P \circ Q$, when imposing the coincidence between these kernels one must specify how x_M and (x, z) are chosen. Depending on this choice, one may formulate different matching problems.

In what follows, we ask the solution Q of the model matching problem to be such that for each initial state x of P and each initial state x_M of M, there exists an initial state z of Q with the property that the Volterra kernels $w^M_{j_1 \cdots j_i}$, evaluated at x_M, and $w^{P \circ Q}_{j_1 \cdots j_i}$, evaluated at (x, z), coincide, for all $i \geqq 1$ and $1 \leqq j_i \leqq m_M$. In other words, each input-output behavior of the model is reproduced from any initial state of the process, provided that the compensator is set in a suitable initial state. Note that we are not requiring the "zero-input" terms w_0^M and $w_0^{P \circ Q}$ to be the same, thus following a practice in use for the linear model matching problem.

In general a global solution (i.e. with a, b, c, d defined on the whole $Z \times X$) may be hard to find. So we restrict ourselves to local solutions, i.e. defined on an open subset of $Z \times X$. In this way we arrive at the following formal statement.

Nonlinear model matching problem (MMP). Given a plant $P = (f, g, h)$, a model $M = (f_M, g_M, h_M)$ and a point $(x^0, x^0_M) \in X \times X_M \subset \mathbb{R}^n \times \mathbb{R}^{n_M}$, find neighborhoods U of

x^0 and U_M of x_M^0, an integer ν, an open subset V of $Z \subset \mathbb{R}^\nu$, a compensator $Q = (a, b, c, d)$ with a, b, c, d real analytic functions defined on $V \times U$, a map $F: U \times U_M \to V$, with the property that

$$w_{j_1 \cdots j_i}^{P \circ Q}(t, \tau_1, \cdots, \tau_i, (x, F(x, x_M))) = w_{j_1 \cdots j_i}^{M}(t, \tau_1, \cdots, \tau_i, x_M)$$

for all $i \geqq 1$, for all $1 \leqq j_i \leqq m_M$ and for all (x, x_M) in $U \times U_M$.

3. An associated disturbance decoupling problem. In this section, we show that the existence of a solution of a given model matching problem is implied by that of a suitable associated disturbance decoupling problem. To this end, we recall that for a nonlinear system of the form

$$\dot{\hat{x}} = \hat{f}(\hat{x}) + \hat{g}(\hat{x})\hat{u} + \hat{p}(\hat{x})\hat{w}, \qquad \hat{y} = \hat{h}(\hat{x}), \tag{3.1}$$

with input $\hat{u}$, disturbance $\hat{w}$ and output $\hat{y}$, the *disturbance decoupling problem with disturbance measurement* is defined in the following way: given a plant $\hat{P} = (\hat{f}, \hat{g}, \hat{h})$, a point $\hat{x}^0$, find a neighborhood $\hat{U}$ of $\hat{x}^0$ and a pair of real analytic functions $\hat{\alpha}$ and $\hat{\gamma}$ defined on $\hat{U}$, with the property that the control law

$$\hat{u} = \hat{\alpha}(\hat{x}) + \hat{\gamma}(\hat{x})\hat{w} \tag{3.2}$$

decouples the output $\hat{y}$ from the disturbance $\hat{w}$. Note that this control mode includes feedback on the state $\hat{x}$ and feedforward on the disturbance $\hat{w}$.

It is known [11] that the disturbance $\hat{w}$ does not influence the output $\hat{y}$ of the composite system (3.1)-(3.2), for every initial state in $\hat{U}$, if and only if the functions $\hat{\alpha}$ and $\hat{\gamma}$ are such as to make the conditions

$$L_{(\hat{g}\hat{\gamma}+\hat{p})} L_{(\hat{f}+\hat{g}\hat{\alpha})}^k \hat{h}(\hat{x}) = 0 \tag{3.3}$$

satisfied for all $\hat{x} \in \hat{U}$ and for all $k \geqq 0$. As a matter of fact, if and only if these conditions are satisfied, the Volterra series expansion of $\hat{y}$ in the composite system (3.1)-(3.2) reduces to the "zero-input" term alone.

In what follows we associate with a given model matching problem a disturbance decoupling problem with disturbance measurement (abbreviated ADDPdm) in this way. We set in (3.1)

$$\hat{x} = (x, x_M), \quad \hat{u} = u, \quad \hat{w} = u_M,$$

and

$$\hat{f}(\hat{x}) = \begin{pmatrix} f(x) \\ f_M(x_M) \end{pmatrix}, \quad \hat{g}(\hat{x}) = \begin{pmatrix} g(x) \\ 0 \end{pmatrix}, \quad \hat{p}(\hat{x}) = \begin{pmatrix} 0 \\ g_M(x_M) \end{pmatrix}, \tag{3.4a}$$

$$\hat{h}(\hat{x}) = h(x) - h_M(x_M). \tag{3.4b}$$

The following result is straightforward.

LEMMA 3.1. *A model matching problem is solvable if the associated disturbance decoupling problem with disturbance measurement is.*

Proof. Suppose the ADDPdm is solved at $\hat{x}^0 = (x^0, x_M^0)$. Then there exist α and γ defined locally around $\hat{x}^0$ such that, in the system

$$\begin{aligned} \dot{x} &= f(x) + g(x)\alpha(x, x_M) + g(x)\gamma(x, x_M)u_M, \\ \dot{x}_M &= f_M(x_M) + g_M(x_M)u_M, \\ \hat{y} &= h(x) - h_M(x_M), \end{aligned}$$

the input u_M does not influence the output $\hat{y}$ for every possible initial state (x, x_M) in

a neighborhood of (x^0, x_M^0). This means that, in the system

$$\dot{x} = f(x) + g(x)\alpha(x, x_M) + g(x)\gamma(x, x_M)u_M,$$
$$\dot{x}_M = f_M(x_M) + g_M(x_M)u_M,$$
$$y = h(x),$$

initialized at (x, x_M), the output $h(x(t))$ equals the output $h_M(x_M(t))$ of the model initialized at x_M, modulo a "zero-input" term.

The last system can be viewed as the process P, initialized at x, composed with a controller Q defined by the equations

$$\dot{z} = f_M(z) + g_M(z)u_M,$$
$$u = \alpha(x, z) + \gamma(x, z)u_M,$$

initialized at $z = x_M$. □

4. Main results. In what follows the possibility of solving a MMP will be expressed in terms of properties of certain (f, g)-invariant distributions.

For convenience, we summarize hereafter some basic facts about nonlinear geometric control theory and we quote, without proof, a series of properties of interest (see [11] for details). We recall that, for the control system (2.1), a distribution Δ is said to be (f, g)-invariant (or controlled invariant) if there exists a nonlinear feedback $u = \alpha(x) + \beta(x)v$, where α and β are defined on X, such that the modified dynamics

$$\dot{x} = f(x) + g(x)\alpha(x) + g(x)\beta(x)v$$

leaves Δ invariant, i.e. such that

$$[f + g\alpha, \Delta](x) \subset \Delta(x), \tag{4.1a}$$

$$[(g\beta)_i, \Delta](x) \subset \Delta(x) \quad \text{for all } 1 \leqq i \leqq m, \tag{4.1b}$$

for all $x \in X$. Conditions (4.1) are sometimes abbreviated as

$$[f + g\alpha, \Delta] \subset \Delta,$$
$$[g\beta, \Delta] \subset \Delta.$$

It is known that, if β is nonsingular,

$$[f, \Delta] \subset \Delta + \mathcal{G}, \tag{4.2a}$$

$$[g, \Delta] \subset \Delta + \mathcal{G}, \tag{4.2b}$$

where $\mathcal{G}$ denotes the distribution

$$\mathcal{G} = \text{span}\{g_1, \cdots, g_m\}.$$

Conversely, if Δ is an involutive distribution of constant dimension and also $\mathcal{G}$ and $\Delta + \mathcal{G}$ have constant dimension, then (4.2) imply the existence of *local* feedback functions α and β such that (4.1) are satisfied.

Let dh denote the codistribution spanned by the differentials of the entries of h, i.e.

$$dh = \text{span}\{dh_1, \cdots, dh_p\}.$$

Moreover, let $\mathbb{I}(f, g, (dh)^\perp)$ denote the class of all distributions satisfying (4.2) and contained in $(dh)^\perp$ and Δ^* its unique maximal element.

The computation of Δ^* is usually performed by means of the so-called Controlled Invariant Distribution Algorithm. With the triplet (f, g, h) one associates a sequence of codistributions defined in the following way:

$$\Omega_0 = dh,$$
$$\Omega_k = \Omega_{k-1} + L_f(\Omega_{k-1} \cap \mathcal{G}^\perp) + \sum_{i=1}^{m} L_{g_i}(\Omega_{k-1} \cap \mathcal{G}^\perp). \tag{4.3}$$

This sequence is clearly increasing and, if $\Omega_{k^*} = \Omega_{k^*+1}$ for some k^*, then $\Omega_k = \Omega_{k^*}$ for all $k > k^*$.

For practical purposes, we shall henceforth assume that the codistributions involved in this algorithm have constant dimension around the point of interest x^0. More precisely, we say that the point x^0 is a *regular point* for the algorithm (2.1) if for all x in a neighborhood of x^0:

(i) the dimension of $\mathcal{G}$ is constant,

(ii) the dimension of Ω_k is constant, for all $k \geqq 0$,

(iii) the dimension of $(\Omega_k \cap \mathcal{G}^\perp)$ is constant, for all $k \geqq 0$.

Note that if x^0 is a regular point for the algorithm (4.3), then there exists an integer $k^* < n$ such that $\Omega_{k^*} = \Omega_{k^*+1}$ and this, as we have seen, implies the convergence of the algorithm (4.3), in a neighborhood of x^0, in a finite number of steps. Moreover, (see [11])

$$\Delta^* = \Omega_{k^*}^\perp.$$

If x^0 is a regular point for the algorithm (4.3), the codistributions Ω_k, for all $k \geqq 0$, are spanned by exact one-forms and, in particular, Δ^* is involutive. The one-forms spanning Ω_k can be recursively computed via an algorithm proposed by Krener [14] and described in the Appendix.

We may now return to the model matching problem. In the following statement, for $\hat{f}$, $\hat{g}$, $\hat{p}$ and $\hat{h}$ we mean the vector fields and the function defined on the right-hand side of (3.4a) and (3.4b), and we set

$$\hat{\mathcal{P}} = \text{span}\{\hat{p}_1, \cdots, \hat{p}_{m_M}\}.$$

THEOREM 4.1. *Let (x^0, x_M^0) be a regular point for the algorithm* (4.3) *for the triplet $(\hat{f}, \hat{g}, \hat{h})$. Let $\hat{\Delta}^*$ be the unique maximal element of $\mathbb{I}(\hat{f}, \hat{g}, (d\hat{h})^\perp)$. If*

$$\hat{\mathcal{P}} \subset \hat{\Delta}^* + \hat{\mathcal{G}} \tag{4.4}$$

in a neighborhood of (x^0, x_M^0), then the MMP *is solvable.*

Proof. As a consequence of the assumptions, the distribution $\hat{\Delta}^*$ is such that:

(i) $\hat{\Delta}^* \subset (d\hat{h})^\perp$;

and, locally around (x^0, x_M^0),

(ii) there exists a function $\hat{\alpha}$ such that:

$$[\hat{f} + \hat{g}\hat{\alpha}, \hat{\Delta}^*] \subset \hat{\Delta}^*;$$

(iii) there exists a function $\hat{\gamma}$ such that:

$$\hat{g}\hat{\gamma} + \hat{p} \in \hat{\Delta}^*$$

(because of (4.4)).

The existence of a distribution which satisfies (i), (ii) and (iii) proves that the conditions (3.3) are satisfied. Then, as a consequence of Lemma 3.1, the result follows. □

In the case of linear systems the condition expressed by this theorem is also necessary for the solvability of an MMP. In the present case of nonlinear systems the proof of the necessity requires some further assumption. As we shall see later on, this is essentially due to the fact that, for nonlinear systems, the invariance condition (4.1a) *alone* does not imply (4.2a) and (4.2b).

In order to proceed with the proof of necessity of (4.4), we introduce some other notations. With the plant P and the model M we associate, in addition to the system (3.4), the system

$$\dot{x}^E = f^E(x^E) + g^E(x^E)u^E, \qquad y^E = h^E(x^E),$$

where $x^E = \hat{x}$ and

$$f^E = \hat{f}, \quad g^E = (\hat{g} \quad \hat{p}), \quad h^E = \hat{h}.$$

Moreover, we consider also a *dynamic extension* of the latter, defined by

$$\dot{\bar{x}} = \bar{f}(\bar{x}) + \bar{g}(\bar{x})\bar{u}, \qquad \bar{y} = \bar{h}(\bar{x}), \tag{4.5}$$

where $\bar{x} = (x^E, z)$,

$$\bar{f}(\bar{x}) = \begin{pmatrix} f^E(x^E) \\ 0 \end{pmatrix}, \quad \bar{g}(\bar{x}) = \begin{pmatrix} g^E(x^E) & 0 \\ 0 & I \end{pmatrix}, \quad \bar{h}(\bar{x}) = h^E(x^E),$$

and $\dim(z) = \nu$. Note that this new system is completely specified by the triplet (f^E, g^E, h^E), which incorporates all the data of the MMP, and by the integer ν.

It is easy to see that the *difference* between the output of the composed system $P \circ Q$ and that of the model M may be interpreted as the output of the system (4.5) subject to a static state-feedback of the form

$$\bar{u} = \bar{\alpha}(\bar{x}) + \bar{\beta}(\bar{x})u_M$$

with

$$\bar{\alpha}(\bar{x}) = \begin{pmatrix} c(z, x) \\ 0 \\ a(z, x) \end{pmatrix}, \qquad \bar{\beta}(\bar{x}) = \begin{pmatrix} d(z, x) \\ I \\ b(z, x) \end{pmatrix}.$$

Thus, if the controller Q solves the MMP, the above feedback is such as to make the output $\bar{y}$ of the system

$$\begin{aligned} \dot{\bar{x}} &= \bar{f}(\bar{x}) + \bar{g}(\bar{x})\bar{\alpha}(\bar{x}) + \bar{g}(\bar{x})\bar{\beta}(\bar{x})u_M, \\ \bar{y} &= \bar{h}(\bar{x}), \end{aligned} \tag{4.6}$$

independent of u_M, whenever the initial state is chosen as

$$\bar{x} = (x, x_M, F(x, x_M)). \tag{4.7}$$

To prove a partial "converse" of Theorem 4.1, we need to consider one of the following assumptions:

(A1) The triplet (f^E, g^E, h^E) is such that, for all $\nu \geqq 1$ and for all α

$$\bigcap_{k \geqq 0} \ker dL^k_{\bar{f}+\bar{g}\alpha}\bar{h} \subset \bar{\Delta}^* \tag{4.8}$$

where $\bar{\Delta}^*$ is the maximal element of $\mathbb{I}(\bar{f}, \bar{g}, (d\bar{h})^\perp)$.

(A2) The triplet (f^E, g^E, h^E) is such that

$$\sum_{i=1}^{m+m_M} L_{g_i^E}(\Omega_k^E \cap \mathcal{G}^{E\perp}) \subset \Omega_k^E$$

for all $k \geqq 0$.

(A3) The model (f_M, g_M, h_M) is linear and the process (f, g, h) can be made linear, from an input-output point of view, via static state-feedback (see [7]).

A short comment on the relative importance of the above assumptions is in order.

LEMMA 4.2. (A3) *implies* (A2) *and* (A2) *implies* (A1).

Proof. (A3) clearly implies that (f^E, g^E, h^E) can be made linear, from an input-output point of view, via static state-feedback. This, in turn, implies (see [16])

$$\sum_{i=1}^{m+m_M} L_{g_i^E}(\Omega_k^E \cap \mathcal{G}^{E\perp}) \subset \Omega_k^E \cap \mathcal{G}^{E\perp}$$

i.e. (A2).

In order to prove that (A2) implies (A1), it is useful to note that, from the algorithm (4.3), one easily obtains

$$\bar{\Omega}_k = \Omega_k^E \times \{0\}$$

for all $k \geqq 0$. Thus, (A2) implies

$$\sum_{i=1}^{m+m_M+\nu} L_{\bar{g}_i}(\bar{\Omega}_k \cap \bar{\mathcal{G}}^{\perp}) \subset \bar{\Omega}_k.$$

This, in turn, implies (A1) (see [17]). □

The distribution Δ on the left-hand side of (4.8) is invariant under $(\bar{f} + \bar{g}\alpha)$ and contained in $(d\bar{h})^{\perp}$. As a matter of fact, it turns out to be the largest distribution invariant under $(\bar{f} + \bar{g}\alpha)$ and contained in $(d\bar{h})^{\perp}$. However, Δ does not necessarily belong to $\mathbb{I}(\bar{f}, \bar{g}, (d\bar{h})^{\perp})$ because, as we already noted, the inclusion (4.1a) alone does not imply both (4.2a) and (4.2b). For this to be true, we would need, e.g., the existence of a nonsingular matrix β such that (4.1b) holds also. In the present case, the solution of an MMP implies the existence of a distribution which is invariant under $\bar{f} + \bar{g}\bar{\alpha}$ and under $\bar{g}\bar{\beta}$, but where $\bar{\beta}$ is a singular matrix. Thus, this distribution does not necessarily belong to the class $\mathbb{I}(\bar{f}, \bar{g}, (d\bar{h})^{\perp})$. This motivates the introduction of the assumption (A1).

However, the assumption (A1) is not testable in practice because it involves a condition to be satisfied for all possible integers ν and functions α. One may wish to use the stronger assumption (A2) which is indeed testable because it involves only a property of the chain of codistributions Ω_k^E generated by means of the algorithm (4.3) for the triplet (f^E, g^E, h^E).

The assumption (A3) is even stronger but it is exactly the one considered in [8] for linear model matching. We mention it here in order to show that the result of [8] can be deduced as a special case of the more general results established in this paper.

THEOREM 4.3. *Suppose the* MMP *is solved by some controller Q. Let* (x^0, x_M^0) *be a regular point for the algorithm* (4.3) *for the triplet* (f^E, g^E, h^E). *Suppose* (A1) *holds. Then*

$$\hat{\mathcal{P}} \subset \Delta^{E*} + \hat{\mathcal{G}} \tag{4.9}$$

where Δ^{E*} *is the maximal element of* $\mathbb{I}(f^E, g^E, (dh^E)^{\perp})$.

Proof. Since, by assumption, the controller Q solves the MMP, the Volterra kernels of (4.6) vanish in the initial state (4.7). In particular, looking at the first-order kernel, this implies that

$$L_{(\bar{g}\bar{\beta})_i} L^k_{(\bar{f}+\bar{g}\bar{\alpha})} \bar{h}(x, x_M, F(x, x_M)) = 0 \tag{4.10}$$

for all $1 \leqq i \leqq m_M$, for all $k \geqq 0$ and for all $(x, x_M) \in U \times U_M$. It is immediate to see that (4.10) implies

$$(\bar{g}\bar{\beta})_i(\bar{x}) \in \bigcap_{k \geqq 0} \ker\,(dL^k_{(\bar{f}+\bar{g}\bar{\alpha})}\bar{h})(\bar{x}) \tag{4.11}$$

for all $1 \leqq i \leqq m_M$ and for all $\bar{x}$ in the subset M of $U \times U_M \times V$

$$M = \{(x, x_M, z) \in U \times U_M \times V \,|\, z = F(x, x_M)\}.$$

As a consequence of the assumption (4.8) it follows that

$$(\bar{g}\bar{\beta})_i(\bar{x}) \in \bar{\Delta}^*(\bar{x}) \tag{4.12}$$

for all $1 \leqq i \leqq m_M$ and for all $\bar{x} \in M$.

Let Q be the natural surjection of $X \times X_M \times Z$ on $X \times X_M$ and Q_* the corresponding differential. Since $\bar{\Delta}^* = \Delta^{E*} \times TZ$, we have

$$Q_* \bar{\Delta}^* = \Delta^{E*} \circ Q.$$

Moreover, from (4.12) we have also

$$Q_*(\bar{g}\bar{\beta})_i(\bar{x}) = \begin{pmatrix} g(x)d(z, x) \\ g_M(x_M) \end{pmatrix}_i \in Q_* \bar{\Delta}^*(\bar{x}) = \Delta^{E*} \circ Q(\bar{x})$$

for all $\bar{x} \in M$. Since $Q|_M$ is onto $U \times U_M$, from this condition one obtains

$$\operatorname{span}\left\{\begin{pmatrix} 0 \\ g_M \end{pmatrix}\right\} \subset \Delta^{E*} + \operatorname{span}\left\{\begin{pmatrix} g \\ 0 \end{pmatrix}\right\}$$

i.e. (see (3.4a))

$$\hat{\mathscr{P}} \subset \Delta^{E*} + \hat{\mathscr{G}}$$

and this completes the proof. □

At this point, the "necessity" of the condition (4.4) is a consequence of the following trivial result.

LEMMA 4.4. *The condition* (4.9) *implies the condition* (4.4).

Proof. Clearly

$$[\hat{f}, \Delta^{E*}] \subset \Delta^{E*} + \mathscr{G}^E = \Delta^{E*} + \hat{\mathscr{G}} + \hat{\mathscr{P}},$$
$$[\hat{g}, \Delta^{E*}] \subset \Delta^{E*} + \mathscr{G}^E = \Delta^{E*} + \hat{\mathscr{G}} + \hat{\mathscr{P}}.$$

If (4.9) holds, then

$$[\hat{f}, \Delta^{E*}] \subset \Delta^{E*} + \hat{\mathscr{G}},$$
$$[\hat{g}, \Delta^{E*}] \subset \Delta^{E*} + \hat{\mathscr{G}},$$

i.e. $\Delta^{E*} \in \mathbb{I}(\hat{f}, \hat{g}, (d\hat{h})^\perp)$. Thus $\Delta^{E*} \subset \hat{\Delta}^*$ and (4.4) follows. □

5. A link with the structure at infinity. In this section we show that the condition (4.9) may be reformulated in terms of properties of the structures at infinity of the triplet (f, g, h) and the triplet (f^E, g^E, h^E). Following Nijmeijer and Schumacher [12] the structure at infinity of a triplet (f, g, h) may be defined as the list of integers

$$r_k = \dim\,(\mathscr{G}^\perp + \Omega_k) - \dim\,(\mathscr{G}^\perp), \qquad k \geqq 0$$

where $\Omega_0, \Omega_1, \cdots, \Omega^*$ is the sequence of codistributions generated by means of the

algorithm (4.3):

$$\Omega_0 = dh,$$

$$\Omega_k = \Omega_{k-1} + \sum_{i=0}^{m} L_{g_i}(\mathscr{G}^{\perp} \cap \Omega_{k-1}),$$

where $g_0 := f$.

In the same way, one can define a list of integers r_k^E, $k \geqq 0$, as the structure at infinity of the triplet (f^E, g^E, h^E), by considering the sequence

$$\Omega_0^E = dh^E,$$

$$\Omega_k^E = \Omega_{k-1}^E + \sum_{i=0}^{m+m_M} L_{g_i^E}(\mathscr{G}^{E\perp} \cap \Omega_{k-1}^E),$$

where $g_0^E := f^E$, and setting

$$r_k^E = \dim(\mathscr{G}^{E\perp} + \Omega_k^E) - \dim(\mathscr{G}^{E\perp}), \qquad k \geqq 0.$$

Clearly, these definitions make sense under the assumption that all the codistributions involved have constant dimension. We shall thus always assume, in what follows, that the point x^0 and the point (x^0, x_M^0) are regular points.

The following statement extends a recent result by Malabre [6].

THEOREM 5.1. *Suppose the sequence r_k and r_k^E, $k \geqq 0$, are defined. The condition* (4.9) *is satisfied if and only if*

$$r_k = r_k^E \tag{5.1}$$

for all $k \geqq 0$.

Proof. Recall that $\Delta^{E*} = (\Omega^{E*})^{\perp}$, so the condition (4.9) may be rewritten as

$$\hat{\mathscr{P}}^{\perp} \supset \Omega^{E*} \cap \hat{\mathscr{G}}^{\perp}. \tag{5.2}$$

Since the sequence of the Ω_k's is nondecreasing, (5.2) is equivalent to

$$\hat{\mathscr{P}}^{\perp} \supset \Omega_k^E \cap \hat{\mathscr{G}}^{\perp} \quad \text{for all } k \geqq 0. \tag{5.3}$$

This, in turn, is equivalent to

$$\hat{\mathscr{G}}^{\perp} \cap \Omega_k^E = \mathscr{G}^{E\perp} \cap \Omega_k^E \quad \text{for all } k \geqq 0 \tag{5.4}$$

because $\hat{\mathscr{G}}^{\perp} \cap \hat{\mathscr{P}}^{\perp} = \mathscr{G}^{E\perp}$.

Therefore, what we need is to prove the necessity and sufficiency of (5.4). To this end, we need a preliminary result. In what follows, we identify the Ω_k with codistributions on $X \times X_M$ (by taking, at $(x, x_M) \in X \times X_M$, the subspace of $T_x^*X \times T_{x_M}^*X_M$ spanned by the pairs $(\omega, 0)$ with $\omega \in \Omega_k(x)$). It is easy to realize that the above codistributions satisfy the recursions:

$$\Omega_0' = dh \times \{0\},$$

$$\Omega_k' = \Omega_{k-1}' + \sum_{i=0}^{m} L_{\hat{g}_i}(\mathscr{G}^{\perp} \cap \Omega_{k-1}'),$$

where $\hat{g}_0 := \binom{f}{0}$.

Moreover, let Γ denote a codistribution on $X \times X_M$ defined by taking

$$\Gamma(x, x_M) = \{0\} \times T_{x_M}^* X_M.$$

These objects will be used in the sequel.

LEMMA 5.2. *If, for some* $k \geqq 0$,

$$\hat{\mathcal{G}}^{\perp} \cap \Omega_k^E = \mathcal{G}^{E\perp} \cap \Omega_k^E \tag{5.5}$$

and

$$\Omega'_k = \Omega_k^E \qquad (\text{mod } \Gamma) \tag{5.6}$$

then

$$\Omega'_{k+1} = \Omega_{k+1}^E \qquad (\text{mod } \Gamma). \tag{5.7}$$

Proof. By definition, the codistributions Ω'_k defined by the algorithm (4.3) are locally spanned by one-forms which do not depend on x_M. Clearly $\hat{\mathcal{G}}^{\perp}$ is spanned by one-forms which do not depend on x_M. Thus $\hat{\mathcal{G}}^{\perp} \cap \Omega'_k$, for $k \geqq 0$, has the same property.

Moreover, since $\Gamma \subset \hat{\mathcal{G}}^{\perp}$, from (5.6) we get

$$\hat{\mathcal{G}}^{\perp} \cap \Omega'_k + \Gamma = \hat{\mathcal{G}}^{\perp} \cap (\Omega'_k + \Gamma) = \hat{\mathcal{G}}^{\perp} \cap (\Omega_k^E + \Gamma) = \hat{\mathcal{G}}^{\perp} \cap \Omega_k^E + \Gamma. \tag{5.8}$$

Thus

$$\begin{aligned}
\Omega'_{k+1} + \Gamma &= \Omega'_k + \sum_{i=0}^{m} L_{\hat{g}_i}(\hat{\mathcal{G}}^{\perp} \cap \Omega'_k) + \Gamma \\
&= \Omega'_k + \sum_{i=0}^{m+m_M} L_{g_i^E}(\hat{\mathcal{G}}^{\perp} \cap \Omega'_k) + \Gamma \\
&= \Omega'_k + \sum_{i=0}^{m+m_M} L_{g_i^E}(\hat{\mathcal{G}}^{\perp} \cap \Omega'_k + \Gamma) + \Gamma \\
&= \Omega'_k + \sum_{i=0}^{m+m_M} L_{g_i^E}(\hat{\mathcal{G}}^{\perp} \cap \Omega_k^E + \Gamma) + \Gamma \quad \text{by (5.8)} \\
&= \Omega'_k + \sum_{i=0}^{m+m_M} L_{g_i^E}(\mathcal{G}^{E\perp} \cap \Omega_k^E + \Gamma) + \Gamma \quad \text{by (5.5)} \\
&= \Omega'_k + \sum_{i=0}^{m+m_M} L_{g_i^E}(\mathcal{G}^{E\perp} \cap \Omega_k^E) + \Gamma \\
&= \Omega_k^E + \sum_{i=0}^{m+m_M} L_{g_i^E}(\mathcal{G}^{E\perp} \cap \Omega_k^E) + \Gamma = \Omega_{k+1}^E + \Gamma \quad \text{by (5.6).}
\end{aligned}$$

This concludes the proof of the lemma. □

At this point it is possible to show that $r_k = r_k^E$, for all $k \geqq 0$, if and only if (5.4) is true, thus concluding the proof of the theorem.

(if). Note that (5.6) is true for $k = 0$. Then, since (5.5) is true for all $k \geqq 0$, (5.6) holds for all $k > 0$. As a consequence, since $\Gamma \subset \hat{\mathcal{G}}^{\perp}$,

$$\hat{\mathcal{G}}^{\perp} + \Omega'_k = \hat{\mathcal{G}}^{\perp} + \Omega'_k + \Gamma = \hat{\mathcal{G}}^{\perp} + \Omega_k^E + \Gamma = \hat{\mathcal{G}}^{\perp} + \Omega_k^E \tag{5.9}$$

for all $k \geqq 0$. It follows that

$$\begin{aligned}
r_k &= \dim (\Omega'_k + \hat{\mathcal{G}}^{\perp}) - \dim \hat{\mathcal{G}}^{\perp} \\
&= \dim (\Omega_k^E + \hat{\mathcal{G}}^{\perp}) - \dim \hat{\mathcal{G}}^{\perp} \qquad \text{(by (5.9))} \\
&= \dim \Omega_k^E - \dim (\Omega_k^E \cap \hat{\mathcal{G}}^{\perp}) \\
&= \dim \Omega_k^E - \dim (\Omega_k^E \cap \hat{\mathcal{G}}^{E\perp}) \\
&= r_k^E
\end{aligned}$$

(only if). Note that the equality $r_k = r_k^E$ may be rewritten

$$\dim (\Omega_k' + \hat{\mathcal{G}}^\perp) - \dim \hat{\mathcal{G}}^\perp = \dim (\Omega_k^E + \mathcal{G}^{E\perp}) - \dim \mathcal{G}^{E\perp}.$$

Suppose that (5.6) is true for some k. Then the following implications hold:

$$\begin{aligned}
\Omega_k' = \Omega_k^E (\text{mod } \Gamma) &\Rightarrow \Omega_k' + \hat{\mathcal{G}}^\perp = \Omega_k^E + \hat{\mathcal{G}}^\perp \quad (\text{since } \Gamma \subset \hat{\mathcal{G}}^\perp) \\
&\Rightarrow \dim (\Omega_k^E + \hat{\mathcal{G}}^\perp) - \dim \hat{\mathcal{G}}^\perp = \dim (\Omega_k^E + \mathcal{G}^{E\perp}) - \dim \mathcal{G}^{E\perp} \\
&\Rightarrow \dim (\Omega_k^E \cap \hat{\mathcal{G}}^\perp) = \dim (\Omega_k^E \cap \mathcal{G}^{E\perp}) \\
&\Rightarrow \Omega_k^E \cap \hat{\mathcal{G}}^\perp = \Omega_k^E \cap \mathcal{G}^{E\perp} \quad (\text{because } \mathcal{G}^{E\perp} \subset \hat{\mathcal{G}}^\perp)
\end{aligned}$$

i.e. (5.5) for k.

Since (5.6) is true for $k = 0$, (5.5) is also true for $k = 0$. Then, by Lemma 5.2, (5.6) is true for all k. As a consequence of the previous sequence of implications, we have that (5.5) is true for all k, i.e. (5.4) is true. □

Remark 5.3. The integers r_k which characterize the structure at infinity of a nonlinear system, may be computed as the ranks of appropriate matrices obtained via the algorithm proposed by Krener for the construction of the maximal element Δ^* of $\mathbb{I}(f, g, (dh)^\perp)$ and described in the Appendix.

This algorithm makes the computation of the r_k possible as follows. Using the notations introduced in the Appendix, it is immediately seen that:

$$\dim \Omega_k = s_k.$$

It is also clear that the intersection $\mathcal{G}^\perp \cap \Omega_k$ has a dimension equal to $s_k - \rho_k$. Thus

$$r_k = \dim \Omega_k - \dim (\Omega_k \cap \mathcal{G}^\perp) = s_k - (s_k - \rho_k) = \rho_k.$$

Using those properties, in a recent paper [19], the authors have proposed a procedure for the computation of the structure at infinity on the input-output data. □

Merging Theorem 5.1 with the results of the previous section, it is seen that if the regularity assumption holds and if (A1) is satisfied, then a necessary and sufficient condition for the MMP to be solvable is that $r_k = r_k^E$, for all k. It should be stressed that the latter condition is no longer necessary whenever (A1) is removed. This is illustrated by the following example.

Example 5.4. Let $X = \mathbb{R}^4$ and

$$f(x) = \begin{pmatrix} 0 \\ x_4 \\ 0 \\ 0 \end{pmatrix}, \quad g(x) = \begin{pmatrix} x_3 & 0 \\ 0 & 1 \\ 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad h(x) = \begin{pmatrix} x_2 - x_3 \\ x_1 \end{pmatrix}. \tag{5.10}$$

Let also $X_M = \mathbb{R}^4$ and

$$f_M(x_M) = \begin{pmatrix} x_{M2} \\ 0 \\ x_{M4} \\ 0 \end{pmatrix}, \quad g_M(x_M) = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix}, \quad h_M(x_M) = \begin{pmatrix} x_{M1} \\ x_{M3} \end{pmatrix}.$$

An easy computation shows that

$$r_0 = 1, \qquad r_k = 2 \quad \text{for } k \geqq 1$$

and

$$r_0^E = 1, \qquad r_k^E = 3 \quad \text{for } k \geqq 1,$$

so the equality (5.1) is not met. However, the MMP is solvable by the dynamic state-feedback

$$\begin{aligned} \dot{z} &= -\frac{z^2}{z+x_3} + \frac{1}{z+x_3}(-z \;\; 1)u_M, \\ u &= \begin{pmatrix} z \\ -\dfrac{z^2}{z+x_3} \end{pmatrix} + \frac{1}{z+x_3}\begin{pmatrix} 0 & 0 \\ x_3 & 1 \end{pmatrix} u_M. \end{aligned} \tag{5.11}$$

In fact it is rather simple to compute the Volterra kernels of (5.10) composed with (5.11), which have the following expression

$$w_1^{P\circ Q}(t, \tau, (x, z)) = \begin{pmatrix} t-\tau \\ 0 \end{pmatrix}, \qquad w_2^{P\circ Q}(t, \tau, (x, z)) = \begin{pmatrix} 0 \\ t-\tau \end{pmatrix}$$

and

$$w_{j_1 \cdots j_i}^{P\circ Q}(t, \tau_1, \cdots, \tau_i, (x, z)) = 0 \quad \text{if } i > 1.$$

Thus, this shows that the MMP is solved. □

6. Concluding remarks. The purpose of this paper was to investigate the problem of designing a dynamic compensator which makes it possible to match a prescribed input-output behavior in a given nonlinear plant. A simple geometric condition, which involves the computation of a suitable maximal controlled invariant distribution, was proved to be sufficient and, in some cases, necessary for the existence of a solution to the problem. Such a condition is the nonlinear version of a similar condition established by Morse in [3]. Moreover, it was shown that the condition in question may be reexpressed in terms of an equality between the structures at infinity of the plant and of a composed system in which the model appears as a "disturbance" on the output of the plant. This extends recent results by Malabre [6].

Appendix.

Krener's algorithm. Suppose the codistribution dh has constant dimension s_0. Let $\lambda_0(x)$ be an s_0-vector whose entries are entries of h such that $d\lambda_{0,1}, \cdots, d\lambda_{0,s_0}$ are linearly independent.

Iteration $(k+1)$. Consider the $s_k \times m$ matrix $A_k(x)$ defined at iteration k whose (i,j)th entry is $\langle d\lambda_{k,i}, g_j \rangle(x)$. Suppose the rank ρ_k of $A_k(x)$ is constant. Then, there exists an $s_k \times s_k$ permutation matrix

$$P_k = \begin{pmatrix} P_{k1} \\ P_{k2} \end{pmatrix}$$

such that the ρ_k rows of $P_{k1}A_k(x)$ are linearly independent. Let $B_k(x)$ denote an s_k vector whose ith element is $\langle d\lambda_{k,i}, f \rangle(x)$. Then, the equations

$$P_{k1}A_k(x)\alpha(x) = -P_{k1}B_k(x), \tag{A.1a}$$

$$P_{k1}A_k(x)\beta(x) = K, \tag{A.1b}$$

where K is a matrix of real numbers of rank ρ_k, are solved by an m-vector α and an $m \times m$ invertible matrix β. Construct the set of functions

$$\Lambda_k = \{\lambda = L_{\tilde{g}_i}\lambda_{k,j} \,|\, 1 \leqq j \leqq s_k, 0 \leqq i \leqq m\}$$

with $\tilde{g}_0 = f + g\alpha$ and $\tilde{g} = g\beta$. Moreover, consider the codistribution

$$\operatorname{span}\{d\lambda \mid \lambda \in \Lambda_k\} + \operatorname{span}\{d\lambda_{k,j} \mid 1 \leqq j \leqq s_k\}.$$

Suppose the dimension s_{k+1} of this codistribution is constant and let

$$\{d\lambda_{k+1,1}, \cdots, d\lambda_{k+1,s_{k+1}}\}$$

be a basis for it. Then, it is possible to prove [14] that at each iteration

$$\Omega_{k+1} = \operatorname{span}\{d\lambda_{k+1,i} \mid 1 \leqq i \leqq s_{k+1}\}.$$

Moreover, there exists an integer k^* such that the algorithm terminates at the k^*th iteration (i.e. $s_{k^*+1} = s_{k^*}$) and one has:

$$\Omega^* = \operatorname{span}\{d\lambda_{k^*,i} \mid 1 \leqq i \leqq s_{k^*}\}.$$

This ends the description of the algorithm. □

REFERENCES

[1] B. C. MOORE AND L. M. SILVERMAN, *Model matching by state feedback and dynamic compensation*, IEEE Trans. Automat. Control, AC-17 (1972), pp. 491-497.

[2] L. M. SILVERMAN, *Inversion of multivariable linear systems*, IEEE Trans. Automat. Control, AC-14 (1969), pp. 270-276.

[3] A. S. MORSE, *Structure and design of linear model following systems*, IEEE Trans. Automat. Control, AC-18 (1973), pp. 346-354.

[4] ———, *Minimal solutions to transfer matrix equations*, IEEE Trans. Automat. Control, AC-21 (1976), pp. 131-133.

[5] E. EMRE AND M. L. J. HAUTUS, *A polynomial characterization of $(\mathcal{A}, \mathcal{B})$-invariant and reachability subspaces*, this Journal, 18 (1980), pp. 420-436.

[6] M. MALABRE, *Structure à l'infini des triplets invariants; application à la poursuite parfaite de modèle*, 5th International Conference on Analysis and Optimization of Systems, INRIA, Paris, 1982, pp. 43-53.

[7] A. ISIDORI AND A. RUBERTI, *On the synthesis of linear input-output responses for nonlinear systems*, Systems Control Lett., 4 (1984), pp. 17-22.

[8] A. ISIDORI, *The matching of a prescribed linear input-output behavior in a nonlinear system*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 258-265.

[9] T. OKUTANI AND K. FURUTA, *Model matching of nonlinear systems*, Preprints IFAC 9th World Congress, Budapest, 1984, vol. IX, pp. 168-172.

[10] M. SHIMA AND Y. ISURUGI, *Variational system theory*, Preprints IFAC 9th World Congress, Budapest, 1984, vol. V, pp. 83-88.

[11] A. ISIDORI, A. J. KRENER, C. GORI-GIORGI AND S. MONACO, *Nonlinear decoupling via feedback: a differential-geometric approach*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 331-345.

[12] H. NIJMEIJER AND J. M. SCHUMACHER, *Zeros at infinity for affine nonlinear control systems*, IEEE Trans. Automat. Control, 30 (1985), pp. 566-573.

[13] M. FLIESS, M. LAMNABHI AND F. LAMNABHI-LAGARRIGUE, *An algebraic approach to nonlinear functional expansions*, IEEE Trans. Circuits and Systems, CAS-30 (1983), pp. 554-570.

[14] A. J. KRENER, (Adf, g), (adf, g) *and locally* (adf, g) *invariant and controllability distributions*, this Journal, 23 (1985), pp. 523-549.

[15] C. MOOG AND A. GLUMINEAU, *Le problème du rejet des perturbations mesurables dans les systèmes non linéaires*, Outils et Modèles Mathématiques pour l'Automatique, l'Analyse des Systèmes et le Traitement du Signal. Vol. 3, Coord. I. D. Landau, CNRS, Paris, 1983.

[16] A. ISIDORI, *Nonlinear Control Systems: an Introduction*, Lecture Notes in Control and Information Science, 72, Springer-Verlag, Berlin, 1985.

[17] M. D. DI BENEDETTO, *A classification of nonlinear systems based on the invariant subdistribution algorithm*, in Algebraic and Geometric Methods in Nonlinear Control Theory, M. Fliess and M. Hazewinkel, eds., Paris, 1985, Reidel Dordrecht.

[18] G. CONTE, M. D. DI BENEDETTO, A. ISIDORI AND A. M. PERDON, *An input-output characterization of the structure at infinity for a nonlinear system*, Proc. MTNS, 85, Stockholm, 1985, to appear.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 5, September 1986

012

ON THE SENSITIVITY MINIMIZATION PROBLEM FOR LINEAR TIME-VARYING PERIODIC SYSTEMS*

AVRAHAM FEINTUCH†, PRAMOD KHARGONEKAR‡ AND ALLEN TANNENBAUM§

Abstract. An optimal control problem is formulated in the context of linear, discrete-time, periodic systems. The cost is the supremum over all exogenous inputs in a weighted ball of plant inputs. The controller is required to be causal, periodic of the fixed order of the system and to achieve internal stability. Existence of an optimal controller is proved and a formula for the minimum cost is derived.

Key words. optimal control, periodic systems, time-varying systems, uniform optimality

AMS(MOS) classifications. 93C25, 93C50, 93C55, 47B35

1. Introduction. Since the classic paper of Zames [18] (in which he formulated a linear-quadratic optimal control problem where the exogeneous signals are not fixed but belong to weighted balls in appropriate function spaces), there has been a great deal of work on various problems of weighted sensitivity minimization.

At first the work was restricted to time-invariant systems (see [4] and [6] for a complete bibliography). Most of the results are presented in a unified fashion in [6]. Feintuch and Francis then showed that these problems can be formulated and solved for time-varying systems in a Hilbert resolution space framework and that in fact one could recover the results of the time-invariant problems as special cases of the more general time-varying problem [4]. In particular, it was shown that for time-invariant systems there is no time-varying controller which is better than an optimal time-invariant one. A special case of this result was also obtained independently by Khargonekar, Poolla and Tannenbaum [11].

While the class of linear time-varying controllers is quite large and unruly, there are certain subclasses that are quite similar in their structure to time-invariant systems and yet different enough so that for problems of robustness, they are superior to time-invariant ones. One such class is that of periodic controllers.

The importance of such systems has been noted in the work of various authors such as Davis [2], Jury and Mullin [10], and Meyer and Burrus [13].

Here we were influenced mainly by the paper of Khargonekar, Poolla and Tannenbaum [11], with one major difference. While the authors have formulated their theory in terms of transfer functions in the frequency domain, we feel that it is often (though not always) more useful to look at operator-representations of systems in the time domain. This is consistent with the approach in [4].

In this paper we formulate and solve an optimal control problem of the Zames-type for N-periodic discrete-time, time-varying systems. The existence of an optimal controller is shown and a formula is obtained for the optimal cost. Finally, it is shown that no better result can be obtained by a time-varying nonperiodic controller. The formulae we obtain of course reduce to the classical formulae obtained by Francis and Zames [7], and Francis et al. in [8] when applied to time-invariant systems.

* Received by the editors April 8, 1985, and in revised form August 15, 1985.

† Department of Mathematics, Ben-Gurion University of the Negev, Beer Sheva 84105, Israel.

‡ Department of Electrical Engineering, University of Minnesota, Minneapolis, Minnesota 55455. This work was supported in part by the National Science Foundation under grants ECS-8451519 and ECS-8400832.

§ Department of Mathematics, Ben-Gurion University of the Negev, Beer Sheva 84105, Israel. Present address, Department of Electrical Engineering, McGill University, Montreal, Quebec, Canada, H3A 2A7.

This paper is similar both in format and spirit to [4], and in fact shows that the methods used there seem to be especially appropriate for attacking problems of this sort.

Finally, it should be noted that the results of this paper again indicate the fundamental importance of the system sensitivity function. Indeed in Khargonekar and Tannenbaum [12], it was shown that in the linear finite-dimensional time-invariant case that certain robust stabilization problems (for example, the gain margin problem, see Tannenbaum [16]) are equivalent to the sensitivity minimization problem of Zames [18]. From the results of Feintuch and Francis [4] and Khargonekar, Poolla and Tannenbaum [11], it can already be seen that in time-varying cases the analogous problem of robust design and sensitivity minimization become dichotomous. However the results of the present paper together with [4] and [11] indicate that sensitivity is a true system *invariant* in the sense that apparently sensitivity can always be minimized within a given class of systems (for example, time-invariant, periodic), and hence sensitivity must be regarded as a fundamental measure of system performance. This is of course precisely the design philosophy of [18].

2. Preliminaries. The purpose of this section is to introduce notation and definitions, and to collect some basic facts regarding linear operators and complex functions.

For an integer $n \geqq 1$, $\mathbb{C}^n$ denotes complex n-dimensional Hilbert space, the inner product of two vectors x and y being x^*y where * denotes complex conjugate transpose. The norm on $\mathbb{C}^n$ is denoted by $\|\cdot\|_2$, that is,

$$\|x\|_2 = (x^*x)^{1/2}.$$

Let $\mathbb{C}^{m\times n}$ denote the set of all $m \times n$ complex matrices. Then the norm on $\mathbb{C}^{m\times n}$, $\|\cdot\|_\infty$, is taken to be that induced by the one on $\mathbb{C}^n$, that is,

$$\|A\|_\infty = \sup\{\|Ax\|_2 : \|x\|_2 \leqq 1\}.$$

The set of all sequences $\{x_k : k \geqq 0\}$ in $\mathbb{C}^n$ is denoted by $\mathcal{A}(\mathbb{C}^n)$. In discussions where the integer n is immaterial, $\mathcal{A}(\mathbb{C}^n)$ will be shortened to $\mathcal{A}$. The subset of $\mathcal{A}(\mathbb{C}^n)$ of all square-summable sequences is denoted by $h_2(\mathbb{C}^n)$, or simply by h_2; that is, $\{x_k\} \in h_2$ if and only if

$$\sum_{k=0}^{\infty} \|x_k\|_2^2 < \infty.$$

Then h_2 is a Hilbert space under the inner product

$$\langle\{x_k\}, \{y_k\}\rangle = \sum x_k^* y_k.$$

The induced norm on h_2 is also denoted by $\|\cdot\|_2$.

Let $F : \mathcal{A}(\mathbb{C}^n) \to \mathcal{A}(\mathbb{C}^n)$ be a linear mapping. Then F has a natural matrix representation $(F_{ij} : i \geqq 0, j \geqq 0)$; $F_{ij} \in \mathbb{C}^{n\times m}$, defined by the equation

$$F\{0, \cdots, 0, x_i 0, \cdots\} = \{F_{0i}x_i, F_{1i}x_i, \cdots\}.$$

The operator F is termed *causal*, *strictly causal* or *time-invariant* if its matrix is lower block triangular, strictly lower block triangular, or constant along block diagonals, respectively.

The Banach space of bounded linear operators from $h_2(\mathbb{C}^m)$ to $h_2(\mathbb{C}^n)$ is denoted by $\mathcal{B}[h_2(\mathbb{C}^m), h_2(\mathbb{C}^n)]$ or simple $\mathcal{B}$. The subspaces of $\mathcal{B}$ of causal operators and of time-invariant operators are denoted by $\mathcal{C}$ and $\mathcal{T}$, respectively, and if $m = n$ these are in fact weakly closed subalgebras of $\mathcal{B}$ [5].

For frequency domain theory, we need some notation and facts concerning complex functions. The space of square-integrable functions defined on the unit circle and taking values in $\mathbb{C}^n$ is denoted by $L_2(\mathbb{C}^n)$ or simply by L_2. The closed subspace of L_2 of functions having analytic continuations into the unit disc is the Hardy space H_2. Its orthogonal complement is denoted by $H_2^{\perp}$. The *inner product* of f and g in L_2 is defined to be

$$(2\pi)^{-1}\int_0^{2\pi} f(e^{i\theta})^* g(e^{i\theta})\, d\theta$$

and the corresponding norm is denoted (also) by $\|\cdot\|_2$. It is a standard fact in the theory of Fourier series that h_2 and H_2 are isomorphic Hilbert spaces.

The set of essentially bounded functions defined on the unit circle and taking values in $\mathbb{C}^{m\times n}$ is denoted by $L_\infty(\mathbb{C}^{m\times n})$ or L_∞. The norm on L_∞ is one induced by that on L_2: For F in $L_\infty(\mathbb{C}^{m\times n})$,

$$\|F\|_\infty := \sup\{\|Fg\|_2 : g\in L_2, \|g\|_2 \leqq 1\}.$$

It can be proved that

$$\|F\|_\infty = \operatorname{ess\,sup}\{\|F(e^{i\theta})\|_\infty : \theta\in[0,2\pi]\}$$

the right-hand norm being the one on $\mathbb{C}^{m\times n}$. We then set H_∞ to be the subset of L_∞ consisting of matrices having analytic continuations into the unit disc.

There is a clear connection between operators which act in the time domain and complex functions in the frequency domain. If F is a time-invariant bounded linear operator on h_2, the matrix of F has the Toeplitz form

$$\begin{bmatrix} F_0 & F_{-1} & F_{-2} & \cdots \\ F_1 & F_0 & F_{-1} & \cdots \\ F_2 & F_1 & F_0 & \cdots \\ \vdots & \vdots & \vdots & \end{bmatrix}.$$

Define the transfer function of F to be $\mathring{F}(e^{i\theta}) := \sum_{-\infty}^{\infty} F_k e^{ki\theta}$. It is a standard fact that $\mathring{F}\in L_\infty$ and $\|F\| = \|\mathring{F}\|_\infty$. If $F\in\mathscr{C}\cap\mathscr{T}$ then the above matrix of F is lower block triangular and $\mathring{F}\in H_\infty$. By slight abuse of notation, we will many times identify F and $\mathring{F}$.

3. Periodic operators. Let Λ denote the right shift operator on $\mathscr{A}(\mathbb{C}^n)$; if

$$x=\{x_k\}\in\mathscr{A}(\mathbb{C}^n),\ \Lambda x = y \quad \text{where } y_0=0,\ y_i = x_{i-1},\ i\geqq 1.$$

DEFINITION 3.1. $T\in\mathscr{C}$ is N-periodic if $T\Lambda^N = \Lambda^N T$.

The fundamental properties of these operators were presented in [11] following the approach in [15]. Here we rework these facts in the time domain. We will present these results for 2-periodic operators in order to simplify notation. *All the results go through immediately for N-periodic operators for* $N>2$. The space of N-periodic operators on h_2 will be denoted by $\mathscr{P}_N$, and in particular the space of 2-periodic operators by $\mathscr{P}_2$.

It is easily checked that the block matrix representation of $T\in\mathscr{P}_2\cap\mathscr{C}$ is of the form

$$\begin{bmatrix} x_0 & 0 & 0 & \cdots \\ x_1 & y_0 & 0 & \cdots \\ x_2 & y_1 & x_0 & \cdots \\ x_3 & y_2 & x_1 & \\ \vdots & \vdots & \vdots & \end{bmatrix},$$

that is, T is determined by two sequences of "Fourier coefficients." As was shown in [11], there is an isometric isomorphism W of h_2 onto $h_2 \oplus h_2 \cong h_2(\mathbb{C}^{2n})$ which has the property that if $\tilde{T} = WTM^{-1} = WTW^*$, then $\tilde{T}\tilde{\Lambda} = \tilde{\Lambda}\tilde{T}$ where $\tilde{\Lambda}$ is the right shift on $h_2(\mathbb{C}^{2n}) = h_2 \oplus h_2$. The matrix representation of $\tilde{T}$ on $h_2(\mathbb{C}^{2n})$ is simply the Toeplitz of the form

$$\begin{bmatrix} T_0 & 0 & \cdots \\ T_1 & T_0 & \cdots \\ T_2 & T_1 & \cdots \\ \vdots & \vdots & \end{bmatrix},$$

where T_i is the 2×2 block matrix given by

$$T_i := \begin{bmatrix} x_{2i} & y_{2i-1} \\ x_{2i+1} & y_{2i} \end{bmatrix}$$

where y_{-1} is understood to be zero. In particular, T_0 is lower triangular and it is easy to see that for $A \in \mathcal{T} \cap \mathcal{C}$ on $h_2(\mathbb{C}^{2n})$, $A = \tilde{T}$ for some $T \in \mathcal{P}_2 \cap \mathcal{C}$ on $h_2(\mathbb{C}^n)$ if and only if A_0 is lower triangular.

If we look at $\tilde{T}$ from the frequency domain point of view, then $\tilde{T} \in H_\infty$ and the matrix-valued functions in H_∞ which correspond to 2-periodic operators are those $A \in H_\infty$ for which $A(0)$ is lower triangular.

For the general N-periodic case, a similar computation to that given above (see [11, (2.6)] for details) shows that with any m-input p-output, N-periodic linear time-varying causal input/output map f, one can canonically associate a $pN \times mN$ transfer function matrix T_f such that $T_f(0)$ is lower triangular. Conversely, any $pN \times mN$ transfer function matrix T_f such that $T_f(0)$ is lower triangular, defines an m-input, p-output, N-periodic linear time-varying causal input/output map f. We should emphasize the fact that it is the lower triangularity of $T_f(0)$ that corresponds to the *causality* of f.

Finally we recall that since W is unitary, the correspondence between T and $\tilde{T}$ is norm preserving, and is an isometric isomorphism between the weakly closed subalgebra $\mathcal{P}_2 \cap \mathcal{C}$ of $\mathcal{C}$ and its image, the weakly closed subalgebra of H_∞ consisting of functions which are lower triangular matrices at the origin.

4. Factorizations. There are two kinds of factorizations that play a role in the theory of minimum sensitivity, one being technical and the other fundamental to the nature of the problem being solved. The first is the inner-outer factorization of functions in H_∞ and the second is the coprime factorization of systems used in the Youla stabilization theory.

Since periodic operators have representations as H_∞ functions, we can obtain both types of factorizations (under the appropriate assumptions [3], [6], [17]) in H_∞. However for such factorizations to be useful it must be shown that the factors correspond to periodic systems; that is, that they are lower triangular at the origin.

THEOREM 4.1. *Suppose $P \in H_\infty$, and that $P(0)$ is lower triangular. Then*

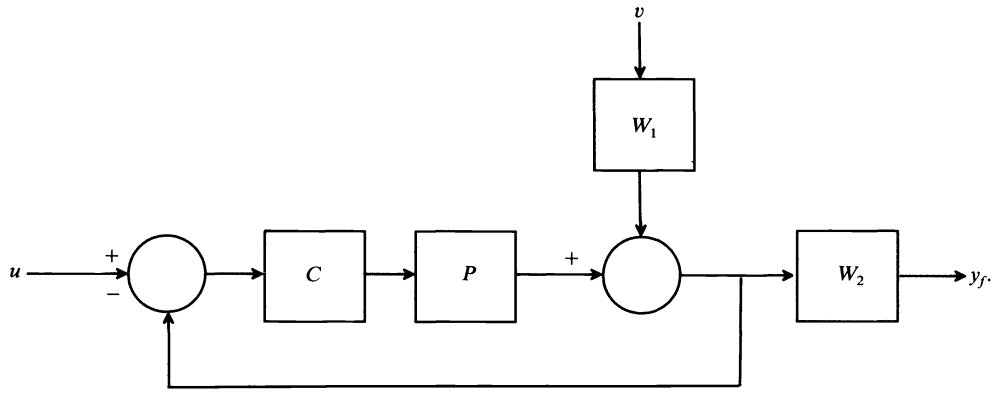
(1) *If P has a left (or right) coprime factorization in H_∞, the factors can be chosen so that they are lower triangular at the origin.*

(2) *P has an inner-outer factorization $P = P_i P_0$ such that $P_i(0)$ and $P_0(0)$ are lower triangular.*

Proof. For (1) suppose $P = AB^{-1}$ where $A, B \in H_\infty$. If $B(0)$ is not lower triangular there exists a constant invertible square matrix F (of appropriate dimension) such that $B(0)F$ is lower triangular. Then $P = (AF)(F^{-1}B^{-1})$, and $AF = PBF$ has the property that $(AF)(0)$ is lower triangular since $P(0)$ and $(BF)(0)$ are lower triangular.

For (2) note that if $P_i P_0$ is an inner-outer factorization for P, there exists a constant unitary matrix U such that $(UP_0)(0)$ is lower triangular. Thus $P=(P_0U^*)(UP_0)$ is an inner-outer factorization with the required properties. Q.E.D.

5. Formulation of the control problem. Consider the discrete-time linear feedback system below



Again for the sake of notational simplicity, we will assume P is a 2-periodic plant, the general N-periodic case being similar. W_1 and W_2 are stable 2-periodic operators with stable inverses (it is clear that W_i^{-1} are periodic when they are defined). We assume that P is such that in its time domain matrix representation $x_0=y_0=0$; that is, P is strictly causal. This assumption is standard and completely technical. Finally again by slight abuse of notation we will identify a time-invariant, time-domain operator F with its transfer function $\mathring{F}$ (as in § 2).

Our objective will be to design C to minimize the energy of y_f for the worst v of unit energy, C being constrained to be causal, 2-periodic and to achieve internal stability. Thus the cost is

$$\text{cost}=\sup\{\|y_f\|_2\colon v\in h_2,\ \|v\|_2\leqq 1\}.$$

In terms of the transfer matrix from v to y_f, we have

$$\text{cost}=\|W_2(I+PC)^{-1}W_1\|,$$

the induced operator norm in h_2.

Let μ denote the infimum of the cost taken over all causal 2-periodic C's achieving internal stability.

Since the family of 2-periodic operators on h_2 is a ring the parameterization of Youla et al. [6] is applicable whenever P has the appropriate right and left coprime factorizations. For example, if it is assumed that $\tilde{P}\in RH_\infty$ (:= the subspace of real-rational functions in H_∞) such factorizations exist and are computable [16], [17]. So assume $\tilde{P}$ has such factorization, and note that by Theorem 4.1 and the isometric isomorphism between $\mathscr{P}_2\cap\mathscr{C}$ and the above-mentioned subalgebra of H_∞, this gives rise to comprime factorizations in $\mathscr{P}_2\cap\mathscr{C}$. So assume $P=\hat{B}^{-1}\hat{A}=AB^{-1}$, such that

$$\hat{A}\hat{X}+\hat{B}\hat{Y}=I,\qquad XA+YB=I,$$

with A, B, $\hat{A}$, $\hat{B}$, X, Y, $\hat{X}$, $\hat{Y}\in\mathscr{P}_2\cap\mathscr{C}$. Then $C=(\hat{X}+BZ)(\hat{Y}-AZ)^{-1}$, $Z\in\mathscr{P}_2\cap\mathscr{C}$ parameterizes all controllers in $\mathscr{P}_2\cap\mathscr{C}$ achieving internal stability. With C so expressed the sensitivity matrix is affine in the parameter Z

$$(I+PC)^{-1}=(\hat{Y}-AZ)\hat{B}.$$

Thus the cost is

$$\text{cost} = \| W_2(\hat{Y} - AZ)\hat{B}W_1 \|.$$

Using the isometric isomorphism defined earlier, we rewrite this as

$$\text{cost} = \| \tilde{W}_2(\tilde{\hat{Y}} - \tilde{A}\tilde{Z})\hat{B}\tilde{W}_1 \|_\infty.$$

To simplify this further, bring in the appropriate inner-outer factorization of $\tilde{W}_2\tilde{A}$ (with periodic factors) and (under appropriate assumptions on P, [6]) an outer-inner factorization of $\tilde{\hat{B}}\tilde{W}_1$. Defining (under the assumption that $(\tilde{W}_2\tilde{A})_i$ is invertible)

$$\tilde{T} = (\tilde{W}_2\tilde{A})_i^* \tilde{W}_2 \tilde{Y}(\tilde{\hat{B}}\tilde{W}_1)_0, \qquad \tilde{V} = (\tilde{W}_2\tilde{A})_0 \tilde{Z}(\tilde{\hat{B}}\tilde{W}_1)_0$$

and using the properties of inner matrices and the assumptions on P, W_1, W_2, we obtain that

$$\text{cost} = \| \tilde{T} - \tilde{V} \|_\infty.$$

Note that $\tilde{V} \in H_\infty$ and corresponds to a periodic system $V \in \mathscr{P}_2 \cap \mathscr{C}$. $\tilde{T}$ is of course not in H_∞. However $\tilde{T} \in L^\infty$ and its periodicity property is preserved in the following way: If we apply the unitary mapping W to obtain $T = W^* \tilde{T} W$, then T has the block matrix representation

$$\begin{bmatrix} x_0 & y_{-1} & x_{-2} & \cdots \\ x_1 & y_0 & x_{-1} & \cdots \\ x_2 & y_1 & x_0 & \cdots \\ \vdots & \vdots & \vdots & \end{bmatrix}.$$

Thus we can write μ as

$$\mu = \inf\{\| T - V \| : V \in \mathscr{P}_2 \cap \mathscr{C}\}$$

which in topological terms is just the distance from the given noncausal 2-periodic operator T to the weakly closed subalgebra $\mathscr{P}_2 \cap \mathscr{C} \subset \mathscr{C}$. The existence of a $V \in \mathscr{P}_2 \cap \mathscr{C}$ for which μ is attained is now (as in [4] only easier) a consequence of the compactness of the unit-ball in the weak operator topology.

In the next section we obtain a formula for μ in terms of the entries of T.

6. The main results. In this section we derive a formula for μ in terms of the parameters given in T. The idea is the same as that in [4]. We make use of the following proposition due to Parrot [14] and, independently, Davis et al. [1]. Let X_i, Y_i, $(i = 1, 2)$ be Hilbert spaces and let

$$\begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix} : X_1 \oplus X_2 \to Y_1 \oplus Y_2$$

be a bounded linear operator.

PROPOSITION 6.1.

$$\inf\left\{\left\|\begin{bmatrix} G_{11} & G_{12} \\ G_{21} - X & G_{22} \end{bmatrix}\right\|\right\} = \max\left\{\|[G_{11}, G_{12}]\|, \left\|\begin{bmatrix} G_{12} \\ G_{22} \end{bmatrix}\right\|\right\},$$

where the infimum is over all bounded operators X from X_1 to Y_2.

Our main result can be stated.

THEOREM 6.2. *Suppose T is an operator on h_2 of the form*

$$T = \begin{bmatrix} x_0 & y_{-1} & x_{-2} & y_{-3} & \cdots \\ x_1 & y_0 & x_{-1} & y_{-2} & \cdots \\ x_2 & y_1 & x_0 & y_{-1} & \cdots \\ x_3 & y_2 & x_1 & y_0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \end{bmatrix}.$$

Then $d(T, \mathscr{P}_2 \cap \mathscr{C})$, the distance from T to $\mathscr{P}_2 \cap \mathscr{C}$ in the norm topology on $\mathscr{B}(\mathscr{H})$, is

$$\max\left\{\left\|\begin{bmatrix} \vdots & \vdots & \vdots & \\ x_{-4} & y_{-5} & x_{-6} & \cdots \\ x_{-3} & y_{-4} & x_{-5} & \cdots \\ x_{-2} & y_{-3} & x_{-4} & \cdots \\ x_{-1} & y_{-2} & x_{-3} & \cdots \end{bmatrix}\right\|, \left\|\begin{bmatrix} \vdots & \vdots & \vdots & \vdots & \\ y_{-3} & x_{-4} & y_{-5} & x_{-6} & \cdots \\ y_{-2} & x_{-3} & y_{-4} & x_{-5} & \cdots \\ y_{-1} & x_{-2} & y_{-3} & x_{-4} & \cdots \end{bmatrix}\right\|\right\}.$$

We begin the proof of Theorem 6.2 with the following lemma.

LEMMA 6.3. *Set*

$$A_1 := \begin{bmatrix} \vdots & \vdots & \vdots & \vdots & \\ x_{-4} & y_{-5} & x_{-6} & y_{-7} & \cdots \\ x_{-3} & y_{-4} & x_{-5} & y_{-6} & \cdots \\ x_{-2} & y_{-3} & x_{-4} & y_{-5} & \cdots \\ x_{-1} & y_{-2} & x_{-3} & y_{-4} & \cdots \end{bmatrix},$$

$$A_2 := \begin{bmatrix} \vdots & \vdots & \vdots & \vdots & \\ y_{-4} & x_{-5} & y_{-6} & x_{-7} & \cdots \\ y_{-3} & x_{-4} & y_{-5} & x_{-6} & \cdots \\ y_{-2} & x_{-3} & y_{-4} & x_{-5} & \cdots \\ y_{-1} & x_{-2} & y_{-3} & x_{-4} & \cdots \end{bmatrix}.$$

Then

$$\inf_{\{a_0, a_1, b_0\}} \left\|\begin{bmatrix} \vdots & \vdots & \vdots & \vdots & \\ x_{-2} & y_{-3} & x_{-4} & y_{-5} & \cdots \\ x_{-1} & y_{-2} & x_{-3} & y_{-4} & \cdots \\ x_0 - a_0 & y_{-1} & x_{-2} & y_{-3} & \cdots \\ x_1 - a_1 & y_0 - b_0 & x_{-1} & y_{-2} & \cdots \end{bmatrix}\right\| = \max\{\|A_1\|, \|A_2\|\}.$$

Proof. The idea is the same as that given in Parrot's proof of Nehari's theorem [14, p. 317]. By Parrot's theorem, Proposition 6.1

$$\inf_{a_0} \left\|\begin{bmatrix} \vdots & \vdots & \vdots & \vdots & \\ x_{-3} & y_{-4} & x_{-5} & y_{-6} & \cdots \\ x_{-2} & y_{-3} & x_{-4} & y_{-5} & \cdots \\ x_{-1} & y_{-2} & x_{-3} & y_{-4} & \cdots \\ x_0 - a_0 & y_{-1} & x_{-2} & y_{-3} & \cdots \end{bmatrix}\right\| = \max\{\|A_1\|, \|A_2\|\},$$

Choose a_0 to attain this infimum.

In the same way one can choose b_0 so that

$$\inf_{b_0}\left\|\begin{bmatrix} \vdots & \vdots & \vdots & \\ y_{-3} & x_{-4} & y_{-5} & \cdots \\ y_{-2} & x_{-3} & y_{-4} & \cdots \\ y_{-1} & x_{-2} & y_{-3} & \cdots \\ y_0-b_0 & x_{-1} & y_{-2} & \cdots \end{bmatrix}\right\| = \max\{\|A_1\|, \|A_2\|\}.$$

Choose a_1 so that

$$\inf_{a_1}\left\|\begin{bmatrix} \vdots & \vdots & \vdots & \\ x_{-4} & y_{-5} & x_{-6} & \cdots \\ x_{-3} & y_{-4} & x_{-5} & \cdots \\ x_{-2} & y_{-3} & x_{-4} & \cdots \\ x_{-1} & y_{-2} & x_{-3} & \cdots \\ x_0-a_0 & y_{-1} & x_{-2} & \cdots \\ x_1-a_1 & y_0-b_0 & x_{-1} & \cdots \end{bmatrix}\right\|$$

$$= \max\left\{\left\|\begin{bmatrix} \vdots & \vdots & \vdots & \\ x_{-3} & y_{-4} & x_{-5} & \cdots \\ x_{-2} & y_{-3} & x_{-4} & \cdots \\ x_{-1} & y_{-2} & x_{-3} & \cdots \\ x_0-a_0 & y_{-1} & x_{-2} & \cdots \end{bmatrix}\right\|, \left\|\begin{bmatrix} \vdots & \vdots & \vdots & \\ y_{-3} & x_{-4} & y_{-5} & \cdots \\ y_{-2} & x_{-3} & y_{-4} & \cdots \\ y_{-1} & x_{-2} & y_{-3} & \cdots \\ y_0-b_0 & x_{-1} & y_{-2} & \cdots \end{bmatrix}\right\|\right\}$$

$$= \max\{\|A_1\|, \|A_2\|\}.$$

This completes the proof. Q.E.D.

Proof of Theorem 6.2.

$$d(T, \mathscr{P}_2 \cap \mathscr{C}) = \inf_{\substack{\{a_i : i\geqq 0\} \\ \{b_i : i\geqq 0\}}}\left\{\left\|\begin{bmatrix} x_0-a_0 & y_{-1} & x_{-2} & \cdots \\ x_1-a_1 & y_0-b_0 & x_{-1} & \cdots \\ x_2-a_2 & y_1-b_1 & x_0-a_0 & \cdots \\ x_3-a_3 & y_2-b_2 & x_1-a_1 & \cdots \\ \vdots & \vdots & \vdots & \end{bmatrix}\right\|\right\}$$

$$= \inf_{a_0,a_1,b_0}\left\{\inf_{\substack{\{a_i : i\geqq 2\} \\ \{b_i : i\geqq 1\}}}\left\{\left\|\begin{bmatrix} X_0 & X_{-1} & X_{-2} & \cdots \\ X_1 & X_0 & X_{-1} & \cdots \\ X_2 & X_1 & X_0 & \cdots \\ \vdots & \vdots & \vdots & \end{bmatrix}\right\|\right\}\right\}$$

$$= \inf_{a_0,a_1,b_0}\left\{\inf_{\{X_i : i\geqq 1\}}\left\{\left\|\begin{bmatrix} X_0 & X_{-1} & X_{-2} & \cdots \\ X_1 & X_0 & X_{-1} & \cdots \\ X_2 & X_1 & X_0 & \cdots \\ \vdots & \vdots & \vdots & \end{bmatrix}\right\|\right\}\right\}$$

where

$$X_1 := \begin{bmatrix} x_2-a_2 & y_1-b_1 \\ x_3-a_3 & y_2-b_2 \end{bmatrix}, \qquad X_0 := \begin{bmatrix} x_0-a_0 & y_{-1} \\ x_1-a_1 & y_0-b_0 \end{bmatrix},$$

etc. The infimum inside the bracket is over Toeplitz matrices and therefore by Nehari's

theorem [14] it is just the norm of the Hankel matrix defined by $\{X_i: i \leqq 0\}$:

$$\left\| \begin{bmatrix} \vdots & \vdots & \vdots & \vdots & \\ X_{-2} & X_{-3} & X_{-4} & X_{-5} & \cdots \\ X_{-1} & X_{-2} & X_{-3} & X_{-4} & \cdots \\ X_0 & X_{-1} & X_{-2} & X_{-3} & \cdots \end{bmatrix} \right\|.$$

Thus

$$d(T, \mathscr{P}_2 \cap \mathscr{C}) = \inf_{a_0, a_1, b_0} \left\{ \left\| \begin{bmatrix} \vdots & \vdots & \\ X_{-2} & X_{-3} & \cdots \\ X_{-1} & X_{-2} & \cdots \\ X_0 & X_{-1} & \cdots \end{bmatrix} \right\| \right\}.$$

By the Lemma 6.3, this is simply $\max\{\|A_1\|, \|A_2\|\}$ and the proof is complete. Q.E.D.

We now ask the question: Can we obtain a lower sensitivity if we allow time-varying controllers? Clearly, since $\mathscr{P}_2 \cap \mathscr{C} \subset \mathscr{C}$, $d(T, \mathscr{P}_2 \cap \mathscr{C}) \geqq d(T, \mathscr{C})$.

However by [4]

$$d(T, \mathscr{C}) = \sup_n \|P_n T(I - P_n)\|.$$

This is again easily computed as in [4, Thm. 4] and we obtain the same number as in Theorem 6.2. Thus $d(T, \mathscr{C}) = d(T, \mathscr{P}_2 \cap \mathscr{C})$ and a nonperiodic controller will *not* improve the minimal sensitivity.

7. The *N*-periodic case. All the results given above relate to the N-periodic case for $N > 2$ (finite). In this case we simply state the result. Recall that N-periodic systems will be determined by N sequences $\{x_i^1\}, \cdots, \{x_i^N\}$. In this case T will be of the form

$$\begin{bmatrix} x_0^1 & x_{-1}^2 & x_{-2}^3 & \cdots & x_{-N+1}^N & x_{-N}^1 & \cdots \\ x_1^1 & x_0^2 & x_{-1}^3 & \cdots & x_{-N+2}^N & x_{-N+1}^1 & \cdots \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \end{bmatrix}.$$

Then

$$d(T, \mathscr{P}_N \cap \mathscr{C}) = \max\{\|A_1\|, \cdots, \|A_N\|\}$$

where

$$A_i = \begin{bmatrix} \vdots & \vdots & \vdots & \\ x_{-2}^i & x_{-3}^{i+1} & x_{-4}^{i+2} & \cdots \\ x_{-1}^i & x_{-2}^{i+1} & x_{-3}^{i+2} & \cdots \end{bmatrix}.$$

We note that, of course, if $N = 1$ our formulas simply reduce to the norm of the Hankel operator determined by the single sequence $\{x_{-1}, x_{-2}, \cdots\}$ which is of course the classical result.

8. Conclusions. The utility of periodic systems in control design has been discussed by a number of authors [2], [11], [13]. In this note, we applied the design philosophy of sensitivity minimization (Zames [18]) to periodic plants. As in the time-invariant case, we showed that apparently one can always minimize sensitivity within a given class of systems, this time for periodic plants.

Moreover, the powerful time-domain input/output techniques of [4] were again shown to be appropriate for the successful solution of sensitivity minimization prob-

lems. We should note that the authors were not successful in pushing through their results using transfer-function techniques and in particular an appropriate version of the commutant lifting theorem (as for the time-invariant case). However given the importance of such techniques in robust system design [11], it seems that this should be a topic of future research.

The reader will notice that even though only the sensitivity minimization problem for a set-up of the kind discussed in § 5 was treated here, the results go through immediately for the more general situation considered in [4]. Moreover in the finite-dimensional case, one can explicitly compute the norms of the Hankel matrices involved in the solution of the sensitivity minimization problem for periodic plants. In the present work however, we have only considered the issue of existence.

Finally, the problem of sensitivity minimization for other interesting classes of time-varying systems should make an interesting subject for future investigation.

Acknowledgment. We would like to thank P. Fuhrmann for some interesting discussions on the problems considered in this paper.

REFERENCES

[1] C. DAVIS, W. M. KAHAN AND H. F. WEINBERGER, *Norm preserving dilations and their applications to optimal error bounds*, SIAM J. Numer. Anal., 10 (1982), pp. 445-469.

[2] J. H. DAVIS, *Stability conditions derived from spectral theory: discrete systems with periodic feedback*, this Journal, 10 (1975), pp. 1-13.

[3] A. FEINTUCH, *Co-prime factorizations for Youla stabilization of discrete time time-varying systems*, System and Control Letters, 1986.

[4] A. FEINTUCH AND B. A. FRANCIS, *Uniformly optimal control of linear feedback systems*, Automatica, 21 (1985), pp. 563-574.

[5] A. FEINTUCH AND R. SAEKS, *System Theory: A Hilbert Space Approach*, Academic Press, New York, 1982.

[6] D. C. YOULA, H. A. JABV AND J. J. BONGIORNO, *Modern Wiener-Hopf design of optimal controllers—Part 2: The multivariable case*, IEEE Trans. Automat. Control, AC-21 (1976), pp. 319-338.

[7] B. A. FRANCIS AND G. ZAMES, *On H^∞-optimal sensitivity for siso feedback systems*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 9-16.

[8] B. A. FRANCIS, J. W. HELTON AND G. ZAMES, *H^∞-optimal feedback controllers for linear multivariable systems*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 888-900.

[9] P. A. FUHRMANN, *Linear Systems and Operators in Hilbert Space*, McGraw-Hill, New York, 1981.

[10] E. I. JURY AND F. J. MULLIN, *The analysis of sampled data control systems with a periodically time-varying sampling rate*, IRE Trans. Automat. Control, AC-24 (1959), pp. 15-21.

[11] P. KHARGONEKAR, K. POOLLA AND A. TANNENBAUM, *Robust control of linear time-invariant plants using periodic compensation*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 1088-1096.

[12] P. KHARGONEKAR AND A. TANNENBAUM, *Noneuclidean metrics and the robust stabilization of systems with parameter uncertainty*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 1005-1013.

[13] R. A. MEYER AND C. S. BURRUS, *A unified analysis of multirate and periodically time-varying digital filters*, IEEE Trans. Circuits and Systems, CAS-22 (1975), pp. 162-168.

[14] S. PARROT, *On a quotient norm and the Sz. Nagy-Foias lifting theorem*, J. Funct. Anal., 30 (1978), pp. 311-328.

[15] B. SZ.-NAGY AND C. FOIAS, *Harmonic Analysis of Operators on Hilbert Space*, American Elsevier, New York, 1970.

[16] A. TANNENBAUM, *Feedback stabilization of plants with uncertainty in the gain factor*, Int. J. Control, 32 (1980), pp. 1-16.

[17] M. VIDYASAGAR, *Control System Synthesis: A Factorization Approach*, M.I.T. Press, Cambridge, Massachusetts, 1985.

[18] G. ZAMES, *Feedback and optimal sensitivity*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 310-320.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 5, September 1986

013

DIFFERENTIAL GAMES AND DIRECTIONAL DERIVATIVES OF VISCOSITY SOLUTIONS OF ISAACS' EQUATIONS II*

P.-L. LIONS† AND P. E. SOUGANIDIS‡

Abstract. Recent work by the authors [this Journal, 23 (1985), pp. 566–583], has demonstrated the connections 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 solutions of the above equations at an arbitrary point. The present note contains a remark and a counterexample which complement the results cited.

Key words. differential games, optimal control, Hamilton–Jacobi equations, directional derivatives, viscosity solutions

AMS (MOS) subject classifications. 35F30, 35L60, 90D25, 49C20

Introduction. We present here a remark and a counterexample which complement the results of P.-L. Lions and P. E. Souganidis [2]. In [2], we considered deterministic control and differential games problems and we studied the relations between two formulations of the associated Bellman and Isaacs' equations. In the case of Lipschitz continuous functions, we compared the notion of viscosity solutions of general Hamilton–Jacobi equations (introduced by M. G. Crandall and P.-L. Lions [1]) and a formulation due to A. I. Subbotin [3] concerning directional derivatives of the value functions. The equivalence of these notions was shown in [2] in the case of control or differential games problems under the Isaacs' condition: in fact, even in the general case of upper and lower value functions, we proved "one half" of the set of the inequalities to be checked. This combined with the Isaacs' condition was enough for the value functions. Here we present a simple example showing that the other half is in general false and we give a positive answer in a very particular situation.

To simplify the presentation, we restrict ourselves to the case of infinite horizon problems without state constraints

$$\inf_{y\in Y}\sup_{z\in Z}\{-f(x,y,z)\cdot DV - l(x,y,z)\} + V(x) = 0 \quad \text{in } R^N, \tag{1}$$

where Y, Z are fixed compact metric spaces and f, l are bounded, uniformly continuous on $R^N \times Y \times Z$ and Lipschitz continuous in $x \in R^N$ uniformly in $(y, z) \in Y \times Z$. As is well known, (1) corresponds to the equation satisfied by the lower value of a differential game (see [2] and below for more details).

In [2] we proved that a Lipschitz continuous function V is a viscosity supersolution (see below for the definition) of (1) if and only if

$$\inf_{y\in Y}\sup_{(f,l)\in K(x,y)}\left\{\overline{\lim_{t\to 0+}}\frac{V(x)-V(x+tf)}{t} - l\right\} + V(x) \geqq 0 \quad \text{in } R^N, \tag{2}$$

where for $x \in R^N$, $y \in Y$ $K(x, y) = \overline{\text{co}}\,\{(f(x, y, z), l(x, y, z))/z \in Z\}$. On the other hand, it is easy to check (see also [2]) that if V satisfies

* Received by the editors April 10, 1985.

† Universite Paris IX-Dauphine, Paris, 75775, Cedex 16, France.

‡ Lefschetz Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. This work was completed while the author was visiting the Institute for Mathematics and its Applications, University of Minnesota, Minneapolis, Minnesota. This work was partially supported by the National Science Foundation under grant MCS-8002946 and the Office of Naval Research under contract N00014-83-K-0542.

$$\inf_{y\in Y}\sup_{(f,l)\in K(x,y)}\left\{\lim_{t\to 0+}\frac{V(x)-V(x+tf)}{t}-l\right\}+V(x)\leqq 0 \quad \text{in } R^N, \tag{3}$$

then V is a viscosity subsolution of (1). The converse statement was left open. In §1 we show that a Lipschitz continuous viscosity subsolution of (1) does not necessarily satisfy (3). Recalling that lower values are always viscosity solutions of (1), this counterexample shows the limitation of the formulation of (1) by the pair of inequalities (2) and (3). Finally, in §2 we describe a particular situation where the converse statement holds. (Some other situations where the equivalence holds are indicated in [2].)

We conclude the introduction by recalling the definition of the viscosity solution. Let $C(\mathcal{O})$ ($C^{0,1}(\mathcal{O})$) denote the set of continuous (Lipschitz continuous) functions defined on $\mathcal{O}$. We have:

DEFINITION. Let $F\in C(R^N\times R\times R^N)$. $u\in C(R^N)$ is a viscosity supersolution (subsolution) of

$$F(x, u, Du)=0 \quad \text{in } R^N$$

if

$$F(x, u(x), D\phi(x))\leqq 0 \qquad (F(x, u(x), D\phi(x))\geqq 0),$$

for every smooth function ϕ in R^N and every local maximum (minimum) x of $u-\phi$ in R^N.

1. A counterexample. Let $l(x)$ be a bounded Lipschitz continuous function on R^N such that

$$|l(x)-l(y)|\leqq |x-y| \quad \text{for all } x, y\in R^N$$

and

$$l(x)=-|x| \quad \text{in a neighborhood of 0.}$$

We take $Y=Z=\{\xi\in R^N, |\xi|\leqq 1\}$ and choose

$$f(x, y, z)=z, \qquad l(x, y, z)=l(x)-(y, z).$$

Thus,

$$\sup_{z\in Z}\{-f\cdot p-l\}=|p-y|-l(x) \quad \text{for all } x, p\in R^N$$

and

$$\inf_{y\in Y}\sup_{z\in Z}\{-f\cdot p-l\}=(|p|-1)^+-l(x).$$

Obviously, $V(x)=l(x)$ is a viscosity solution of (1). On the other hand, if we try to check (3) at $x=0$ we obtain

$$\sup_{(f,l)\in K(0,y)}\left\{\lim_{t\to 0+}\frac{V(0)-V(0+tf)}{t}-l\right\}+V(0)=\sup_{z\in Z}\{|z|+(y, z)\}=1+|y|.$$

The infimum over y yields 1. Hence, (3) does not hold at $x=0$.

2. A particular situation. In this section we prove that if $N=1$ and $l(x, y, z)=l(x)$, then any viscosity subsolution $V\in C^{0,1}(R)$ of (1) satisfies (3). We will prove this claim following a general scheme of proof and indicating where the above restrictive assumptions are needed.

Let $P(Z)$ be the set of probability measures on Z and identify Z with the subspace of $P(Z)$ consisting of Dirac measures. For $\mu\in P(Z)$, $y\in Y$, $x\in R^N$ we set

$$f(x, y, \mu)=\int_Z f(x, y, z)\,d\mu(z),$$

$$l(x, y, \mu)=\int_Z l(x, y, z)\,d\mu(z).$$

Let M and N be the set of measurable functions y_s and z_s from $[0, \infty)$ into Y and $P(Z)$, respectively. We consider the set Δ of strategies β which are maps from M into N such that for all $t \geqq 0$

$$\beta(y^1)_s = \beta(y^2)_s \text{ on } [0, t] \quad \text{if } y_s^1 = y_s^2 \text{ on } [0, t].$$

Observing that for all $x \in R^N$, $y \in Y$, $p \in R^N$

$$\sup_{z \in Z} \{-f(x, y, z) \cdot p - l(x, y, z)\} = \sup_{\mu \in P(z)} \{-f(x, y, \mu) \cdot p - l(x, y, \mu)\},$$

we deduce from [2] that if $V \in C(R^N)$ is a viscosity subsolution of (1), then for all $h > 0$ and $x \in R^N$ we have

$$V(x) \leqq \inf_{\beta \in \Delta} \sup_{y \in M} \left\{ \int_0^h e^{-s} l(x_s, y_s, \beta[y]_s)\, ds + V(x_h) e^{-h} \right\}, \tag{4}$$

where x_t is the solution of

$$\frac{dx_t}{dt} = f(x_t, y_t, \beta[y]_t) \quad \text{on } [0, \infty), \qquad x_0 = x.$$

To prove (3) we argue by contradiction. Indeed, we assume that there exist $x \in R^N$ and $\delta > 0$ such that

$$\inf_{y \in Y} \sup_{(f,l) \in K(x,y)} \left\{ \lim_{t \to 0+} \frac{V(x) - V(x + tf)}{t} - l \right\} + V(x) \geqq \delta > 0.$$

Hence, for every $y \in Y$, there exist $t_y > 0$, $m_y \in N^+$ (positive integers), $z(y) \in Z$ and $\theta_i(y) \in (0, 1)$ with $i = 1, \cdots, m_y$ such that

$$\sum_{i=1}^{m_y} \theta_i(y) = 1,$$

and

$$\frac{V(x) - V(x + tf_y)}{t} - l_y + V(x) \geqq \frac{\delta}{2} > 0 \quad \text{for } 0 < t < t_y,$$

where

$$f_y = \sum_i^{m_y} \theta_i(y) f(x, y, z_i(y)), \qquad l_y = \sum_i^{m_y} \theta_i(y) l(x, y, z_i(y)).$$

Here we used the fact that $V \in C^{0,1}(R^N)$ and that $(V(x) - V(x + tf))/t$ is continuous in f uniformly for $t > 0$. If we set

$$f_{\bar{y},y} = \sum_{i=1}^{m_y} \theta_i(y) f(x, \bar{y}, z_i(y)), \qquad l_{\bar{y},y} = \sum_{i=1}^{m_y} \theta_i(y) l(x, \bar{y}, z_i(y)),$$

the properties of f and l imply that for every $y \in Y$ there exists $r_y > 0$ such that

$$\frac{1}{t} \{V(x) - V(x + tf_{\bar{y},y})\} - l_{\bar{y},y} + V(x) \geqq \frac{\delta}{4} > 0$$

for $0 < t < t_y$ and $d(\bar{y}, y) < r_y$, where d denotes the metric of Y. Using the compactness of Y we obtain the existence of $M \geqq 1$, $\bar{t} > 0$, $y_1, \cdots, y_M \in Y$ and $r_1, \cdots, r_M > 0$ such that $Y \subset \cup_{k=1}^M B(y_k, r_k)$ and

$$\frac{1}{t} \{V(x) - V(x + t\bar{f}_y)\} - \bar{l}_y + V(x) \geqq \frac{\delta}{4} \quad \text{for } 0 < t < \bar{t}, \tag{5}$$

where for $y \in B(y_j, r_j) \setminus \cup_{k=1}^{j-1} B(y_k, r_k)$, $\bar{f}_y$, $\bar{l}_y$ are defined by

$$\bar{f}_y = \sum_{i=1}^{m_{y_j}} \theta_i(y_j) f(x, y, z_i(y_j)), \qquad \bar{l}_y = \sum_{i=1}^{m_{y_j}} \theta_i(y_j) l(x, y, z_i(y_j)).$$

Next, we introduce a map $\tilde{\beta}$ from Y into $P(z)$ defined by

$$\tilde{\beta}(y) = \sum_{i=1}^{m_{y_j}} \theta_i(y_j)\delta_{z_i(y_j)} \quad \text{if } y \in B(y_j, r_j) \setminus \bigcup_{k=1}^{j-1} B(y_k, r_k).$$

This map obviously defines a strategy of $\beta \in \Delta$ by

$$\beta[y]_s = \tilde{\beta}(y_s) \quad \text{for } s \geqq 0.$$

Moreover,

$$f(x, y_s, \beta[y]_s) = \bar{f}_{y_s}, \qquad l(x, y_s, \beta[y]_s) = \bar{l}_{y_s}.$$

Hence, for any $y \in M$, we have

$$\begin{aligned}
&\frac{1}{t}\{V(x) - V(x_t)e^{-t}\} - \frac{1}{t}\int_0^t e^{-s} l(x_s, y_s, \beta[y]_s)\, ds \\
&\quad \geqq -Ct + \frac{1}{t}\{V(x) - V(x_t)\} - \frac{1}{t}\int_0^t l(x_s, y_s, \beta[y]_s)\, ds + V(x) \\
&\quad \geqq -Ct + V(x) + \frac{1}{t}\left\{V(x) - V\left(x + \int_0^t f(x, y_s, \beta[y]_s)\, ds\right)\right\} - \frac{1}{t}\int_0^t l(x, y_s, \beta[y]_s)\, ds \\
&\quad \geqq -Ct + V(x) + \frac{1}{t}\left\{V(x) - V\left(x + \int_0^t \bar{f}_{y_s}\, ds\right)\right\} - \frac{1}{t}\int_0^t \bar{l}_{y_s}\, ds,
\end{aligned}$$

where C denotes various constants independent of t, $y \in M$.

Next observe that the suboptimality principle (4) yields that the left-hand side of the preceding inequality is nonpositive. We wish to reach a contradiction by showing that (5) implies that the right-hand side is nonnegative for t small. The counterexample of §1 shows that this is not possible in general. However, in the particular situation where $l(x, y, z) = l(x)$ (hence $l_y = l(x)$ for all y) and $N = 1$, there exists $s_0 \in (0, t)$ such that

$$\begin{aligned}
&V(x) + \frac{1}{t}\left\{V(x) - V\left(x + \int_0^t f_{y_s}\, ds\right)\right\} - \frac{1}{t}\int_0^t l_{y_s}\, ds \\
&\quad = \frac{1}{t}\{V(x) - V(x + t\bar{f}_{y_{s_0}})\} + V(x) - l(x) \geqq \frac{\delta}{4}
\end{aligned}$$

for $t < \bar{t}$. The contradiction proves our claim.

REFERENCES

[1] M. G. CRANDALL AND P.-L. LIONS, *Viscosity solutions of Hamilton–Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1–42.

[2] P.-L. LIONS AND P. E. SOUGANIDIS, *Differential games, optimal control and directional derivatives of viscosity solutions of Bellman's and Isaacs' equations*, this Journal, 23 (1985), pp. 566–583.

[3] 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. 24, No. 6, November 1986

TRANSCENDENTAL AND INTERPOLATION METHODS IN SIMULTANEOUS STABILIZATION AND SIMULTANEOUS PARTIAL POLE PLACEMENT PROBLEMS*

B. K. GHOSH†

Abstract. In this paper, we investigate the existence of a compensator which simultaneously renders a given r-tuple of multiinput multioutput $p \times m$ linear dynamical systems internally stable. In particular we parametrize a set of simultaneously stabilizable r-tuples of plants and show that provided $r \leqq \max(m, p)$, the above set is semialgebraic and dense in the space Σ of r-tuples of plants. Furthermore, we also consider an extension of the classical pole placement and stabilization problems and investigate the simultaneous partial pole placement problem. Consequently, we consider a suitable topology in Σ and obtain a necessary condition and a sufficient condition for the generic partial pole placement problem. Surprisingly enough, the problem of simultaneously stabilizing a triplet of $m \times m$ plants, chosen generically, is shown equivalent to the problem of partially pole placing a single $m \times m$ plant by a stable minimum phase compensator. On the other hand the problem of simultaneously stabilizing $m+2$ tuple of $1 \times m$ plants, chosen generically, is shown equivalent to the problem of partially pole placing a $1 \times m$ plant by a stable minimum phase compensator. Investigating the $1 \times m$ plants in details, we parametrize the set of all compensators simultaneously stabilizing a m tuple of $1 \times m$ plants chosen generically. Consequently, we obtain a necessary and sufficient condition for simultaneously stabilizing a $m+1$ tuple of $1 \times m$ plants chosen generically. Lastly we construct two numerical examples. The first example is a triplet of simultaneously unstabilizable single input single output plants, each of McMillan degree 1 that are simultaneously stabilizable in pairs. The second example is a triplet of 1×3 plants that are not simultaneously stabilizable.

Key words. partial pole placement, simultaneous stabilization, generic, semialgebraic

AMS(MOS) subject classifications. 14, 30, 93

1. Introduction.

In order to introduce and analyze the problems considered in this paper we need the following notation.

$\mathbb{C}$:	the complex plane
$\mathbb{R}$:	the real line
$\mathbb{C}_s$:	a self-conjugate subset of $\mathbb{C}$ which intersects $\mathbb{R}$
$\mathbb{C}_u$:	$[\mathbb{C}-\mathbb{C}_s] \cup \{\infty\}$
$\mathbb{R}_u$:	$\mathbb{R} \cap \mathbb{C}_u$
$\mathbb{C}^-$:	open left half of the complex plane
H:	ring of proper rational functions with real co-efficients with poles in $\mathbb{C}_s$
$H^{p \times m}$:	set of $p \times m$ matrices whose elements belong to H
J:	set of multiplicative units in H
F:	quotient field of H [21, pp. 88-90].

Let us consider a set $G_1(s), \cdots, G_r(s)$ of r real, linear time invariant, proper $p \times m$ dynamical systems, each of a given fixed McMillan degree n_i, $i = 1, \cdots, r$ with m inputs and p outputs and ask the following problem.

Problem 1.1 (simultaneous stabilization). "When does there exist a nonswitching p input m output real, linear, time-invariant, proper compensator of arbitrary large but finite McMillan degree q, which stabilizes each of the r given plants either in discrete time or in continuous time?"

* Received by the editors July 3, 1984, and in revised form September 16, 1985. This work was partially supported by National Aeronautics and Space Administration grant NSG-2265 while the author was at Harvard University, Cambridge, Massachusetts 02138.

† Department of Systems Science and Mathematics, Washington University, St. Louis, Missouri 63130.

The simultaneous stabilization problem 1.1 has been originally addressed by Birdwell, Castanon and Athans [2] for the case $m = p = 1$ and is followed up subsequently by Saeks and Murray [26] as a problem in reliability and in the design of a multi-mode control system. Subsequently, Vidyasagar and Viswanadham [34] addressed the above problem for the multiinput multioutput case and showed that if $\min(m, p) > 1$, a generic pair of $p \times m$ plants is simultaneously stabilizable in a suitably chosen topology. The case $m = p = 1$ is more difficult and it is shown [27], [34] that the problem of simultaneously stabilizing a pair of single input single output plants is equivalent to the well-known problem considered by Youla, Bongiorno and Lu [35]: When can a single plant be stabilized by a stable compensator? Moreover the problem of stabilizing a triplet of single input single output plant chosen generically has been shown by Ghosh [14] to be equivalent to the problem of partially pole placing a single input single output plant by a stable minimum phase compensator. The problem of simultaneous partial pole placement, introduced and analyzed for the single input single output plants in [14], consists in answering the following problem:

Problem 1.2 (simultaneous partial pole placement). "Given a r tuple $G_1(s), \cdots, G_r(s)$ of $p \times m$ proper transfer functions of degree n_i, $i = 1, \cdots, r$, respectively, together with a r tuple of proper rational functions $\psi_i(s)$ of degree d_i, $i = 1, \cdots, r$ respectively from H. Does there exist a proper $m \times p$ compensator $K(s)$ of degree $q \geqq \max_i [d_i - n_i]$ such that the closed loop systems $G_1(s)[I + K(s)G_1(s)]^{-1}, \cdots, G_r(s)[I + K(s)G_r(s)]^{-1}$ have, respectively, d_i poles precisely where $\psi_i(s)$ vanishes and all but the above d_i poles are in $\mathbb{C}_s$?"

As explained in [14], the partial pole placement problem 1.2 is an extension of the pole placement and stabilization problems. In fact, if we assume $d_1 = d_2 = \cdots = d_r = 0$ and $\mathbb{C}_s = \mathbb{C}^-$ then the problem 1.2 reduces to the simultaneous stabilization problem 1.1. On the other hand, if we choose $d_i = n_i + q$, one obtains the simultaneous pole placement problem described in [12] and [18]. Frequently however, it is necessary to choose d_i in between 0 and $n_i + q$ for all $i = 1, \cdots, r$. Such a choice, we remark, satisfies the need to shape the response of the closed loop system to the extent that the designer can place an arbitrary number of self conjugate poles in the closed loop while restricting the remaining poles in the region $\mathbb{C}_s$. The above problem 1.2 also includes the simultaneous version of the well-known dominant pole placement problem and the gain margin problem. It also appears in the analysis of a class of robust stabilization problems described in [15], [16], the reliable stabilization problem described in [11] and many other design problems described in [17].

The new ingredient in this paper is the application of interpolation and transcendental methods in analyzing the simultaneous partial pole placement problem. This is now described as follows.

Interpolation problem 1.3. "Given a self conjugate set of pairs of complex numbers (s_i, z_i), $i = 1, \cdots, t$, does there exist a $\Delta(s) \in J$ such that $\Delta(s_i) = z_i$, for all $i = 1, \cdots, t$?"

The problem of interpolation by a rational function is known to be very important in network theory and electrical engineering. For accounts of this see Youla-Saito [36], Zeheb-Lempel [39], Helton [20] and many other references therein. In control theory, similar interpolation methods have been successfully applied by Tannenbaum [30], Zames and Francis [37], Kimura [22] based on classical work of Nevanlinna [24] and Pick [25]. More recently, Vidyasagar and Davidson [32] have used interpolation methods to characterize the set of stable stabilizing compensators for a single input single output system.

Assuming $\mathbb{C}_s$ to be the open left half of the complex plane, the interpolation problem 1.3 has been posed and solved by Youla, Bongiorno and Lu [35] in order to

analyze the stabilizability problem of a single input single output system by a stable compensator, in the continuous time. Complete solution to the problem 1.3 when $\mathbb{C}_s$ is arbitrary is given in [14]. Moreover generalizing the results due to Saeks and Murray [26], the problem of simultaneously stabilizing (partially pole assigning) a pair of single input single output plants is completely solved in [14]. Surprisingly however, as shown in [12], [13] and [14], in order to analyze the simultaneous stabilizability problem 1.1 for a triplet of single input single output systems, it is not enough to analyze the interpolation problem 1.3 and one asks the following transcendental problem.

Transcendental problem 1.4. "Given a t-tuple of rational functions $\delta_1(s), \cdots, \delta_t(s)$ from H, when does there exist a t tuple of rational functions $\Delta_1(s), \cdots, \Delta_t(s)$ in J such that

$$\Delta_1(s)\delta_1(s)+\cdots+\Delta_t(s)\delta_t(s)\equiv 0." \tag{1.1}$$

As shown in [11], a triplet of single input single output plants $x_1(s)/y_1(s)$, $x_2(s)/y_2(s)$, $x_3(s)/y_3(s)$ chosen generically in the topology described in [3] is simultaneously stabilizable iff there exists a triplet of stable, minimum phase rational functions $\Delta_1(s)$, $\Delta_2(s)$, $\Delta_3(s)$ such that

$$\begin{aligned}\Delta_1(s)(x_1(s)y_3(s)-x_3(s)y_1(s))+\Delta_2(s)(x_2(s)y_3(s)-x_3(s)y_2(s))\\ +\Delta_3(s)(x_2(s)y_1(s)-x_1(s)y_2(s))\equiv 0.\end{aligned} \tag{1.2}$$

It may be remarked that (1.2) is satisfied just in case the plant $[x_1(s)y_3(s)-x_3(s)y_1(s)]/[x_2(s)y_3(s)-x_3(s)y_2(s)]$ is partially pole assignable at the zeros of $[x_1(s)y_2(s)-x_2(s)y_1(s)]$ in $\mathbb{C}_u$ by a stable, minimum phase compensator, Δ_1/Δ_2. Thus we remark that the partial pole assignability problem by a stable minimum phase compensator is an important problem in system theory which is addressed, among others, by the transcendental problem 1.4.

This paper is organized as follows. In § 2 we parametrize a set of simultaneously stabilizable r-tuples of plants and show that provided $r\leqq\max(m, p)$, the above set is semialgebraic and dense in the set of r-tuples of plants. In § 3 we analyze square systems, i.e. when $m=p$ and show that the problem of simultaneously stabilizing r tuples ($r\geqq 3$) of $m\times m$ systems chosen generically is equivalent to the problem of partially pole assigning a $r-2$ tuple of $m\times m$ systems by a stable, minimum phase compensator. Consequently, we obtain a necessary condition for the simultaneous partial pole placement of three or more square systems chosen generically. Finally in § 4 we discuss in considerable details the case when $\min(m, p)=1$, i.e. when the number of inputs or the number of outputs is one. In § 5 we construct two folklore examples which do not pre-exist in the literature. Section 6 concludes this paper with a discussion on the prospects of this rapidly growing simultaneous system design methodology.

2. A generic and/or semialgebraic parametrization of the partially pole assignable r-tuples of $p\times m$ systems.

2.1. We would first like to consider the basic mathematical setup, for the details of which we refer to Saeks et al. [26], [27], Vidyasagar et al. [33], [34] and Desoer et al. [9]. Every single input single output system, viewed as an element of F, can be written as $x(s)/y(s)$, where $x(s)$, $y(s)\in H$. Similarly, a $p\times m$ multiinput multioutput plant $G(s)$ has the left coprime representation $D(s)^{-1}N(s)$, where $N(s)\in H^{p\times m}$, $D(s)\in H^{p\times p}$. Moreover a r-tuple of m input p output plants can be written as

$$G_i(s)=D_i(s)^{-1}N_i(s), \qquad i=1,\cdots,r \tag{2.1}$$

where $N_i(s) \in H^{p\times m}$, $D_i(s) \in H^{p\times p}$, for all i. Of course if n_i is the McMillan degree of $G_i(s)$, $i=1,\cdots,r$ respectively, we consider the space Σ of r tuples of plants given by

$$\Sigma = \Sigma_{m,p}^{n_1} \times \cdots \times \Sigma_{m,p}^{n_r} \tag{2.2}$$

where

$$\Sigma_{m,p}^{n_i} = \{p \times m \quad G_i(s);\ \text{degree } G_i(s) = n_i\}. \tag{2.3}$$

As has been described in [12], Σ is a quasiaffine algebraic variety in the affine space $\mathbb{R}^{(2n_1+1)mp} \times \cdots \times \mathbb{R}^{(2n_r+1)mp}$ and inherits the subspace topology.

Our main results of this paper concern the following pair of questions.

Question 2.1 (generic partial pole placement problem). Fix m, p, r and n_i. Let $\psi_1(s),\cdots,\psi_r(s) \in H$ be given. Is the set Σ_p, of r-tuples $G_1(s),\cdots,G_r(s)$ which can be simultaneously partially pole assigned at the zeros of $\psi_1(s),\cdots,\psi_r(s)$ respectively by some compensator, open and dense in Σ?

Question 2.2 (semialgebraic parametrization). Does there exist a semialgebraic subset $\mathcal{S}$ in Σ which is open and dense in Σ_p?

A semialgebraic (see [4] for details) set is a finite union and intersection of sets defined by algebraic equations and inequations. It is a classical result by Tarski [31] and Seidenberg [28], that the property of being semialgebraic is preserved by a rational map. Anderson, Bose and Jury [1] have used this fact to show that the set of plants, of a given McMillan degree, which can be stabilized by some feedback gain is an open, semialgebraic subset in the space of all plants.

Before we state and prove the main results of this section, we refer to the following two well-known results, which also illustrate the significance of semialgebraicity. Routh and Hurwitz parametrized the set of monic polynomials of degree n in one variable, with real co-efficients that have roots in the open left half of the complex plane. They showed that the above set of polynomials is semialgebraic in $\mathbb{R}^n$. This result is surprising because their proof involves complex analytic methods. As a second application of semialgebraicity we note the following. Youla, Bongiorno and Lu [35] considered the problem of parametrizing the set of plants in $\Sigma_{m,p}^n$ which can be stabilized by a stable compensator. Their result shows that the above set may be described by an interlacing condition involving the poles and the blocking zeros of the plant on the nonnegative real axis. It is therefore a semialgebraic subset of $\Sigma_{m,p}^n$. This result is again surprising because the compensators under consideration is of a priori unbounded McMillan degree and therefore "decision-algebra" methods [1] are not applicable. In fact the parametrization problem involves solving a matrix analogue of the interpolation problem 1.3 which is once again a problem in "complex analysis."

At present a semialgebraic parametrization, of the space $\pi_{i=1}^t [\Sigma_{1,1}^{n_i} \cap H]$ of t tuples of rational functions $\delta_1(s),\cdots,\delta_t(s)$ which satisfy (1.1) for some $\Delta_1,\cdots,\Delta_t$ in J, does not exist. Thus it is unknown if Σ_p is semialgebraic. It is therefore of interest to answer Question 2.2. To clarify further, it is not enough to know that there exists a set of r-tuples of plants $\mathcal{S}$ that can be partially pole assignable by some compensator, where $\mathcal{S}$ is open and dense (generic) in Σ. It is also of interest to know if there exists a means to describe $\mathcal{S}$ as well—in particular, is $\mathcal{S}$ semialgebraic in Σ?

If $r=1$, every plant is pole assignable and is therefore partially pole assignable by some compensator. In this section we assume $r>1$ and prove the following:

THEOREM 2.4. *Assume* $\psi_1(s),\cdots,\psi_r(s) \in H^{p\times p}$ *such that* $\det \psi_i(s)$, $i=1,\cdots,r$ *have zeros only in* $\mathbb{C}_u$ *and do not have a zero in common in* $\mathbb{C}_u$. *Assume*

$$1 < r \leqq \lfloor (m+p)/\min(m,p) \rfloor. \tag{2.4}$$

There exists an open, semialgebraic subset $\mathcal{S}$ in Σ which is open and dense in Σ_p, the set of partially, pole assignable r-tuples of proper plants in Σ. ($\lfloor n \rfloor$ is defined to be the largest integer less than or equal to n.)

Remark. In principle, one might obtain the semialgebraic set $\mathcal{S}$ using the elimination methods of Tarski [31] and Seidenberg [28]. Indeed, Byrnes and Anderson [6] have successfully used the concept of elimination, for the generic output feedback stabilization problem. For the purpose of computation, however, it is of interest to know explicitly the semialgebraic set $\mathcal{S}$, without going through "elimination" since it is known [10] to be computationally inefficient. Theorem 2.4 is a result on such an explicit parametrization as its proof would show.

It is of theoretical interest, however, to use "elimination methods" and prove the following:

THEOREM 2.5. *Assume $r>1$, $\psi_i(s)$, $i=1,\cdots,r$ belong to $H^{p\times p}$ and satisfy the condition mentioned in Theorem 2.4. A sufficient condition that there exists an open, dense, semialgebraic subset $\mathcal{S}$ (in Σ) of r-tuples of $p\times m$ proper plants $G_1(s),\cdots,G_r(s)$ of McMillan degrees n_i, $i=1,\cdots,r$ respectively which can be partially pole assigned at the zeros of* $\det \psi_i(s)$, $i=1,\cdots,r$ *respectively by an arbitrary large but finite proper compensator q is given by*

$$r \leqq \max(m,p). \tag{2.5}$$

Remark. Except for the case $\min(m,p)=1$ we note that the hypothesis (2.4) is stronger than (2.5).

It is interesting to note that by choosing $\det \psi_i(s) \equiv 1$, $i=1,\cdots,r$, we have the following corollary from Theorem 2.5.

COROLLARY 2.6 (Vidyasagar and Viswanadham [34]). *Assume* $\max(m,p)>1$. *A generic pair of $p\times m$ proper plants is simultaneously stabilizable.*

Note that the Theorems 2.4 and 2.5 describe only the sufficiency conditions. We now obtain the following necessary condition for the partial pole placement problem.

THEOREM 2.7. *Assuming $r>1$, $\psi_i(s)$, $i=1,\cdots,r$ as in Theorem 2.4. A necessary condition for generic simultaneous partial pole placement of a r-tuple of proper plants at the zeros of a generic r tuple of stable rational functions* $\det \psi_1(s),\cdots,\det \psi_r(s)$ *in $\mathbb{C}_u$ by some proper compensator of arbitrary large but finite McMillan degree q is given by*

$$q(m+p)+mp \geqq \sum_{i=1}^{r} \alpha_i \tag{2.6}$$

where α_i is the number of zeros of $\det \psi_i(s)$, $i=1,\cdots,r$ *respectively.*

Remark. The inequality (2.6) is not surprising and is precisely what one would guess from a "dimension count."

From Theorem 2.7 we now deduce the following.

COROLLARY 2.8. *Assume $\alpha_i>q$, $r>1$, $\psi_i(s)$, $i=1,\cdots,r$ as in Theorem 2.4. A necessary condition for generic simultaneous partial pole placement of a r-tuple of proper plants is given by*

$$r \leqq \max[mp/(\min_i \beta_i), m+p-1] \tag{2.7}$$

where $\alpha_i = q+\beta_i$, $i=1,\cdots,r$.

We remark that Corollary 2.8 includes the simultaneous pole placement problem as a special case when $\alpha_i = n_i+q$, $i=1,\cdots,r$. It is interesting to observe that (2.7) is independent of q and in particular we have the following.

COROLLARY 2.9 (Saeks and Murray [26], Vidyasagar and Viswanadham [34]). *Assume $m = p = 1$, $r = 2$. A generic pair of single input single output proper plants is not simultaneously pole assignable.*

Finally we note from Theorem 2.5 and Corollary 2.8 that the condition (2.5) is sharp in the following sense.

COROLLARY 2.10. *If $\min(m, p) = 1$, $r > 1$ and $\alpha_i > q$ and $\psi_i(s)$, $i = 1, \cdots, r$ as in Theorem 2.4, the inequality (2.5) is a necessary and sufficient condition for generic simultaneous partial pole placement of a r tuple of proper plants.*

To summarize the theorems and corollaries of this section, we make the following remark: If r is sufficiently small (given by (2.5)) a generic r tuple of multiinput and multioutput plants is simultaneously partially pole assignable. On the other hand under the assumptions of Corollary 2.8, if r is sufficiently large and fails to satisfy (2.7), a generic r tuple of multiinput multioutput plants is not simultaneously partially pole assignable. It may also be noted that for the generic simultaneous stabilization problem, one assumes $\alpha_i = 0$, $i = 1, \cdots, r$, so that Corollary 2.8 is inapplicable. In fact, no other necessary condition to the problem is known to exist. This constitutes an outstanding open problem in system theory.

We now sketch the proof of the theorems.

2.2. Proof of Theorem 2.4. First of all, we need to prove the following lemmas.

LEMMA 2.1. *Given a set of t vectors $v_1, \cdots, v_t$ in $\mathbb{R}^{m+p}$ with nonzero components. Then there exists another set of t vectors $u_1, \cdots, u_t$ such that u_i is orthogonal to v_i, for all $i = 1, \cdots, t$ and such that each component of u_i has the same sign for all $i = 1, \cdots, t$ respectively iff the set of vectors $v_1, \cdots, v_t, -v_1, \cdots, -v_t$ misses an orthant in $\mathbb{R}^{m+p}$.*

Proof. (If part). Let θ be the orthant which does not contain the vectors $v_1, \cdots, v_t$, $-v_1, \cdots, -v_t$. For a given vector v_i belonging to the orthant θ_1 it is well known that there exists a vector orthogonal to v_i in the orthant θ, since θ is different from θ_1, for all $i = 1, \cdots, t$. Thus in particular it is possible to choose $u_1, \cdots, u_t \in \theta$ such that $u_i \cdot v_i = 0$, $i = 1, \cdots, t$.

(Only if part). Assume that the set of vectors $v_1, \cdots v_t, -v_1, \cdots, -v_t$ does not miss any orthant in $\mathbb{R}^{m+p}$. It is clear that there does not exist any orthant θ in $\mathbb{R}^{m+p}$ such that $u_i \in \theta$ and $u_i \cdot v_i = 0$, for otherwise there would exist two nonzero vectors in the same orthant θ of $\mathbb{R}^{m+p}$ that are orthogonal to each other, which is absurd. Thus there does not exist a set of t vectors $u_1, \cdots, u_t$ with the above mentioned property. Q.E.D.

LEMMA 2.2. *Given a self conjugate set of tuples (s_i, M_i), $i = 1, \cdots, t$ where $s_i \in \mathbb{C}$ and M_i is a $p \times p$ complex matrix, $i = 1, \cdots, t$ respectively. Moreover for all $i \in \{1, \cdots, t\}$ for which $s_i \in \mathbb{R}_u$, assume that* $\det M_i > 0$. *There exists $\Delta(s) \in H^{p \times p}$, $\det \Delta(s) \in J$ such that*

$$\Delta(s_i) = M_i \tag{2.8}$$

for all $i = 1, \cdots, t$.

Remark. The main idea of Lemma 2.2 is originally due to Vidyasagar and Viswanadham [34] and the proof that we now sketch is an adaptation of their procedure.

Proof. Assume first of all that $\mathbb{C}_s$ contains the open left half of the complex plane. Let us define $N_p(s) \in H^{p \times p}$ such that

$$N_p(s_j) = 0, \qquad j = 1, \cdots, t \tag{2.9}$$

and that $N_c(s)$ has no other blocking zero in $\mathbb{C}_u$. A blocking zero is a point where the matrix function $N_p(s)$ vanishes as a matrix. Let $D_p(s)$ be such that

$$D_p(s_j) = M_j, \qquad j = 1, \cdots, t \tag{2.10}$$

and that $N_p(s)$, $D_p(s)$ are co-prime. It has been shown by Youla, Bongiorno and Lu [35] that the plant $N_p(s)D_p(s)^{-1}$ is stabilizable by a stable compensator, since at the zeros s_j, $j=1,\cdots,t$ of $N_p(s)$ in $\mathbb{C}_u$, $\det D_p(s_j)=\det M_j$, $j=1,\cdots,t$ have the same sign. Thus there exists $N_c(s)$, $D_c(s)$, $\Delta_1(s)\in H^{p\times p}$, $\det D_c(s)\in J$, $\det \Delta_1(s)\in J$ such that

$$N_c(s)N_p(s)+D_c(s)D_p(s)=\Delta_1(s). \tag{2.11}$$

Clearly we have

$$D_c^{-1}(s_j)\Delta_1(s_j)=D_p(s_j)=M_j, \qquad j=1,\cdots,t. \tag{2.12}$$

Defining $\Delta(s)=D_c^{-1}(s)\Delta_1(s)$, we have the required matrix function $\Delta(s)$.

In general if $\mathbb{C}_s$ does not contain the open left half of the complex plane, assume that there exists ball B_ε of radius ε with center at $c_1\in\mathbb{R}$ contained in $\mathbb{C}_s$. Note that since $\mathbb{C}_s$ intersects $\mathbb{R}$, such a ball would always exist. Let us now conformally transform B_ε onto the open left half of the complex plane by the map

$$\psi: z\mapsto\frac{s-c_1+\varepsilon}{s-c_1-\varepsilon}. \tag{2.13}$$

Define $z_j=(s_j-c_1+\varepsilon)/(s_j-c_1-\varepsilon)$, $j=1,\cdots,t$ and construct $\Delta(z)\in H^{p\times p}$, $\det\Delta(z)\in J$ such that $\Delta(z_j)=M_j$, $j=1,\cdots,t$, $\det\Delta(1)\neq 0$. It follows that $\Delta[(s-c_1+\varepsilon)/(s-c_1-\varepsilon)]$ is the required matrix function. Q.E.D.

LEMMA 2.3. *Given a pair of nonzero vectors* $(\alpha_1,\cdots,\alpha_p)$, $(\beta_1,\cdots,\beta_p)\in\mathbb{R}^p$, $p>1$ *there exists a* $m\times m$ *matrix* M *such that*

$$(\alpha_1,\cdots,\alpha_p)M=(\beta_1,\cdots,\beta_p) \tag{2.14}$$

and

$$\det M>0. \tag{2.15}$$

Proof. Let the ijth entry of M be given by a_{ij}. We may expand $\det M$ as

$$\det M=\sum_{i=1}^{p} a_{1i}\Delta_{1i} \tag{2.16}$$

where Δ_{1i} is the adjoint of the $1i$th entry in M. Assume without any loss of generality that $\alpha_1\neq 0$. One can rewrite (2.14) as follows:

$$a_{1i}=-\sum_{j=2}^{p}\frac{\alpha_j}{\alpha_1}a_{ji}+\frac{1}{\alpha_1}\beta_i. \tag{2.17}$$

Substituting a_{1i} in (2.16), we may check by straightforward computation that

$$\det M=\frac{1}{\alpha_1}\sum_{i=1}^{p}\beta_i\Delta_{1i}. \tag{2.18}$$

Clearly for a given α_1 and $\beta_1,\cdots,\beta_p$ we can write (2.18) as

$$\det M=a_{21}a'+a'' \tag{2.19}$$

where a', a'' are nonzero multinomial expressions in a_{ij}, $i=2,\cdots,p$; $j=1,\cdots,p$; independent of a_{21}. For any choice of $a_{22},\cdots,a_{2p},a_{31},\cdots,a_{3p},\cdots,a_{pp}$, such that $a'\neq 0$, it is possible to choose a_{21} such that $\det M>0$. The entries $a_{11},\cdots,a_{1p}$ are to be computed using (2.17) completing the proof. Q.E.D.

We would now proceed to prove Theorem 2.4. Assume $p \leqq m$ without any loss of generality. Let the r tuple of plants be given by (2.1). Let us represent the compensator as

$$N_c(s)D_c(s)^{-1} \tag{2.20}$$

where $N_c(s) \in H^{m\times p}$, $D_c(s) \in H^{p\times p}$, $N_c(s)$, $D_c(s)$ are coprime. It is well known that the compensator (2.20) partially pole assigns the r-tuple of plants (2.1) at the zeros of $\det \psi_i(s)$, $i=1,\cdots,r$ iff

$$N_i(s)N_c(s)+D_i(s)D_c(s)=\psi_i(s)\Delta_i(s) \tag{2.21}$$

for some $\Delta_i(s)$, $\psi_i(s) \in H^{p\times p}$, $\det \Delta_i(s) \in J$, $i=1,\cdots,r$ respectively. We may write (2.21) in matrix form as

$$\begin{bmatrix} N_1(s) & D_1(s) \\ \vdots & \vdots \\ N_r(s) & D_r(s) \end{bmatrix}\begin{bmatrix} N_c \\ D_c \end{bmatrix}=\begin{bmatrix} \psi_1(s)\Delta_1(s) \\ \vdots \\ \psi_r(s)\Delta_r(s) \end{bmatrix}. \tag{2.22}$$

Let us denote the $(rp)\times(m+p)$ matrix in the left-hand side of (2.22) by $M(s)$. Let us also define a set of r-tuples of plants $(G_1,\cdots,G_r)$ as follows.

(2.23) $\Omega=\{(G_1,\cdots,G_r) | M(s)$ in (2.22) has full rank for all points in $\mathbb{C}\}$.

Note first of all that Ω can be defined as union and intersection of sets of the type

$$f_\alpha(g_1,\cdots,g_r) \neq 0 \tag{2.24}$$

where $f_\alpha(\cdot)$ are polynomials in the coefficients of $g_1(s),\cdots,g_r(s)$. Since $f_\alpha \not\equiv 0$, (2.24) describes a union and intersection of open and dense subsets of Σ. Thus Ω is open, semialgebraic and dense in Σ, the space of r-tuples of plants. We now show that when $r < \lfloor (m+p)/\min(m,p) \rfloor$, $\Omega \subset \Sigma_p$ so that it is enough to define the required set $\mathcal{S}=\Omega$. On the other hand we consider the case $\min(m,p)>1$, $r\min(m,p)=m+p$ and the case $\min(m,p)=1$, $r=m+p$ and show the existence of an open, semialgebraic and dense subset $\mathcal{S}$ of Σ_p.

Case 1 ($r<\lfloor (m+p)/\min(m,p)\rfloor$). Since H is a principal ideal domain [38], for every r tuple of plants in Ω, $M(s)$ has a right inverse $M'(s)$ in $H^{(m+p)\times rp}$ [33] provided $rp \leqq m+p$. Thus for a r-tuple of plants in Ω, (2.22) is indeed solvable for $N_c(s)$, $D_c(s)$. We need to check, however, that the solution would yield $N_c(s)$, $D_c(s)$ which are coprime and that the compensator $N_c(s)D_c(s)^{-1}$ is proper.

Assume that $N_c(s)$, $D_c(s)$ are not coprime. It follows that there exists $s^* \in \mathbb{C}_u$ such that $\operatorname{rank}[N_c^T D_c^T]^T(s^*) < p$. It follows from (2.22) that $\det[\psi_i(s^*)\Delta_i(s^*)]=0$ for $i=1,\cdots,r$. Since $\det \Delta_i(s) \in J$, it follows that $\det[\psi_i(s^*)]=0$ for $i=1,\cdots,r$ which contradicts the assumption on $\psi_i(s)$, $i=1,\cdots,r$. Hence $N_c(s)$, $D_c(s)$ are coprime.

To see that $N_c(s)D_c(s)^{-1}$ is proper, it is enough to show that $\det D_c(\infty)\neq 0$. We claim, first of all, that for every r-tuple of plants in Ω, the associated matrix $M(s)$ has a right inverse $M^1(s)$ with the property that if $[P_1(s)\cdots P_r(s)]$ is the last p row of $M^1(s)$, (where $P_i(s)$, $i=1,\cdots,r$ are $p\times p$ matrices), $\det P_i(\infty)\neq 0$ for all $i=1,\cdots,r$. First of all, let us assume the claim. Since $\det \psi_j(\infty)\neq 0$ for some $j=1,\cdots,r$, it follows that $\psi_j(\infty)\Delta_j(\infty)$ is an arbitrary $p\times p$ matrix for some $j=1,\cdots,r$ and for an arbitrary choice of $\Delta_j(s)$. Since $D_c(s)=\sum_{k=1}^r P_k(s)\psi_k(s)$, it follows from (2.22) that for a suitable choice of $\Delta_1(s),\cdots,\Delta_r(s)$, $\det D_c(\infty)\neq 0$.

Now we proceed to prove the claim. If $M^1(s)$ does not satisfy the above property, let P be a $(m+p)\times rp$ real matrix which satisfies the above property. Let us define $M_\varepsilon^1(s)=M(s)+\varepsilon P$ where ε is a nonzero real number. Clearly for all but finitely many

values of ε, $M_\varepsilon^1(s)$ satisfies the above property. Moreover for small ε we have $\det[M(s)M_\varepsilon^1(s)] \in J$ since by construction $\det[M(s)M^1(s)] \in J$. Thus there exists $\varepsilon = \varepsilon^*$ such that $M_{\varepsilon^*}^1(s)$ is the required right inverse.

Case 2 ($\min(m, p) > 1$, $r \min(m, p) = m + p$). Assume without any loss of generality the case $p \leqq m$ and consider the worst case when $rp = m + p$. The case $rp < m + p$ is analogous to Case 1 above. Note that the matrix $M(s)$ in (2.22) is square and we can write

$$[N_c(s)^T D_c(s)^T]^T = M(s)^{-1}[\Delta_1^T \psi_1^T \cdots \Delta_r^T \psi_r^T]. \tag{2.25}$$

We now define the required subset $\mathcal{S}$ of Ω as follows:

$$\begin{aligned}\mathcal{S} \triangleq \{(G_1, \cdots, G_r) &| \det M(s) \text{ has simple zeros at } s_1, \cdots, s_t \text{ in } \mathbb{C} \text{ and} \\ &\text{does not vanish at } \infty; \text{ Adj } M(s_i),\ i = 1, \cdots, t \text{ are of rank 1; if the} \\ &\text{rows of Adj } M(s_i) \text{ are spanned by the nonzero vector } [\mathbf{c}_{i1}, \cdots, \mathbf{c}_{ir}] \text{ then} \\ &\mathbf{c}_{ij}\psi_j(s_i) \neq \mathbf{0}, \qquad i = 1, \cdots, t; j = 1, \cdots, r.\}\end{aligned} \tag{2.26}$$

It is clear that $\mathcal{S}$ is open, dense and semialgebraic in Ω. Moreover, in order that $N_c(s) \in H^{m \times p}$, $D_c(s) \in H^{p \times p}$, it is necessary and sufficient that

$$\text{Adj } M(s_i)[\Delta_1^T \psi_1^T(s_1), \cdots, \Delta_r^T \psi_r^T(s_i)]^T = 0 \tag{2.27}$$

for all s_i where $\det M(s_i) = 0$, $i = 1, \cdots, t$. It follows from the definition of $\mathcal{S}$ in (2.26) that a necessary and sufficient condition for the solvability of (2.27) is that

$$\sum_{j=1}^{r} \mathbf{c}_{ij}\psi_j(s_i)\Delta_j(s_i) = 0 \tag{2.28}$$

for $i = 1, \cdots, t$. That indeed (2.28) is satisfied by a set of $\Delta_j(s)$, $j = 1, \cdots, r$ follows from Lemma 2.2 and Lemma 2.3.

Case 3 ($\min(m, p) = 1$, $r = m + p$). The computation of Case 2 can be repeated so that we obtain an equation of the type (2.28). Since $\mathbf{c}_{ij}\psi_j(s_i) \neq 0$, it follows from Lemmas 2.1 and 2.2 that indeed there exist $\Delta_1, \cdots, \Delta_r(s)$ which satisfy equation (2.28), provided for $s_i \in \{s_1, \cdots, s_t\} \cap \mathbb{R}$, $i = 1, \cdots, t_1$ the following condition is satisfied: Let us define $\{s_1, \cdots, s_{t_1}\} \triangleq \{s_1, \cdots, s_t\} \cap \mathbb{R}_u$. The real vectors

$$v_i = [\mathbf{c}_{i,1}\psi_1(s_i), \cdots, \mathbf{c}_{i,r}\psi_r(s_i)] \tag{2.29}$$

for $i = 1, \cdots, t_1$ and $-v_i$, $i = 1, \cdots, t_1$ miss an orthant in $\mathbb{R}^{m+p}$. Since "missing an orthant" in $\mathbb{R}^{m+p}$ may be described by open semialgebraic conditions, it follows that the partially pole assignable r tuples of plants in $\mathcal{S}$ is open, semialgebraic. Q.E.D.

Remark. The distinction between the results in Case 2 and Case 3 of Theorem 2.4 is to be noted. In Case 2, we claim that every r tuple of plants in the semialgebraic set $\mathcal{S}$ is simultaneously partially pole assignable. In Case 3, we only claim that the set of simultaneously partially pole assignable plants in $\mathcal{S}$ is semialgebraic and is given precisely by the above condition of "missing an orthant."

2.3. Proof of Theorem 2.5. The case $\min(m, p) = 1$ follows from Case 1 of Theorem 2.4. The case $\min(m, p) > 1$ proceeds by a reduction to the case $\min(m, p) = 1$ by a procedure called "vectoring down" adopted from Stevens' thesis [29] and from Brasch-Pearson [5]. The following two lemmas are now stated without proof (see [12], [18] for a proof).

LEMMA 2.4. *Given a r-tuple of $p \times m$ plants $G_i(s)$ of degrees n_i, each with n_i simple poles. There is an open dense set of $1 \times p$ vectors $v \in \mathbb{R}^p$ such that $vG_i(s)$ has degree n_i, for all i.*

LEMMA 2.5. *Given a r-tuple of $p \times m$ plants $G_i(s)$, $i = 1, \cdots, r$. There exists a constant gain output feedback K such that the closed-loop systems $G_i(s)[I + KG_i(s)]^{-1}$ have distinct simple poles.*

Thus by choosing any $(v, K) \in \mathbb{R}^p \times \mathbb{R}^{mp}$, we have a mapping

$$\phi: \Sigma \times \mathbb{R}^p \times \mathbb{R}^{mp} \to \Sigma_1, \tag{2.30}$$

$$\phi(G_1(s), \cdots, G_r(s), v, k) = (vG_i(s)[I + KG_i(s)]^{-1})_{i=1}^{r} \tag{2.31}$$

where Σ is the space of r-tuples of $p \times m$ plants and Σ_1 is the space of r-tuples of $1 \times m$ plants.

Since ϕ is rational in the coefficients of $G_1, \cdots, G_r, v, K$; the inequalities (2.24) together with the mapping ϕ would define semialgebraic, open, dense subset Ω_1 of $\Sigma \times \mathbb{R}^p \times \mathbb{R}^{mp}$ given by

$$f_\alpha(vG_1[I + KG_1]^{-1}, \cdots, vG_r[I + KG_r]^{-1}) \neq 0. \tag{2.32}$$

By considering the projection

$$\text{Proj}: \Sigma_{m,p}^{n_1} \times \cdots \times \Sigma_{m,p}^{n_r} \times \mathbb{R}^p \times \mathbb{R}^{mp} \to \Sigma_{m,p}^{n_1} \times \cdots \times \Sigma_{m,p}^{n_r} \tag{2.33}$$

we have the set

$$\mathcal{S} \triangleq \text{Proj}\, \Omega_1. \tag{2.34}$$

By the Tarski [31]-Seidenberg [28] theory of elimination over $\mathbb{R}$, $\mathcal{S}^1$ is semialgebraic. Moreover $\mathcal{S}^1$ is dense, since Ω_1 is dense. Thus $\mathcal{S}^1$ is semialgebraic, dense and is defined by union and/or intersection of sets given by polynomial equations or inequality

$$g_\alpha > 0 \quad \text{and} \quad [g_{\beta_1} > 0 \text{ or } g_{\beta_2} = 0] \quad \text{and} \quad g_\gamma = 0. \tag{2.35}$$

Since $g_\gamma(\mathcal{S}^1) = 0 \Rightarrow g_\gamma \equiv 0$, hence $\mathcal{S}$ defined by union and/or intersection of sets of the type $[g_\alpha > 0$ and $g_{\beta_1} > 0]$ is semialgebraic, open and dense in Σ. Note that $\mathcal{S}$ is dense in $\mathcal{S}^1$ since $\mathcal{S}$ is obtained by removing the set $[g_{\beta_2} \neq 0]$ from $\mathcal{S}^1$.

Thus a max (m, p) tuple of plants in $\mathcal{S}$ can be partially pole assigned by the vectoring down technique. Q.E.D.

2.4. Proof of Theorem 2.7. Let $G_1(s), \cdots, G_r(s)$ be the given r tuple of proper plants in Σ. Let $K(s)$ be a compensator in $\Sigma_{p,m}^{q}$. The associated return difference equation, $\det[I + K(s)G_i(s)] = 0$ is given by

$$\Pi_i(s) \overset{\Delta}{\equiv} \sum_{j=0}^{n_i+q} c_{ij} s^j \quad \text{for all } i = 1, 2, \cdots, r. \tag{2.36}$$

A generic r-tuple of plants defines a smooth mapping χ, between the compensator parameters and the coefficient of the return difference polynomials given by

$$\chi: \Sigma_{p,m}^{q} \to \mathbb{RP}^{n_1+q+1} \times \cdots \times \mathbb{RP}^{n_r+q+1}, \tag{2.37}$$

$$\chi(K) = ([c_{1,0}, \cdots, c_{1,n_1+q}], \cdots, [c_{r,0}, \cdots, c_{r,n_r+q}]) \tag{2.38}$$

where the right-hand side of (2.38) has been defined in homogeneous co-ordinates. It is well known [7], [8], [19] that $\Sigma_{p,m}^{q}$ is a manifold of dimension $q(m+p) + mp$. To say that there exists a compensator which partially pole assigns a generic r set of α_i self conjugated poles is to say that image χ contains a $\sum_{i=1}^{r} \alpha_i$ dimensional submanifold of $\mathbb{RP}^{n_1+q+1} \times \cdots \times \mathbb{RP}^{n_r+q+1}$. By Sard's theorem [23] a necessary condition is given by (2.6) concluding the proof. Q.E.D.

2.5. Proof of the corollaries. The Corollary 2.6 is immediate from Theorem 2.5 assuming $\psi_i(s) \equiv 1$, $i = 1, \cdots, r$. We now prove the Corollary 2.8. Substituting $\alpha_i = q + \beta_i$ in (2.6), we have

$$q(m+p-r)+mp \geqq \sum_{i=1}^{r} \beta_i \geqq r \min_i \beta_i. \tag{2.39}$$

A necessary condition that there exists some $q \in \mathbb{N}$ satisfying (2.39) is given by

$$r \leqq m+p-1 \quad \text{or} \quad r \min_i \beta_i \leqq mp. \tag{2.40}$$

Thus (2.7) is indeed a necessary condition concluding the proof of Corollary 2.8. Corollary 2.9 is a consequence of Corollary 2.8. Finally, the Corollary 2.10 follows from Theorems 2.5, 2.7 and Corollary 2.8.

3. Simultaneous partial pole placement of square systems.

3.1. The purpose of this section is to analyze the partial pole placement of a square system and obtain various equivalent characterization of the problem. Results of this type has been originally obtained by Saeks and Murray in [26] where they have shown that the problem of simultaneously stabilizing a pair of single input single output plants is equivalent to the problem of stabilizing a plant by a stable compensator. Likewise, it has been shown by Vidyasagar and Viswanadham [34] that the problem of simultaneously stabilizing r multiinput multioutput plants is equivalent to the problem of stabilizing $r-1$ plants by a stable compensator.

In this section we make contact with the "matrix version" of the transcendental problem described in the introduction. Adapting the proof of the "strong stabilization problem" in [35], we obtain a necessary condition for the solvability of the transcendental problem. We remark, however, that a necessary and sufficient condition for the transcendental problem is unknown, although for the scalar case, a separate necessary condition and a sufficient condition has been reported in [14]. The necessary condition also appears in the proof of Theorem 2.4 (Case 3). An interesting zero interlacing property of the necessary condition is described in § 5.

Let us now prove the following:

THEOREM 3.1. *A necessary and sufficient condition for the simultaneous partial pole placement of a r tuple* ($r \geqq 3$) *of* $m \times m$ *plants* $G_1(s), \cdots, G_r(s)$, *chosen generically, at the zeros of* $\det \psi_1(s), \cdots, \det \psi_r(s)$ *in* $\mathbb{C}_u$ *by a nonswitching* $m \times m$ *compensator is given by the existence of* $\Delta_1(s), \cdots, \Delta_r(s) \in H^{m\times m}$, $\det \Delta_i(s) \in J$, $i = 1, \cdots, r$ *satisfying*

$$[\Delta_1\psi_1(s) \quad \Delta_2\psi_2(s)]\begin{bmatrix} N_1(s) & N_2(s) \\ D_1(s) & D_2(s) \end{bmatrix}^{-1}\begin{bmatrix} N_i(s) \\ D_i(s) \end{bmatrix} = \Delta_i(s)\psi_i(s) \tag{3.1}$$

for $i = 3, \cdots, r$. *Here* $G_i(s) = N_i(s)D_i(s)^{-1}$ *is a coprime fraction representation of the ith plant and* $\psi_1(s), \cdots, \psi_r(s) \in H^{m\times m}$ *with the property that* $\det \psi_i(s)$, $i = 1, \cdots, r$ *do not have a common zero in* $\mathbb{C}_u$ *and have zeros only in* $\mathbb{C}_u$.

Remark. Let us define

$$[\mathcal{K}_{i1}^T(s) \quad \mathcal{K}_{i2}^T(s)]^T \triangleq \left\{\operatorname{Adj}\begin{bmatrix} N_1(s) & N_2(s) \\ D_1(s) & D_2(s) \end{bmatrix}\right\}\begin{bmatrix} N_i(s) \\ D_i(s) \end{bmatrix}, \tag{3.2}$$

$$\mathcal{K}_3(s) = \det\begin{bmatrix} N_1(s) & N_2(s) \\ D_1(s) & D_2(s) \end{bmatrix} \tag{3.3}$$

and rewrite (3.1) as

$$\Delta_1\psi_1\mathcal{K}_{i1} + \Delta_2\psi_2\mathcal{K}_{i2} = \Delta_i\psi_i\mathcal{K}_3, \tag{3.4}$$

$i=3,\cdots,r$. The matrix version of the transcendental Problem 1.4 is to solve (3.4) for a suitable $\Delta_i(s)$, where $\det \Delta_i(s) \in J$, $i=1,\cdots,r$. We now state the following necessary condition for the solvability of (3.4).

THEOREM 3.2. *Let $s_1,\cdots,s_t$ be a set of blocking zeros of $\psi_{i_0}(s)$, $i_0 \in \{3,\cdots,r\}$ in a connected component of $\mathbb{R}_u$. A necessary condition that the r-tuple of plants is simultaneously partially pole assignable at the zeros of* $\det \psi_i(s)$ *in $\mathbb{C}_u$, $i=1,\cdots,r$ respectively (where* $\det \psi_i(s)$, $i=1,\cdots,r$ *do not have a common zero in $\mathbb{C}_u$) is given by*

(3.5) "sign $[\det \psi_1 \mathcal{K}_{i_01}(s_j)] \times [\det \psi_2 \mathcal{K}_{i_02}(s_j)$ *is the same for $j=1,\cdots,t$.*"

3.2. Proof of Theorem 3.1. Let $K(s)=D_c(s)^{-1}N_c(s)$ be the coprime representation of the required compensator. (only if): Assume simultaneous partial pole assignability of the r-tuple of plants $G_1,\cdots,G_r$ by the compensator $K(s)$. Clearly there exists $\Delta_1(s),\cdots,\Delta_r(s) \in H^{m\times m}$, $\det \Delta_i(s) \in J$, $i=1,\cdots,r$ respectively such that

$$N_c(s)N_i(s)+D_c(s)D_i(s)=\Delta_i(s)\psi_i(s) \tag{3.6}$$

for all $i=1,\cdots,r$. Eliminating $D_c(s)$ and $N_c(s)$ from (3.6), we obtain (3.1).

(if): Assume that there exists $\Delta_1(s),\cdots,\Delta_r(s) \in H^{m\times m}$, $\det \Delta_i(s) \in J$ satisfying (3.1). Let us denote

$$P(s)=\begin{bmatrix} N_1(s) & N_2(s) \\ D_1(s) & D_2(s) \end{bmatrix}, \tag{3.7}$$

$$[N_c(s) \quad D_c(s)] \triangleq [\Delta_1\psi_1(s) \quad \Delta_2\psi_2(s)]P(s)^{-1}. \tag{3.8}$$

We need to show that $N_c(s)$, $D_c(s) \in H^{m\times m}$; $N_c(s)$, $D_c(s)$ are coprime, $N_c(s)$, $D_c(s)$ satisfy (3.6) and finally $D_c^{-1}N_c(s)$ is proper. Assume generically that

(3.9) a. $P(s)$ has simple zeros in $\mathbb{C}_u$ at $s_1,\cdots,s_t$,

(3.10) b. Adj $P(s_i)$, $i=1,\cdots,t$ are of rank 1,

(3.11) c. If the rows of Adj $P(s_i)$ are spanned by $[r_{i,1}\cdots r_{i,2m}]$, then
$$[r_{i1} \quad r_{i,2m}][N_j^T \quad D_j^T]^T(s_i) \neq 0 \quad \text{for } j=3,\cdots,r,$$

(3.12) d. $\det P(\infty) \neq 0$.

To see that $N_c(s)$, $D_c(s) \in H^{m\times m}$, we consider the following. From (3.1), (3.9) it follows that

$$[\Delta_1\psi_1(s_1) \quad \Delta_2\psi_2(s_i)] \text{ Adj } P(s_i)[N_j(s_i)^T \quad D_j(s_i)^T]^T = 0 \tag{3.13}$$

for $i=1,\cdots,t$; $j=3,\cdots,r$. From (3.10), (3.11) and (3.13) one infers that

$$[\Delta_1\psi_1(s_i) \quad \Delta_2\psi_2(s_i)] \text{ Adj } P(s_i) = 0. \tag{3.14}$$

From (3.8), (3.9) and (3.14) it follows that $N_c(s)$, $D_c(s) \in H^{m\times m}$. Moreover from (3.1), (3.8), it is clear that N_c, D_c satisfy (3.6). Also $N_c(s)$, $D_c(s)$ are coprime for otherwise there exists $s^* \in \mathbb{C}_u$ such that $[N_c(s^*), D_c(s^*)]$ has rank $< m$. This implies that for every $i=1,\cdots,r$,

$$\det [N_c(s)N_i(s)+D_c(s)D_i(s)] \tag{3.15}$$

vanishes at s^*. It follows that $\det \psi_i(s^*)=0$, which contradicts the hypothesis on $\psi_i(s)$, $i=1,\cdots,r$. Finally from (3.8) and (3.12) it follows that $D_c^{-1}N_c$ is proper. Q.E.D.

3.3. Proof of Theorem 3.2. From (3.4) one infers that

$$\Delta_1(s_j)\psi_1\mathcal{H}_{i_01}(s_j) = -\Delta_2(s_j)\psi_2\mathcal{H}_{i_02}(s_j) \tag{3.16}$$

for all $j = 1, \cdots, t$. Thus

$$\text{sign}\,[\det \Delta_1(s_j) \times \det \Delta_2(s_j)] = \text{sign}\,[\det (\psi_1\mathcal{H}_{i_01}(s_j))\ \det (\psi_2\mathcal{H}_{i_02}(s_j))]. \tag{3.17}$$

The proof now follows from the observation that the sign of the left-hand side of (3.17) does not change for $j = 1, \cdots, t$. Q.E.D.

4. Simultaneous partial pole placement of r min $(m, p) = 1$ plants.

4.1. In this section we consider the simultaneous stabilization problem of r single input or single output systems. The case $r \leqq \max(m, p)$ has been considered in Theorem 2.5. Here we restrict attention to $r > \max(m, p)$. First of all we parametrize the set of all compensators simultaneously stabilizing a generic m tuple of $1 \times m$ plants and prove the following.

THEOREM 4.1. *The problem of simultaneously stabilizing a $m+p$ tuple of* min $(m, p) = 1$ *systems is equivalent to the problem of stabilizing a* min $(m, p) = 1$ *system by a minimum phase compensator.*

In particular for $m = p = 1$ we obtain the following corollary.

COROLLARY 4.2. *The following three problems are equivalent.*

(i) *The problem of stabilizing a pair of single input single output systems.*

(ii) *The problem of stabilizing a single input single output system by a stable compensator.*

(iii) *The problem of stabilizing a single input single output system by a minimum phase compensator.*

Of course the equivalence between (i) and (ii) was originally obtained by Saeks and Murray [26], Vidyasagar and Viswanadham [34].

The importance of the partial pole placement problem is further emphasized by the following.

THEOREM 4.3. *Assume* min $(m, p) = 1$. *The problem of simultaneously stabilizing $m+p+k$, systems (assume $k \geqq 1$) chosen generically is equivalent to the problem of simultaneous partial pole placement of k systems by a stable minimum phase compensator.*

In particular for $m = p = k = 1$ we obtain

COROLLARY 4.4. (Ghosh [14]). *The problem of simultaneously stabilizing 3 single input single output plants chosen generically is equivalent to the problem of partially pole placing one single input single output plants by a stable minimum phase compensator.*

With reference to Theorem 4.3, we remark that the simultaneous partial pole placement problem of a single input single output plant by a stable, minimum phase compensator is precisely the transcendental Problem 1.4 and has been analyzed by Ghosh [14]. In particular conditions have been obtained which are separately necessary and sufficient. These conditions can be analogously stated for the min $(m, p) = 1$ case and we refer to [12] for details.

4.2. Parametrizing the set of all compensators simultaneously stabilizing a generic m-tuple of $1 \times m$ systems. Let us represent an m-tuple of $1 \times m$ plants as

$$g_i(s) = \left[\frac{x_{i,1}(s)}{x_{i,m+1}(s)}\ \frac{x_{i,2}(s)}{x_{i,m+1}(s)} \cdots \frac{x_{i,m}(s)}{x_{i,m+1}(s)}\right] \tag{4.1}$$

where $x_{i,j}(s) \in H$, $i = 1, \cdots, m$; $j = 1, \cdots, m+1$. Let us also represent a $m \times 1$ compensator as

$$k(s) = \left[\frac{y_1(s)}{y_{m+1}(s)}\ \frac{y_2(s)}{y_{m+1}(s)} \cdots \frac{y_m(s)}{y_{m+1}(s)}\right]^T \tag{4.2}$$

where $y_j(s) \in H$, $j = 1, \cdots, m+1$. To say that the above m-tuple of systems (4.1) is stabilizable by the compensator (4.2) is to say that there exists $\Delta_1(s), \cdots, \Delta_m(s) \in J$ such that

$$\sum_{j=1}^{m+1} x_{i,j}(s) y_j(s) = \Delta_i(s) \tag{4.3}$$

$i = 1, \cdots, m$. Let

$$\mathbf{y}^h(s) = [y_1^h(s) \cdots y_{m+1}^h(s)] \tag{4.4}$$

be the solution of the homogeneous equation

$$\sum_{j=1}^{m+1} x_{i,j}(s) y_j(s) = 0 \tag{4.5}$$

$i = 1, \cdots, m$ and

$$\mathbf{y}^{p_i} = [y_1^{p_i}(s), \cdots, y_{m+1}^{p_i}(s)] \tag{4.6}$$

be a particular solution of the equation.

$$\begin{bmatrix} x_{1,1}(s) & \cdots & x_{1,m+1}(s) \\ \vdots & & \vdots \\ x_{m,1}(s) & & x_{m,m+1}(s) \end{bmatrix} \cdot \begin{bmatrix} y_1(s) \\ \vdots \\ y_{m+1}(s) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \leftarrow i\text{th spot.} \tag{4.7}$$

A complete solution of (4.3) is given by

$$\mathbf{y}(s) = \delta(s)\mathbf{y}^h(s) + \sum_{i=1}^{m} \Delta_i(s)\mathbf{y}^{p_i}(s) \tag{4.8}$$

where $\delta(s) \in H$, $\Delta_i(s) \in J$ for $i = 1, \cdots, m$. The set of all compensators simultaneously stabilizing the m tuple of $\min(m, p) = 1$ systems is given by

$$\left[\frac{\delta(s)y_1^h(s) + \sum_{i=1}^m \Delta_i(s) y_1^{p_i}(s)}{\delta(s)y_{m+1}^h(s) + \sum_{i=1}^m \Delta_i(s) y_{m+1}^{p_i}(s)}, \cdots, \frac{\delta(s)y_m^h(s) + \sum_{i=1}^m \Delta_i(s) y_m^{p_i}(s)}{\delta(s)y_{m+1}^h(s) + \sum_{i=1}^m \Delta_i(s) y_{m+1}^{p_i}(s)} \right] \tag{4.9}$$

for some $\delta(s) \in H$, $\Delta_i(s) \in J$, $i = 1, \cdots, m$.

4.3. Proof of Theorem 4.1. Consider a set of $m+1$ tuple of $1 \times m$ plants given by (4.1). The set of all compensators that would stabilize a m-tuple of $1 \times m$ plants is given by (4.9). In order that it also stabilizes the $(m+1)$th plant, we need to satisfy the equation

$$\sum_{j=1}^{m+1} x_{m+1,j}(s) \left[\delta(s) y_j^h(s) + \sum_{i=1}^{m} \Delta_i(s) y_j^{p_i}(s) \right] = \Delta_{m+1}(s) \tag{4.10}$$

for some $\Delta_1, \cdots, \Delta_{m+1} \in J$, $\delta(s) \in H$.

Equation (4.10) can be rewritten as

$$\delta(s) \left[\sum_{j=1}^{m+1} x_{m+1,j}(s) y_j^h(s) \right] + \sum_{i=1}^{m} \Delta_i(s) \left[\sum_{j=1}^{m+1} x_{m+1,j}(s) y_j^{p_i}(s) \right] = \Delta_{m+1}(s). \tag{4.11}$$

Thus if a $m+1$ tuple of plants is simultaneously stabilizable, the plant

$$\left[\frac{\sum_{j=1}^{m+1} x_{m+1,j}(s) y_j^{p_1}(s)}{\sum_{j=1}^{m+1} x_{m+1,j}(s) y_j^h(s)} \cdots \frac{\sum_{j=1}^{m+1} x_{m+1,j}(s) y_j^{p_m}(s)}{\sum_{j=1}^{m+1} x_{m+1,j}(s) y_j^h(s)} \right] \tag{4.12}$$

is stabilizable by the minimum phase compensator

$$\left[\frac{\Delta_1(s)}{\delta(s)}, \cdots, \frac{\Delta_m(s)}{\delta(s)}\right]. \tag{4.13}$$

Conversely the minimum phase compensator (4.13) also stabilizes the $1 \times m$ plants

$$[kl_1(s), 0, \cdots, 0], [0, kl_2(s), 0, \cdots, 0], \cdots, [0, 0, \cdots, 0, kl_m(s)] \tag{4.14}$$

for $l_i(s) \in J$ and where $k \in \mathbb{R}$ is to be chosen arbitrarily large. □

4.4. Proof of Theorem 4.3. First of all, we shall need the following lemma.

LEMMA 4.1. *The problem of simultaneously stabilizing a pair of* $\min(m, p) = 1$ *systems chosen generically by a minimum phase compensator is equivalent to the problem of partially pole placing a* $\min(m, p) = 1$ *system by a stable, minimum phase compensator.*

Proof. As before we consider a pair of $1 \times m$ systems given by (4.1). Assume that these two systems are stabilizable by a minimum phase compensator (4.2) where $y_i(s) \in J$, $i = 1, \cdots, m$ and $y_{m+1}(s) \in H$, and

$$\begin{aligned} x_{1,1}(s)y_1(s) + \cdots + x_{1,m+1}(s)y_{m+1}(s) &= \Delta_1, \\ x_{2,1}(s)y_1(s) + \cdots + x_{2,m+1}(s)y_{m+1}(s) &= \Delta_2. \end{aligned} \tag{4.15}$$

Assume $x_{1,m+1}(s) \not\equiv 0$, $x_{2,m+1}(s) \not\equiv 0$; we have

$$\begin{aligned} y_{m+1}(s) &= \frac{1}{x_{1,m+1}(s)}\Delta_1(s) - \sum_{j=1}^{m} \frac{x_{1,j}(s)}{x_{1,m+1}(s)} y_j(s) \\ &= \frac{1}{x_{2,m+1}(s)}\Delta_2(s) - \sum_{j=1}^{m} \frac{x_{2,j}(s)}{x_{2,m+1}(s)} y_j(s) \end{aligned} \tag{4.16}$$

or

$$\begin{aligned} x_{2,m+1}(s)\Delta_1(s) - \sum_{j=1}^{m} x_{2,m+1}(s)x_{1,j}(s)y_j(s) &= x_{1,m+1}(s)\Delta_2(s) \\ &\quad - \sum_{j=1}^{m} x_{1,m+1}(s)x_{2,j}(s)y_j(s) \end{aligned} \tag{4.17}$$

or equivalently

$$x_{2,m+1}(s)\Delta_1(s) + \sum_{j=1}^{m} [x_{1,m+1}(s)x_{2,j}(s) - x_{2,m+1}(s)x_{1,j}(s)]y_j(s) = x_{1,m+1}(s)\Delta_2(s). \tag{4.18}$$

We now claim the following. To say that (4.15) is solvable is to say that (4.18) is solvable for some $\Delta_1(s)$, $\Delta_2(s)$, $y_j(s) \in J$, $j = 1, \cdots, m$. Since (4.18) is the algebraic condition for the partial pole placement of a $\min(m, p) = 1$ plant, we have the lemma. Conversely note that if (4.18) is solvable then there exists a $y_{m+1}(s) \in H$ defined by (4.16) which satisfies (4.15), provided there does not exist $s^* \in \mathbb{C}_u$ where $x_{1,m+1}$, $x_{2,m+1}$, $(x_{1,m+1}x_{2,j} - x_{2,m+1}x_{1,j})$, $j = 1, \cdots, m$ vanish. Q.E.D.

The proof of Theorem 4.3 now follows by the following argument. Let $g_1, \cdots, g_{m+p+k}$ be a $m+p+k$ tuple of $\min(m, p) = 1$ systems. Consider the following k, $m+p+1$ tuples of systems

$$(g_1, \cdots, g_{m+p}, g_{m+p+1}), \cdots, (g_1, \cdots, g_{m+p}, g_{m+p+j}).$$

By Theorem 4.1 and Lemma 4.1, stabilizability of each of the above $m+p+1$ tuple of plants is equivalent to the partial pole assignment of a $\min(m, p) = 1$ plant by a stable

minimum phase compensator. Thus to say that all the k, $m+p+1$ tuples are stabilizable is to say that a k tuple of min $(m, p) = 1$ plants are partially pole assignable by a stable, minimum phase compensator. Q.E.D.

5. Examples. In this section we construct the following two examples:

5.1. *Example* 1. Assume $\mathbb{C}_u$ to be the closed right half of the complex plane. Consider the following triplet of single input single output plants of McMillan degrees 1.

$$\frac{s-7}{s-4.6}, \quad \frac{s-2}{2s-2.6}, \quad \frac{s-6}{4.8s-24.6}. \tag{5.1}$$

We claim that every pair of plants in the above triplet are simultaneously stabilizable. This may be trivially checked, using the necessary and sufficient condition due to Saeks and Murray [26]. On the other hand, the above triplet is simultaneously stabilizable iff there exists $\Delta_1(s)$, $\Delta_2(s)$, $\Delta_3(s) \in J$ such that

$$\begin{aligned}&\Delta_1(s)[2.8(s-4)(s-3)]/[(s-1)(s-9)]\\ &\quad+\Delta_2(s)[(-3.8s^2+47.6s-144.6)]/[(s-1)(s-9)]+\Delta_3(s)\equiv 0.\end{aligned} \tag{5.2}$$

A necessary condition that (6.2) is satisfied is that the vectors

$$\pm\left[\frac{2.8(s-4)(s-3)}{(s-1)(s-9)} \quad \frac{-3.8s^2+47.6s-144.6}{(s-1)(s-9)} \quad 1\right] \tag{5.3}$$

miss an orthant in $\mathbb{R}^3$ for all $s \in \mathbb{C}_u$. For $s = 0$, 2, 3.5, 6 the above vector (5.3) may be computed as

$$\begin{aligned}&\pm[3.73 \quad -16.066 \quad 1], \quad \pm[-.8 \quad 9.2286 \quad 1],\\ &\pm[.0501 \quad 1.7854 \quad 1], \quad \pm[-1.12 \quad -.28 \quad 1],\end{aligned}$$

respectively. Thus it follows that the triplet of plants in (5.1) is simultaneously unstabilizable.

Remark. The transcendental problem which arises in the simultaneous stabilization of a triplet of plants $x_i(s)/y_i(s)$, $i = 1, 2, 3$ is given by (1.2). A necessary condition that (1.2) is solvable is that every pair in the above triplet is simultaneously stabilizable. If the triplet is chosen generically, then it is easy to show that: "every pair in the above triplet is simultaneously stabilizable iff the total number of zeros of $\eta_{ij}(s)$, $\eta_{ik}(s)$ in between every two consecutive zeros of $\eta_{jk}(s)$ in any connected component of $\mathbb{R}_u$ is even, for $i, j, k \in \{1, 2, 3\}$, $i \neq j \neq k$, where $\eta_{ij} = x_i y_j - x_j y_i$." The three plants in (5.1) are an example of such a triplet. It is interesting however that even when the above interlacing condition is satisfied, if the zeros of η_{ij}, η_{ik}, η_{jk} are interlaced in a connected component of $\mathbb{R}_u$ as shown in Fig. 5.1, the equation (1.2) cannot be solved and the triplet under consideration is simultaneously unstabilizable.

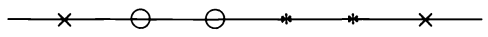


FIG. 5.1. *The zeros of η_{ij}, η_{jk}, η_{ik} in a connected component of $\mathbb{R}_u$ are denoted by "×", "○" and "*". The zero distribution represents an example of a triplet of plants that are not simultaneously stabilizable but are simultaneously stabilizable in pairs.*

Example 2. Consider the following triplet of 1×3 plants

$$\left[\frac{-26}{18s-70}\quad\frac{54}{18s-70}\quad\frac{18s-42}{18s-70}\right], \tag{5.4}$$

$$\left[\frac{9s+10}{9s+2}\quad\frac{-9s}{9s+2}\quad\frac{9s+12}{9s+2}\right], \tag{5.5}$$

$$\left[\frac{-9s+9}{9s-9}\quad\frac{9s+9}{9s-9}\quad\frac{9s+9}{9s-9}\right]. \tag{5.6}$$

We now show that the above triplet is not simultaneously stabilizable.

Suppose not, and let (4.2) be the stabilizing compensator. Clearly we need to satisfy

$$\begin{bmatrix} \frac{-26}{s-a} & \frac{-54}{s-a} & \frac{-18s+42}{s-a} & \frac{-18s+42}{s-a} \\ \frac{9s+10}{s-a} & \frac{-9s}{s-a} & \frac{-9s}{s-a} & \frac{9s+2}{s-a} \\ \frac{-9s+9}{s-a} & \frac{9s+9}{s-a} & \frac{9s+9}{s-a} & \frac{9s-9}{s-a} \end{bmatrix} \begin{bmatrix} y_1(s) \\ y_2(s) \\ y_3(s) \\ y_4(s) \end{bmatrix} = \begin{bmatrix} \Delta_1(s) \\ \Delta_2(s) \\ \Delta_3(s) \end{bmatrix} \tag{5.7}$$

for some $y_i(s)\in H$, $i=1,2,3,4$ and $\Delta_i(s)\in J$, $i=1,2,3$. In (5.7) "a" is chosen in such a way that $s-a$ vanishes in $\mathbb{C}_s$. It may be checked that the rows of the matrix in (5.7) are linearly dependent so that Δ_1, Δ_2, Δ_3 must satisfy

$$\Delta_1(s)[(s^2-1)/(s^2-s-6)]+\Delta_2(s)[(s^2-6s+8)/(s^2-s-6)]+\Delta_3(s)=0. \tag{5.8}$$

By evaluating the vector

$$[(s^2-1)\quad(s^2-6s+8)\quad(s^2-s-6)] \tag{5.9}$$

at $s=1.5$, 2.5, 3.5, 4.5 and using Lemma 2.1, we may infer that the triplet (5.4), (5.5), (5.6) is not simultaneously stabilizable.

Remark. Considering the fact that a generic triplet of 1×3 systems is simultaneously stabilizable, the triplet of plants in Example 2 is an "exceptional triplet." This serves to illustrate that the transcendental technique introduced in this paper can analyze "nongeneric" problems as well.

6. Conclusion. In this paper, we discuss in great details the application of scalar and matrix interpolation and transcendental methods in system design. In particular the transcendental problem arises in the partial pole assignment of a multiinput multioutput plant by a stable, minimum phase compensator. The latter problem seems to arise in the simultaneous stabilization problems and also in the nonswitching compensator problem [15]. The interpolation problem on the other hand is well known in control theory and has already been introduced in [26], [27], [34] and [35].

Among the results that we have obtained in this paper, the most significant result is that if $r\min(m,p)\leqq m+p$, the simultaneous partial pole assignment problem may be analyzed via interpolation methods and one obtains a semialgebraic parametrization of the partially pole assignable r-tuples of plants. On the other hand if $r\min(m,p)>m+p$ (as would be the case if $m=p$, $r\geqq 3$), the simultaneous partial pole assignment problem is to be analyzed via transcendental methods of the type introduced in this paper. The proposed transcendental approach, we hope, would become a new simultaneous system design methodology and may be generalized to include sensitivity minimization and other system design criterion as well.

Acknowledgments. It is a pleasure to acknowledge the reviewers' comments on an earlier draft of this paper. Encouragement of Profs. C. I. Byrnes and M. Vidyasagar is gratefully acknowledged.

REFERENCES

[1] B. D. O. ANDERSON, N. K. BOSE AND E. I. JURY, *Output feedback stabilization and related problems—solution via decision methods*, IEEE Trans. Automat. Control, AC-20 (1975), pp. 53-66.

[2] J. D. BIRDWELL, D. CASTANON AND M. ATHANS, *On reliable control system designs with and without feedback reconfigurations*, Proc. 17th CDC, San Diego, CA, 1979, pp. 419-426.

[3] R. W. BROCKETT, *Some geometric questions in the theory of linear systems*, IEEE Trans. Automat. Control, AC-21 (1976), pp. 449-464.

[4] G. BRUMFIEL, *Partially ordered rings and semialgebraic geometry*, Lecture Note Series of the London Math. Soc., Cambridge Univ. Press, Cambridge, 1979.

[5] F. M. BRASCH AND J. B. PEARSON, *Pole placement using dynamic compensator*, IEEE Trans. Automat. Control, AC-15 (1970), pp. 34-43.

[6] C. I. BYRNES AND B. D. O. ANDERSON, *Output feedback and generic stabilizability*, this Journal, 22 (1984), pp. 362-380.

[7] C. I. BYRNES AND N. E. HURT, *On the moduli of linear dynamical systems*, Adv. in Math., Suppl. Series, Vol. 4, (1978), pp. 83-122; also in Modern Mathematical Systems Theory, MIR Press, Moscow, 1978 (in Russian).

[8] J. M. C. CLARK, *The consistent selection of local coordinates in linear system identification*, Proc. JACC, Purdue Univ., W. Lafayette, IN, 1976, pp. 576-580.

[9] C. A. DESOER, R. W. LIU, J. J. MURRAY AND R. SAEKS, *Feedback system design: the fractional representation approach to analysis and synthesis*, IEEE Trans. Automat. Control, AC-25 (1980), pp. 401-412.

[10] M. J. FISCHER AND M. O. RABIN, *Super-exponential complexity of Presburger arithmetic*, MIT, MAC Tech. Memo. 43, February 1974; also in Complexity of Computation, Proc. Symposium, New York, 1973, SIAM-AMS Proceedings, Vol. VII, American Mathematical Society, Providence, RI, 1974, pp. 27-41.

[11] B. K. GHOSH, *A robust reliable stabilization scheme for single input single output systems using transcendental methods*, Systems Control Lett., 5 (1984), pp. 111-115.

[12] ———, *Simultaneous stabilization and pole placement of a multimode linear dynamical system*, Ph.D. dissertation, Harvard Univ., Cambridge, MA, 1983.

[13] ———, *Simultaneous stabilization and its connection with the problem of interpolation by rational functions*, NASA Contractor Report 166441, contract NSG-2265, Feb. 1983.

[14] ———, *Simultaneous partial placement—a new approach to multimode system design*, IEEE Trans. Automat. Control (May, 1986).

[15] ———, *Some new results on the simultaneous stabilizability of a family of single input single output systems*, Systems Control Lett., 6 (1985), pp. 39-45.

[16] ———, *A geometric approach to simultaneous system design: robust stabilization as a parameterization problem*, IMA J. Math. Control Inform., special issue on parameterization problems, D. Hinrichsen and J. Willems, eds., 1986.

[17] ———, *Simultaneous stabilization, sensitivity minimization, tracking and disturbance rejection by interpolation methods*, 22nd Annual Allerton Conference on Communication, Control and Computing, 1984, pp. 625-634.

[18] B. K. GHOSH AND C. I. BYRNES, *Simultaneous stabilization and simultaneous pole-placement by non-switching dynamic compensation*, IEEE Trans. Automat. Control, AC-28 (1983), pp. 735-741.

[19] M. HAZEWINKEL AND R. E. KALMAN, *On invariants, canonical forms and moduli for linear constant finite-dimensional dynamical systems*, in Lecture Notes in Economics and Mathematics 131, System Theory, Springer-Verlag, Berlin, 1976, pp. 48-60.

[20] J. W. HELTON, *Non-Euclidean functional analysis and electronics*, Bulletin (new series) Amer. Math. Soc., 7 (1982), pp. 1-64.

[21] N. JACOBSON, *Lectures in Abstract Algebra*, Vol. I, Van Nostrand, New York, 1953.

[22] H. KIMURA, *Robust stabilizability for a class of transfer function*, 22nd IEEE Conference on Decision and Control, 1983.

[23] J. W. MILNOR, *Topology from the Differential Viewpoint*, University Press of Virginia, Charlottesville, VA, 1965.

[24] R. NEVANLINNA, *Uber bescharnkte Funktionen, die ingegebenen Punkten Vorgeschriebene werte Annehmen*, Ann. Acad. Sci., Fenn., 13, No. 1 (1919).

[25] G. PICK, *Uber die beschrankungen analytischer Funktionen, welche durch Vorgegebenen Funktionswerte bewiskt Sind*, Math. Ann., 77 (1916), pp. 7-23.

[26] R. SAEKS AND J. J. MURRAY, *Fractional representation, algebraic geometry and the simultaneous stabilization problem*, IEEE Trans. Automat. Control, AC-27, (1982), pp. 895-903.

[27] R. SAEKS, J. J. MURRAY, O. CHUA AND C. KARMOKOLIAS, *Feedback system design, the single variate case*, unpublished notes, Texas Tech. Univ., Lubbock, TX, 1980.

[28] A. SEIDENBERG, *A new decision method for elementary algebra*, Ann. Math., 60 (1954), pp. 365-374.

[29] P. STEVENS, *Algebro-geometric methods for linear multivariable feedback systems*, Ph.D. dissertation, Harvard Univ., Cambridge, MA, 1982.

[30] A. TANNENBAUM, *Invariance and System Theory: Algebraic and Geometric Aspects*, Springer-Verlag, Berlin, 1981.

[31] A. TARSKI, *A decision method for elementary algebra and geometry*, Berkeley Note, 1951.

[32] M. VIDYASAGAR AND K. R. DAVIDSON, *A characterization of all stable stabilizing compensators for single input single output systems* (to be published).

[33] M. VIDYASAGAR, H. SCHNEIDER AND B. FRANCIS, *Algebraic and topological aspects of feedback stabilization*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 880-894.

[34] M. VIDYASAGAR AND N. VISWANADHAM, *Algebraic design techniques for reliable stabilization*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 1085-1095.

[35] D. YOULA, J. BONGIORNO AND C. LU, *Single loop feedback stabilization of linear multivariable dynamic plant*, Automatica, 10 (1974), pp. 155-173.

[36] D. YOULA AND M. SAITO, *Interpolation with positive real functions*, J. Franklin Institute, 284 (1967), pp. 77-108.

[37] G. ZAMES AND B. A. FRANCIS, *Feedback minimax sensitivity, and optimal robustness*, IEEE Trans. Automat. Control, AC-28 (1983), pp. 585-601.

[38] O. ZARISKI AND P. SAMUEL, *Commutative Algebra*, Vol. I, Springer-Verlag, Berlin, 1979.

[39] E. ZEHEB AND A. LEMPEL, *Interpolation in the network*, IEEE Trans. Circuit Theory, CT-13 (1966), pp. 118-119.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 6, November 1986

OPTIMAL CONTROL WITH STATE-SPACE CONSTRAINT II*

HALİL METE SONER†

Abstract. Optimal control of piecewise deterministic processes with state space constraint is studied. Under appropriate assumptions, it is shown that the optimal value function is the only viscosity solution on the open domain which is also a supersolution on the closed domain. Finally, the uniform continuity of the value function is obtained under a condition on the deterministic drift.

Key words. viscosity solutions, stochastic control, state-space constraint, piecewise deterministic processes

AMS(MOS) subject classifications. 93E20, 35J65, 35K60, 60J60

Introduction. We are interested in the optimal control of jump processes with a state-space constraint. By that we mean the trajectories of the controlled process have to stay within a given subset θ of $\mathbb{R}^n$. These kinds of processes arise naturally in some applications [5], [9], [10]. The deterministic counterpart of this problem is studied in [11] and the optimal value function is characterized as the viscosity solution of the corresponding Hamilton-Jacobi-Bellman (HJB) equation. Also the concept of viscosity solutions, introduced by M. G. Crandall and P.-L. Lions [2], was used to identify the boundary conditions satisfied by the optimal value function. For more information about viscosity solutions see [1], [3], [7], [8] and references therein.

In this paper we generalize the results mentioned above to a certain class of jump processes, namely piecewise deterministic processes. These kinds of processes are introduced by M. Davis [4] and used by D. Vermes in [12]. Let us summarize the construction of the piecewise deterministic processes.

Let u be a Borel measurable map of $\bar{\theta} = [0, \infty)$ into a compact, separable metric space U and $y_0(x, s; t, u)$ be the solution of

$$\frac{d}{dt} y_0(x, s, t, u) = b(y_0(x, s, t, u), u(x, t-s)) \quad \text{for } t \geqq s \tag{0.1}$$

with initial data $y(x, s, s, u) = x$. Pick the first jump time T_1 so that the jump rate is $\lambda(y_0(x, 0, t, u))$. Then construct the post-jump location Y_1 such that $Q(y_0(x, 0, \tau, u), u(x, \tau), \cdot)$ is its conditional distribution given $T_1 = \tau$. Starting from Y_1 at time T_1 select the inter-jump time $T_2 - T_1$ and the second post-jump location Y_2 similarly. Set $T_0 = 0$, $Y_0 = x$ and iterate the procedure above to obtain $\{(T_n, Y_n): n = 0, 1, \cdots\}$. Between the jumps T_n and T_{n+1} the process $y(x, t, u)$ follows the deterministic trajectory passing through (Y_n, T_n), i.e.

$$y(x, t, u) = y_0(Y_n, T_n, t, u) \quad \text{if } t \in [T_n, T_{n+1}). \tag{0.2}$$

Moreover, $\{(Y_n, T_n)\}$ satisfies

$$\begin{aligned} &P(T_{n+1} - T_n \geqq t \mid T_1, Y_1, \cdots, T_n, Y_n) \\ &\qquad = \exp\left\{-\int_{T_n}^{t+T_n} \lambda(y(x, s, u), u(Y_n, s - T_n))\, ds\right\}, \end{aligned} \tag{0.3}$$

* Received by the editors March 11, 1985, and in revised form August 6, 1985. This research was supported by the National Science Foundation under grant MCS 8121940.

† Lefschetz Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912.

$$P(Y_{n+1} \in A | T_1, Y_1, \cdots, Y_n, T_{n+1}) = Q(y_0(Y_n, T_n, T_{n+1}, u), u(Y_n, T_{n+1} - T_n), A) \quad \text{for all } A \subset \bar{\theta}. \tag{0.4}$$

The process $y(x, \cdot, u)$ is a strong Markov process and the following version of Ito's lemma is proved in [4]. Set $y(t) = y(x, t, u)$ and $u_n(t) = u(Y_n, t - T_n)$. Then for any $\psi \in C^1(\bar{\theta} \times [0, T])$ we have

$$\begin{aligned} E\psi(y(T), T) &= \psi(x, 0) + E\left\{ \sum_{n=0}^{\infty} \int_{T_n \wedge T}^{T_{n+1} \wedge T} \left[b(y(t), u_n(t)) \nabla \psi(y(t), t) \right.\right. \\ &\qquad + \frac{\partial}{\partial t} \psi(y(t), t) + \lambda(y(t), u_n(t)) \\ &\qquad \left.\left. \cdot \int_{\bar{\theta}} [\psi(z, t) - \psi(y(t), t)] Q(y(t), u_n(t), dz) \right] dt \right\}. \end{aligned} \tag{0.5}$$

We assume that the post-jump locations are in $\bar{\theta}$. Then one can define the set of admissible strategies $\mathcal{A}_{ad}$ as:

$$\mathcal{A}_{ad} := \left\{ \begin{array}{l} u: \bar{\theta} \times [0, \infty) \to U, \text{ Borel measurable and} \\ P(y(x, t, u) \in \bar{\theta} \text{ for all } t \geqq 0) = 1, \text{ for all } x \in \bar{\theta} \end{array} \right\}. \tag{0.6}$$

The optimal value is given by

$$v(x) := \inf_{u \in \mathcal{A}_{ad}} E\left\{ \sum_{n=0}^{\infty} \int_{T_n}^{T_{n+1}} e^{-t} f(y(x, t, u), u(Y_n, t - T_n))\, dt \right\}. \tag{0.7}$$

It is shown, in § 2, that v is the only viscosity solution of the corresponding HJB equation, satisfying the same boundary condition as in the deterministic case [11]. This result holds if the optimal value is in $BUC(\bar{\theta})$ and the dynamic programming relation (0.8) is satisfied.

$$v(x) = \inf_{u \in \mathcal{A}_{ad}} E\left\{ \int_0^{T \wedge T_1} e^{-t} f(y(x, t, u), u(x, t))\, dt + e^{-T \wedge T_1} v(y(x, T \wedge T_1, u)) \right\} \quad \text{for all } T \geqq 0 \text{ and } x \in \bar{\theta}. \tag{0.8}$$

Finally, in § 3 we show that under assumptions (A2)-(A4) v is in $BUC(\bar{\theta})$ and satisfies the dynamic programming relation (0.8). Note that these assumptions yield that the optimal value of the corresponding deterministic problem is in $BUC(\bar{\theta})$. By an induction argument one can extend this result to piecewise deterministic processes with finitely many jumps. We eventually pass to the limit to conclude.

1. Main result. Let θ be an open subset of $\mathbb{R}^n$ with connected boundary satisfying:

(A.1) There are positive constants h, r and $\mathbb{R}^n$-valued bounded-uniformly continuous map η of $\bar{\theta}$ such that

$$B(x + t\eta(x), tr) \subset \bar{\theta} \quad \text{for all } x \in \bar{\theta} \text{ and } t \in (0, h].$$

Here $B(x, R)$ denotes the ball with center x and radius R.

Remark 1.1. If θ is bounded and $\partial\theta$ is C^1, then (A.1) is satisfied. Also boundaries with corners may satisfy (A.1), for example, $\theta = \{(x, y) \in R^2 : x > 0, y > 0\}$.

The strategies take values in U which is a compact, separable metric space. Also, we assume the following throughout the paper. Let x and y be in $\bar{\theta}$.

$$\sup_{\alpha\in U}|\gamma(x,\alpha)-\gamma(y,\alpha)|\leqq L(\gamma)|x-y|, \qquad \gamma=b,f \text{ or } \lambda, \tag{1.1}$$

$$\sup_{\substack{\alpha\in U\\ x\in\bar{\theta}}}|\gamma(x,\alpha)|\leqq K(\gamma), \qquad \gamma=b,f \text{ or } \lambda. \tag{1.2}$$

For each bounded, continuous function h on $\bar{\theta}$, there is a continuous function W_h with $W_h(0)=0$ such that

$$\sup_{\alpha\in U}\left|\int_{\bar{\theta}}h(z)Q(x,\alpha,dz)-\int_{\bar{\theta}}h(z)Q(y,\alpha,dz)\right|\leqq W_h(|x-y|). \tag{1.3}$$

$$Q(x,\alpha,\bar{\theta})=1 \quad \text{for all } x\in\bar{\theta} \text{ and } \alpha\in U. \tag{1.4}$$

$$\lambda(x,\alpha)\geqq 0 \quad \text{for all } x\in\bar{\theta} \text{ and } \alpha\in U. \tag{1.5}$$

The corresponding Hamiltonian H is a continuous map of $\bar{\theta}\times\mathbb{R}^n\times BUC(\bar{\theta})$ given as:

$$H(x,p,\psi)=\sup_{\alpha\in U}\left\{-b(x,\alpha)\cdot p-f(x,\alpha)-\lambda(x,\alpha)\int_{\bar{\theta}}[\psi(z)-\psi(x)]Q(x,\alpha,dy)\right\}. \tag{1.6}$$

This Hamiltonian is a nonlocal operator but still one can define a notion of viscosity solutions.

DEFINITION. Let K be a subset of $\mathbb{R}^n$ and $v\in BUC(\bar{K})$.

(i) We say v is a *viscosity subsolution* of $v(x)+H(x,Dv(x),v)=0$ on K if $v(x_0)+H(x_0,\nabla\psi(x_0),v)\leqq 0$ whenever $\psi\in C^1(N_{x_0})$ and $(v-\psi)$ has a global maximum, relative to K, at $x_0\in K$, where N_{x_0} is a neighborhood of x_0.

(ii) We say v is a *viscosity supersolution* of $v(x)+H(x,Dv(x),v)=0$ on K if $v(x_0)+H(x_0,\nabla\psi(x_0),v)\geqq 0$ whenever $\psi\in C^1(N_{x_0})$ and $(v-\psi)$ has a global minimum, relative to K, at $x_0\in K$, where N_{x_0} is a neighborhood of x_0.

Remark 1.2. This is an obvious generalization of the original notion introduced by M. G. Crandall and P.-L. Lions [2]. The definition we used above is analogous to one of the definitions introduced in [1].

We are interested in the following notion of viscosity solutions.

DEFINITION. $v\in BUC(\bar{\theta})$ is said to be a *constrained viscosity solution* of $v(x)+H(x,Dv(x),v)=0$ on $\bar{\theta}$ if it is a subsolution on θ and supersolution on $\bar{\theta}$.

Remark 1.3. The fact that v is a supersolution on the closed domain imposes a certain boundary condition. Suppose that v is smooth and a constrained viscosity solution. Then $H(x,\nabla v(x)+\alpha v(x),v)\geqq H(x,\nabla v(x),v)$ for all $x\in\partial\theta$ and $\alpha\geqq 0$ ($v(x)$ is the exterior normal vector). This effect is discussed in [11].

THEOREM 1.1. *Suppose* (A.1), (1.1)-(1.5) *hold. Then there is at most one constrained viscosity solution of* $v(x)+H(x,Dv(x),v)=0$ *on* θ. *Moreover if* $v\in BUC(\bar{\theta})$ *and dynamic programming relation* (0.8) *holds, then the optimal value function* v *is a constrained viscosity solution.*

2. Proof of the main theorem. We need the following lemma:

LEMMA 2.1. $v\in BUC(\bar{\theta})$ *is a viscosity subsolution of* $v(x)+H(x,Dv(x),v)=0$ *on* θ (*or supersolution on* $\bar{\theta}$) *if and only if*

$$v(x_0)+H(x_0,\nabla\psi(x_0),\psi)\leqq 0$$

(*or* $\geqq 0$) *whenever* $\psi\in C^1(N_{x_0})$ *and* $v-\psi$ *has a global maximum relative to* $\bar{\theta}$ *at* $x_0\in\theta$ (*or minimum at* $x_0\in\bar{\theta}$ *respectively*), *where* N_{x_0} *is a neighborhood of* x_0.

Proof. We will prove the statement for subsolutions only, the other statement is proved exactly the same way.

Necessity. Suppose $v \in BUC(\bar{\theta})$ is a viscosity subsolution and ψ, x_0 are as above, i.e.

$$v(x_0) - \psi(x_0) = \max_{x \in \theta} v(x) - \psi(x).$$

Then we have $v(x_0) - v(z) \leqq \psi(x_0) - \psi(z)$ for all $z \in \bar{\theta}$. Then (1.5) yields:

$$H(x_0, \nabla\psi(x_0), \psi) \leqq H(x_0, \nabla\psi(x_0, v).$$

Hence the viscosity property of v gives the result.

Sufficiency. Let $\psi \in C^1(N_{x_0})$ and $(v-\psi)(x_0) = \max_{x \in \theta}\{(v-\psi)(x)\} = 0$. For each $\varepsilon > 0$ we define Φ^ε as follows:

$$\Phi^\varepsilon(x) = \psi(x)\chi^\varepsilon(x) + v(x)(1-\chi^\varepsilon(x)) \quad \text{for } x \in \bar{\theta} \tag{2.1}$$

where χ^ε is a smooth function satisfying

$$\begin{aligned} &0 \leqq \chi^\varepsilon \leqq 1, \\ &\chi^\varepsilon(x) = 1 \quad \text{if } x \in B(x_0, \varepsilon), \\ &\chi^\varepsilon(x) = 0 \quad \text{if } x \in R^n \setminus B(x_0, 2\varepsilon). \end{aligned} \tag{2.2}$$

Observe $v(x_0) - \Phi^\varepsilon(x_0) = 0$ and $v(x) - \Phi^\varepsilon(x) = (v(x) - \psi(x))\chi^\varepsilon(x) \leqq 0$. Hence

$$v(x_0) - \Phi^\varepsilon(x_0) = \max_{x \in \bar{\theta}} \{v(x) - \Phi^\varepsilon(x)\}. \tag{2.3}$$

Thus the hypothesis of the lemma and $\nabla\Phi^\varepsilon(x_0) = \nabla\psi(x_0)$ yields

$$v(x_0) + H(x_0, \nabla\psi(x_0), \Phi^\varepsilon) \leqq 0. \tag{2.4}$$

The following estimate follows (1.5):

$$\begin{aligned} &|H(x_0, \nabla\psi(x_0), \Phi^\varepsilon) - H(x_0, \nabla\psi(x_0), v)| \\ &\leqq \sup_{\alpha \in U} \left\{ \lambda(x_0, \alpha) \int |\Phi^\varepsilon(x_0) - \Phi^\varepsilon(y) - v(x_0) + v(y)| Q(x_0, \alpha, dy) \right\}. \end{aligned} \tag{2.5}$$

Observe that $\Phi^\varepsilon(x_0) = v(x_0)$ and $\Phi^\varepsilon(x_0 + y) = v(x_0 + y)$ for $y \notin B(0, 2\varepsilon)$. Also for $y \in B(0, 2\varepsilon)$

$$\begin{aligned} |\Phi^\varepsilon(x_0+y) - v(x_0+y)| &= |\psi(x_0+y) - v(x_0+y)|\chi^\varepsilon(x_0+y) \\ &\leqq |\psi(x_0+y) - \psi(x_0)| + |v(x_0) - v(x_0+y)| \\ &\leqq \|\nabla\psi\|_\infty |y| + \omega_v(|y|) \\ &\leqq 2\|\nabla\psi\|_\infty \varepsilon + \omega_v(2\varepsilon). \end{aligned} \tag{2.6}$$

Here ω_v is the modulus of continuity of v and we used $\psi(x_0) = v(x_0)$ in the second inequality. Combine (2.4)–(2.6) to conclude that v is a viscosity subsolution. □

Remark 2.1. It is easy to prove that in Lemma 2.1 we may replace $\psi \in C^1(N_{x_0})$ by $\psi \in C^1(\bar{\theta})$ (see [1]).

Proof of Theorem 1.1. Suppose v_1 and v_2 are two solutions in $BUC(\bar{\theta})$. For $i = 1, 2$ define f_i and $\bar{H}_i$ as follows

$$f_i(x, \alpha) = f(x, \alpha) + \lambda(x, \alpha) \int_{\bar{\theta}} (v_i(y) - v_i(x)) Q(x, \alpha, dy) \quad \text{for } x \in \bar{\theta},\ \alpha \in U, \tag{2.7}$$

$$\bar{H}_i(x, p) = \sup_{\alpha \in U} \{-b(x, \alpha) \cdot p - f_i(x, \alpha)\} \quad \text{for } x \in \bar{\theta},\ p \in \mathbb{R}^n. \tag{2.8}$$

Notice that $\bar{H}_i$ is the Hamiltonian of the corresponding deterministic problem with running cost f_i. Using (1.1)-(1.4), one can show that the f_i's are uniformly continuous in x uniformly with respect to α. Pick $z_\delta \in \bar{\theta}$ such that

$$v_1(x) - v_2(x) \leqq v_1(z_\delta) - v_2(z_\delta) + \delta \quad \text{for all } x \in \bar{\theta}.$$

Then Corollary 2.3 in [11] yields

$$v_1(z_\delta) - v_2(z_\delta) \leqq c\delta + \omega_{f_1}(c\delta) + \omega_{f_2}(c\delta) + \sup_{\alpha \in U} \{f_1(z_\delta, \alpha) - f_2(z_\delta, \alpha)\}. \tag{2.9}$$

But $f_1(z_\delta, \alpha) - f_2(z_\delta, \alpha) \leqq \delta\lambda(z_\delta, \alpha) \leqq \delta K(\lambda)$ for every α. Substitute this into (2.9) and send δ to zero, to prove the uniqueness.

Let $\psi \in C^1(\bar{\theta})$ and $x_0 \in \theta$ such that $(v-\psi)(x_0) = \max\{(v-\psi)(x); x \in \bar{\theta}\} = 0$, where v is the optimal value. For any $u \in \mathcal{A}_{ad}$ the dynamic programming relation (0.8) yields

$$\psi(x_0) = v(x_0) \leqq E\left\{\int_0^{t \wedge T_1} e^{-s} f(y(x_0, s, u), u(x_0, s))\, ds + e^{-t \wedge T_1} v(y(x_0, t \wedge T_1, u))\right\}.$$

Set $y(s) = y(x_0, s, u)$ and $u(s) = u(x_0, s)$. Use $v \leqq \psi$ to obtain:

$$\psi(x_0) \leqq E\left\{\int_0^{t \wedge T_1} e^{-s} f(y(s), u(s))\, ds + e^{-t \wedge T_1} \psi(y(t \wedge T_1))\right\}. \tag{2.10}$$

Ito's formula (0.5) on $e^{-s\psi(y(s))}$ yields

$$E \int_0^{t \wedge T_1} e^{-s}[\psi(y(s)) - b(y(s), u(s)) \cdot \nabla\psi(y(s)) - \bar{f}(y(s), u(s))]\, ds \leqq 0 \tag{2.11}$$

where

$$\bar{f}(x, \alpha) = f(x, \alpha) + \lambda(x, \alpha) \int_{\bar{\theta}} (\psi(z) - \psi(x)) Q(x, \alpha, dz).$$

Observe that on $[0, t \wedge T_1)$ $y(\cdot)$ is a deterministic trajectory. Thus standard estimates on $y(\cdot)$ and (1.1)-(1.4) yield

$$\frac{1}{t} E \int_0^{t \wedge T_1} [\psi(x_0) - b(x_0, u(s)) \cdot \nabla\psi(x_0) - \bar{f}(x_0, u(s))]\, ds \leqq h(t) \tag{2.12}$$

where h is a continuous function with $h(0) = 0$. Since $x_0 \in \theta$, for any $\alpha \in U$ there is a strategy $u \in \mathcal{A}_{ad}$ such that $u(x_0, t) = \alpha$ for all $t \leqq \operatorname{dist}(x_0, \partial\theta)/K(b)$. Use this strategy in (2.12) to obtain

$$[\psi(x_0) - b(x_0, \alpha) \cdot \nabla\psi(x_0) - \bar{f}(x_0, \alpha)] E[(t \wedge T_1)/t] \leqq h(t).$$

Hence $\psi(x_0) + H(x_0, \nabla\psi(x_0), \psi) \leqq 0$. So Lemma 2.1 and Remark 2.1 imply that v is a subsolution on θ.

Now let $\psi \in C^1(\bar{\theta})$ and $(v-\psi)(x_0) = \min\{(v-\psi)(x): x \in \bar{\theta}\} = 0$ for some $x_0 \in \bar{\theta}$. The dynamic programming relation and $v \geqq \psi$ yield

$$v(x_0) = \psi(x_0) \geqq \inf_{u \in \mathcal{A}_{ad}} E\left[\int_0^{t \wedge T_1} e^{-s} f(y(s), u(s))\, ds + e^{-t \wedge T_1} \psi(y(t \wedge T_1))\right]. \tag{2.13}$$

For $t = 1/m$ one can pick $u_m \in \mathcal{A}_{ad}$ such that

$$\psi(x_0) + \left(\frac{1}{m}\right)^2 \geqq E\left[\int_0^{1/m \wedge T_1} e^{-s} f(y(s), u_m(s))\, ds + e^{-T_1 \wedge 1/m} \psi\left(y\left(T_1 \wedge \frac{1}{m}\right)\right)\right].$$

First, use Ito's lemma (0.5) on $e^{-s}\psi(y(s))$ then (1.1)-(1.4), as in (2.10)-(2.12) above to obtain

$$mE\int_0^{T_1\wedge 1/m}[\psi(x_0)-b(x_0,u_m(s))\cdot\nabla\psi(x_0)-\bar f(x_0,u_m(s))]\,ds\geqq K(m) \tag{2.14}$$

where $K(m)$ is converging to zero as m tends to infinity. Rewrite (2.14) as

$$[\psi(x_0)-B(m)\cdot\nabla\psi(x_0)-F(m)]E(mT_1\wedge 1)\geqq K(m) \tag{2.15}$$

where

$$B(m)=\left(E\left(T_1\wedge\frac{1}{m}\right)\right)^{-1}E\int_0^{T_1\wedge 1/m}b(x_0,u_m(s))\,ds,$$

$$F(m)=\left(E\left(T_1\wedge\frac{1}{m}\right)\right)^{-1}E\int_0^{T_1\wedge 1/m}\bar f(x_0,u_m(s))\,ds.$$

Observe that $(B(m),F(m))\in\overline{\text{co}}\{(b(x_0,\alpha),\bar f(x_0,\alpha)):\alpha\in U\}:=\overline{\text{co}}[BF(x_0)]$. Hence there is $(B,F)\in\overline{\text{co}}\,[BF(x_0)]$ such that $(B(m),F(m))$ converges to (B,F) on a subsequence, denoted by m again. Pass to the limit in (2.15) to obtain

$$\psi(x_0)+\sup\{-B\cdot\nabla\psi(x_0)-F:(B,F)\in\overline{\text{co}}\,[BF(x_0)]\}\geqq 0.$$

Also

$$\sup\{-B\cdot\nabla\psi(x_0)-F:(B,F)\in\overline{\text{co}}\,[BF(x_0)]\}=H(x_0,\nabla\psi(x_0),\psi).$$

Thus, Lemma 2.1 and Remark 2.1 imply that v is a viscosity supersolution on $\bar\theta$. □

3. Uniform continuity and dynamic programming. We assume the following.

(A.2) There is a Borel measurable map α of $\partial\theta$ into U and β positive satisfying $b(x,\alpha(x))\cdot\nu(x)\leqq-\beta<0$, where ν is the exterior normal vector.

(A.3) The boundary of θ is of class C^2.

(A.4) If $\partial\theta$ is not compact, there are constants ρ and l such that for any $x\in\partial\theta$ there is a $C^2(B(x,\rho))$ function T with $C^1(B(x,\rho))$ inverse T^{-1} satisfying

$$\begin{aligned}&\text{(i)}\quad T(B(x,\rho)\cap\theta)\subset\{y\in R^n:y_n>0\},\\ &\text{(ii)}\quad T(B(x,\rho)\cap\partial\theta)\subset\{y\in R^n:y_n=0\},\\ &\text{(iii)}\quad \|T\|_{C^2(B(x,\rho))}+\|T^{-1}\|_{C^1(B(x,\rho))}\leqq l.\end{aligned} \tag{3.1}$$

The subscript n denotes the nth component.

Remark 3.1. The assumption (A.2) holds with some $\beta>0$ if

$$\sup_{x\in\partial\theta}\min_{\alpha\in U}b(x,\alpha)\cdot\nu(x)<0.$$

Note that we do not assume that $b(x,\alpha)$ points inwards the domain θ for all α and $x\in\partial\theta$. Thus there may be controls that allow the deterministic process to reach the boundary.

Remark 3.2. The assumptions (A.2)-(A.4) are used to obtain the uniform continuity of the corresponding deterministic problem [11]. In particular see [11, Lemma 3.2].

Let v_0 be the optimal value of the deterministic problem and define v^N as follows

$$v^N(x)=\inf_{u\in\mathcal{A}_{ad}}J^N(x,u)$$

where

$$J^N(x,u)=E\left[\int_0^{T_1}e^{-t}f(y(x,t,u),u(x,t))\,dt+e^{-T_1}v^{N-1}(Y_1)\right]. \tag{3.2}$$

LEMMA 3.1. *Let $v^{N-1} \in BUC(\bar{\theta})$, then the dynamic programming relation holds for v^N, i.e., for all $T>0$*

$$
\begin{aligned}
v^N(x) = \inf_{u \in \mathcal{A}_{ad}} E\Bigg\{ & \int_0^{T \wedge T_1} e^{-T} f(y(x, t, u), u(x, t))\, dt \\
& + e^{-T_1} v^{N-1}(Y_1) \chi_{[0,T]}(T_1) + e^{-T} v^N(y(x, T, u)) \chi_{(T,\infty)}(T_1) \Bigg\}
\end{aligned} \tag{3.3}
$$

where χ_A is the indicator function of set A.

Proof. Fix $x \in \bar{\theta}$ and T positive. Let $I^N(x, u)$ be the right-hand side of (3.3) before taking the infimum. To simplify the notation, put $y(t) = y_0(x, 0; t, u)$, $u(t) = u(x, t)$, $\lambda(t) = \lambda(y(t), u(t))$ and $\Lambda(t) = \exp\{-\int_0^t \lambda(s)\, ds\}$. Recall that y_0 is the corresponding deterministic trajectory given by (0.1). In terms of these quantities $I^N(x, u)$ is given by

$$
\begin{aligned}
I^N(x, u) = & \int_0^\infty \lambda(t)\Lambda(t) \Bigg[\int_0^{T \wedge t} e^{-s} f(y(s), u(s))\, ds \\
& + e^{-t} \int_{\bar{\theta}} v^{N-1}(z) Q(y(t), u(t), dz) \chi_{[0,T]}(t) \\
& + e^{-T} v^N(y(T)) \chi_{(T,\infty)}(t) \Bigg] dt \\
& + \Lambda(\infty) \Bigg[\int_0^T e^{-t} f(y(t), u(t))\, dt + e^{-T} v^N(y(T)) \Bigg] \\
= & \int_0^T \lambda(t)\Lambda(t) \Bigg[\int_0^t e^{-s} f(y(s), u(s))\, ds + e^{-t} \int_{\bar{\theta}} v^{N-1}(z) Q(y(t), u(t), dz) \Bigg] dt \\
& + \Lambda(T) \int_0^T e^{-s} f(y(s), u(s))\, ds + \Lambda(T)\, e^{-T} v^N(y(T)).
\end{aligned} \tag{3.4}
$$

Since $y(T)$ is a deterministic quantity determined by x, T and u, one can pick $u^* \in \mathcal{A}_{ad}$ such that $v^N(y(T)) \geqq J^N(y(T), u^*) - \delta$. Now we define $\bar{u}$ as follows:

$$
\bar{u}(z, t) = u(z, t) \chi_{[0,T]}(t) + u^*(y_0(z, 0, T, u), t - T) \chi_{[T,\infty)}(t). \tag{3.5}
$$

Define $\bar{y}(t) = \bar{y}(x, 0; t, \bar{u})$ and $\bar{\lambda}$, $\bar{\Lambda}$ similarly. Then we have

$$
\begin{aligned}
J^N(y(T), u^*) = (\Lambda(T))^{-1} \Bigg\{ & \int_0^\infty \bar{\lambda}(t+T)\bar{\Lambda}(t+T) \\
& \cdot \Bigg[\int_0^t e^{-s} f(\bar{y}(s+T), \bar{u}(x, s+T))\, ds \\
& + e^{-t} \int_{\bar{\theta}} v^{N-1}(z) Q(\bar{y}(t+T), \bar{u}(x, t+T), dz) \Bigg] dt \\
& + \bar{\Lambda}(\infty) \int_0^\infty e^{-s} f(\bar{y}(s+T), \bar{u}(x, s+T))\, ds \Bigg\}.
\end{aligned} \tag{3.6}
$$

Change variables in (3.6) and use $v^N(y(T)) \geqq J^N(y(T), u^*) - \delta$ to obtain

$$e^{-T}\Lambda(T)v^N(y(T)) \geqq \int_T^\infty \bar{\lambda}(t)\bar{\Lambda}(t)\Big[\int_T^t e^{-s}f(\bar{y}(s), \bar{u}(x, s))\,dx + e^{-t}\int_{\bar{\theta}} v^{N-1}(z)Q(\bar{y}(t), \bar{u}(x, t), dz)\Big]\,dt + \bar{\Lambda}(\infty)\int_T^\infty e^{-s}f(\bar{y}(s), \bar{u}(x, s))\,ds - e^{-T}\Lambda(T)\delta.$$

The fact that $\int_T^\infty \bar{\lambda}(t)\bar{\Lambda}(t)\,dt = \Lambda(T) - \bar{\Lambda}(\infty)$ and the above inequality yield

$$\Lambda(T)\int_0^T e^{-s}f(y(s), u(s))\,ds + \Lambda(T)\,e^{-T}v^N(y(T)) \geqq \int_T^\infty \bar{\lambda}(t)\bar{\Lambda}(t)\Big[\int_0^t e^{-s}f(\bar{y}(s), \bar{u}(x, s))\,ds + e^{-t}\int_{\bar{\theta}} v^{N-1}(z)Q(\bar{y}(t), \bar{u}(x, t), dz)\Big]\,dt + \bar{\Lambda}(\infty)\int_0^\infty e^{-s}f(\bar{y}(s), \bar{u}(x, s))\,ds - \delta.$$

Substitute the above inequality into (3.4) and use the fact $\lambda(t) = \bar{\lambda}(t)$, $\Lambda(t) = \bar{\Lambda}(t)$ for $t \in [0, T]$ to obtain

$$\begin{aligned} I^N(x, u) &\geqq \int_0^\infty \bar{\lambda}(t)\bar{\Lambda}(t)\Big[\int_0^t e^{-s}f(\bar{y}(s), \bar{u}(x, s))\,ds + e^{-t}\int_{\bar{\theta}} v^{N-1}(z)Q(\bar{y}(t), \bar{u}(x, t), dz)\Big]\,dt \\ &\quad + \bar{\Lambda}(\infty)\int_0^\infty e^{-s}f(\bar{y}(s), \bar{u}(x, s))\,ds - \delta \\ &= J^N(x, \bar{u}) - \delta \geqq v^N(x) - \delta. \end{aligned} \tag{3.7}$$

Thus, $v^N(x) \leqq \inf_{u \in \mathscr{A}_{ad}} I^N(x, u)$. One can prove the other inequality similarly. □

The following lemma is an analogue of Lemma 3.2 in [11].

LEMMA 3.2. *Let $v^{N-1} \in BUC(\bar{\theta})$ and* (1.1)-(1.4), (A.2)-(A.4) *hold. Then for any T positive, there is a positive function h_T and a projection $\mathscr{P}_T(u)$ of any Borel measurable map u of $\bar{\theta} \times [0, \infty)$ into U such that $\mathscr{P}_T(u) \in \mathscr{A}_{ad}$ and*

$$|J_T^N(x, u) - J_T^N(x, \mathscr{P}_T(u))| \leqq h_T(\sup\{\operatorname{dist}(y_0(x, 0, t, u), \bar{\theta}): t \in [0, T]\}) \tag{3.8}$$

where h_T is a continuous function with $h_T(0) = 0$ and

$$J_T^N(x, u) = E\Big\{\int_0^{T \wedge T_1} e^{-s}f(y(x, s, u), u(x, s))\,ds + e^{-T_1}v^{N-1}(Y_1)\chi_{[0,T]}(T_1)\Big\}. \tag{3.9}$$

Proof. Define $t_0(x, u) = \inf\{t \geqq 0: y_0(x, 0, t, u) \in \partial\theta\}$ or infinity. Let t^* and k be as in Lemma 3.2 of [11], i.e.

$$\begin{aligned} t^* &= \min\{\rho/K(\bar{b}), l\beta/L(\bar{b})K(\bar{b}), \ln(1 + \beta l/4K(\bar{b}))/L(\bar{b})\}, \\ k &= 2/l\rho \end{aligned} \tag{3.10}$$

where $K(\bar{b}) = lK(b)$ and $L(\bar{b}) = l^2K(b) + l^2L(b)$. We now construct u^1 as in Lemma 3.2 of [11]

$$u^1(x, t) = \begin{cases} u(x, t) & \text{if } t \leqq t_0(x, u) \quad \text{or } t \geqq t_0(x, u) + k\varepsilon_0(x, u), \\ \alpha(y_0(x, 0, t_0(x, u), u)) & \text{if } t_0(x, u) < t < t_0(x, u) + k\varepsilon_0(x, u) \end{cases} \tag{3.11}$$

where α as in (A.2) and $\varepsilon_0(x, u) = \sup\{\text{dist}(y_0(x, 0, t, u), \bar{\theta}): t \in [0, t^*]\}$. Since $x \to t_0(x, u)$ is a Borel measurable map u^1 is measurable. Also in Lemma 3.2 of [11] it is proved that

$$y_0(x, 0, t, u^1) \in \bar{\theta} \quad \text{for all } t \in [0, t^*]. \tag{3.12}$$

Now construct a sequence of strategies $\{u^n: n = 1, 2, \cdots\}$ by the following recursive formula

$$t_n(x, u^n) = \inf\{t \geqq nt^*: y_0(x, 0, t, u^n) \in \partial\theta\} \quad \text{or infinity},$$

$$\varepsilon_n(x, u^n) = \sup_{t \in [0,(n+1)t^*]} [\text{dist}(y_0(x, 0, t, u^n), \bar{\theta})],$$

then

$$u^{n+1}(x, t) = \begin{cases} u^n(x, t) & \text{if } t \leqq t_n(x, u^n) \quad \text{or } t \geqq t_n(x, u^n) + k\varepsilon_n(x, u^n), \\ \alpha(y_0(x, 0, t_n(x, u^n), u^n)) & \text{if } t_n(x, u^n) < t < t_n(x, u^n) + k\varepsilon_n(x, u^n). \end{cases}$$

Iterate (3.12) to get $y_0(x, 0, t, u^n) \in \bar{\theta}$ for $t \in [0, nt^*]$. We have to estimate the Lebesgue measure of the following set

$$M^n(x) = \{t \in [0, nt^*]: u^n(x, t) \neq u(x, t)\}. \tag{3.13}$$

For every t we have

$$|y_0(x, 0, t, u^n) - y_0(x, 0, t, u)|$$

$$\leqq \int_{[0,t] \cap M^n(x)} 2K(b) + \int_{[0,t] \setminus M^n(x)} L(b)|y_0(x, 0, s, u) - y_0(x, 0, s, u^n)|\, ds;$$

thus the Gronwall's inequality implies

$$\varepsilon_n(x, u^n) \leqq \varepsilon_n(x, u) + 2K(b)\, e^{(n+1)L(b)t^*} \text{ meas } M^n(x).$$

The construction of u^n yields

$$\begin{aligned} \text{meas } M^n(x) &\leqq k \sum_{i=0}^{n-1} \varepsilon_i(x, u^i) \\ &\leqq nk\varepsilon_{n-1}(x, u) + k2K(b)\, e^{nL(b)t^*} \sum_{i=0}^{n-1} \text{meas } M^i(x). \end{aligned} \tag{3.14}$$

Iterate this inequality to obtain $C(n)$, depending only on n, $K(b)$ and $L(b)$, such that

$$\text{meas } M^n(x) \leqq C(n)\varepsilon_{n-1}(x, u). \tag{3.15}$$

Now, for given T, choose $\bar{n}$ so that $T \leqq \bar{n}t^*$. Then define $\mathscr{P}_T u$ as

$$\mathscr{P}_T u(x, t) = \begin{cases} u^{\bar{n}}(x, t) & \text{for } t \leqq T, \\ \tilde{u}(y_0(x, 0, T, u^{\bar{n}}), t - T) & \text{for } t > T \end{cases}$$

where $\tilde{u}$ is any strategy in $\mathscr{A}_{ad}$.

Observe that for any u

$$J_T^N(x, u) = \int_0^T \lambda(t)\Lambda(t)\left[\int_0^t e^{-s} f(y(s), u(s))\, ds + e^{-t} \int_{\bar{\theta}} v^{N-1}(z) Q(y(t), u(t), dz)\right] dt$$

$$+ \Lambda(T) \int_0^T e^{-s} f(y(s), u(s))\, ds.$$

By using (1.1)-(1.4) one can show

$$|J_T^N(x,u)-J_T^N(x,\mathscr{P}_T(u))|\leqq CT[d_T(x)+W_{v^{N-1}}(d_T(x))+\text{meas } M^{\bar\theta}(x)] \tag{3.16}$$

where C is a positive constant and

$$\begin{aligned} d_T(x) &= \sup_{t\in[0,T]}\{|y_0(x,0,t,u)-y_0(x,0,t,\mathscr{P}_T(u))|\} \\ &\leqq 2K(b)\, e^{L(b)T}\,\text{meas } M^{\bar\theta}(x). \end{aligned}$$

Plug this into (3.16) together with (3.15) to conclude Lemma 3.2. □

LEMMA 3.3. *Let* (A.2)-(A.4) *and* (1.1)-(1.5) *hold*; *then* $v^N\in BUC(\bar\theta)$ *and there is a* δ*-optimal strategy* u^* *such that*

$$J^N(x,u^*)\leqq v^N(x)+\delta \quad \textit{for all } x\in\bar\theta.$$

Proof. It is proved that $v^0\in BUC(\bar\theta)$ [11, Thm. 3.3]. Now suppose $v^{N-1}\in BUC(\bar\theta)$ and define

$$\omega(r)=\sup\{|v^N(x)-v^N(y)|: x,y\in\bar\theta \text{ and } |x-y|<r\} \quad \text{for } r>0.$$

At the origin $\omega(0)=\lim_{r\downarrow 0}\omega(r)$. Using Lemmas 3.1 and 3.2, we can conclude, as in Theorem 3.3 of [11], that for some t positive,

$$\omega(r)\leqq h_t(Cr)+e^{-t}\omega(\tilde C r)+Ctr \tag{3.17}$$

where $\tilde C>1$, $C>0$ and h_t is as in Lemma 3.2. Iterate (3.17) to obtain

$$\omega(\tilde C^{-n})\leqq e^{-nt}\omega(1)+\sum_{m=0}^{n-1} e^{-mt}[h_t(C\tilde C^{(m-n)})+C\tilde C^{(m-n)}t].$$

Use dominated convergence theorem to get $\lim_{n\to\infty}\omega(\tilde C^{-n})=0$. Hence $v^N\in BUC(\bar\theta)$.

Pick $\{x_m: n=1,2,\cdots\}\subset\bar\theta$ such that $\bar\theta\subset\bigcup_m B(x_m,r)$ where r to be chosen. Then select $\{u_n: n=1,2,\cdots\}\subset\mathscr{A}_{ad}$

$$J^N(x_m,u_m)\leqq v^N(x_m)+\frac{\delta}{2}. \tag{3.18}$$

Now define u as follows

$$u(x,t)=u_m(x_m,t) \quad \text{if } x\in\theta_m=\bigcup_{n=1}^{m} B(x_n,r)\Big\backslash\bigcup_{n=1}^{m-1} B(x_n,r).$$

Let $u^*=\mathscr{P}_T(u)$ where T to be chosen. Note that u^* depends both on r and T but this dependence is suppressed in the notation. For every $x\in\theta_m$ we have

$$\begin{aligned} |J^N(x,u^*)-J^N(x_m,u)| &\leqq |J_T^N(x,u^*)-J_T^N(x,u)|+|J_T^N(x,u)-J_T^N(x_m,u)| \\ &\quad +2\sup_{\substack{x\in\theta\\ u\in\mathscr{A}_{ad}}}|J^N(x,u)-J_T^N(x,u)|:=I_1+I_2+I_3. \end{aligned} \tag{3.19}$$

The construction of u and standard ODE estimates yield

$$\sup_{\substack{x\in\theta_m\\ t\in[0,T]}}|y_0(x,0;t,u)-y_0(x_m,0;t,u)|\leqq C(T)r \tag{3.20}$$

where $C(T)$ is a positive constant. Since $y_0(x_m,0;t,u)\in\bar\theta$ for all $t\geqq 0$ above inequality implies that

$$\sup_{\substack{x\in\bar\theta\\ t\in[0,T]}} d(y_0(x,0;t,u),\bar\theta)\leqq C(T)r. \tag{3.21}$$

Now, in the case of I_1 use (3.8), (3.21) and in the case of I_2 use (3.20), (1.1)-(1.4) and the continuity of v^{N-1} to obtain

$$I_1 + I_2 \leqq h^N(T, r) \tag{3.22}$$

where $h^N \in C([0,\infty)\times[0,\infty))$ with $\lim_{r\downarrow 0} h^N(T, r) = 0$ for every T and N. Also (3.2) and (3.9) imply

$$|J^N(x,u) - J_T^N(x,u)| = \left| E\left[e^{-T_1} v^{N-1}(Y_1)\chi_{(T,\infty)}(T_1) + \int_{T\wedge T_1}^{T_1} e^{-s} f(y(x,s,u), u(x,s))\, ds \right] \right|$$
$$\leqq e^{-T}[\|v^{N-1}\|_{L^\infty(\bar\theta)} + K(f)].$$

Recall that $K(f)$ is the sup-norm of f and it is easy to show that v^N is bounded by $K(f)$ for every N. Hence we have

$$I_3 \leqq 4K(f)\, e^{-T}. \tag{3.23}$$

Substitute (3.22)-(3.23) into (3.19) to get for all $x_m \in \theta_m$

$$|J^N(x,u^*) - J^N(x_m,u)| \leqq h^N(T,r) + 4K(f)\, e^{-T}. \tag{3.24}$$

The continuity of v^N, (3.18) and the above inequality imply

$$J^N(x,u^*) \leqq v^N(x) + \frac{\delta}{2} + \omega_{v^N}(r) + h^N(T,r) + 4K(f)\, e^{-T}.$$

Recall that $\lim_{r\downarrow 0} h^N(T,r) = 0$; thus we can choose T and r so that $J^N(x,u^*) \leqq v^N(x) + \delta$. □

THEOREM 3.4. *If* (A.2)-(A.4), (1.1)-(1.5) *hold, then* $v \in BUC(\bar\theta)$ *and the dynamic programming relation* (0.8) *holds.*

Proof. Iterating the second assertion of the previous lemma, one gets

$$v^N(x) = \inf_{\{u_1,\cdots,u_N\}\subset \mathcal{A}_{ad}} E\left\{ \sum_{n=0}^{N-1} \int_{T_n}^{T_{n+1}} e^{-t} f(y_0(Y_n, T_n, t, u_n), u_n(Y_n, t-T_n))\, dt \right.$$
$$\left. + \int_{T_N}^{\infty} e^{-t} f(y_0(Y_N, T_N, t, u_N) u_N(Y_N, t-T_N)) \right\}. \tag{3.25}$$

Now define v^∞ by

$$v^\infty(x) = \inf_{\{u_n:\, n=1,2,\cdots\}\subset \mathcal{A}_{ad}} E\left\{ \sum_{n=0}^{\infty} \int_{T_n}^{T_{n+1}} e^{-t} f(y_0(Y_n, T_n, t, u_n), u_n(Y_n, t-T_n))\, dt \right\}. \tag{3.26}$$

Hence we have

$$\sup_{x\in\bar\theta} |v^N(x) - v^\infty(x)| \leqq 2K(f) \sup_{\{u_1,\cdots,u_N\}\subset \mathcal{A}_{ad}} E(e^{-T_N}).$$

To prove that v^N converges to v^∞ it suffices to show that $E(e^{-T_N})$ is decreasing to zero independent of control.

Let $\{u_n:\, n=1,\cdots\} \subset \mathcal{A}_{ad}$ and set $\lambda_n(t) = \lambda(y_0(Y_{n-1}, 0, t, u_n), u_n(Y_{n-1}, t))$.

$$E(e^{-T_n+T_{n-1}} | Y_1, \cdots, Y_{n-1}, T_{n-1}) = \int_0^\infty e^{-t}\lambda_n(t) \exp\left\{ -\int_0^t \lambda_n(s)\, ds \right\} dt$$
$$= 1 - \int_0^\infty e^{-t} \exp\left\{ -\int_0^t \lambda_n(s)\, ds \right\} dt$$
$$\leqq 1 - \int_0^\infty e^{(-(1+K(\lambda))t)}\, dt := \gamma < 1.$$

Observe that γ is independent of control and strictly less than one. Hence

$$\sup_{\{u_1,\cdots,u_N\}\subset \mathcal{A}_{ad}} E(e^{-T_N}) \leqq \gamma^N. \tag{3.27}$$

Therefore v^N converges to v^∞ uniformly on $\bar{\theta}$. So $v^\infty \in BUC(\bar{\theta})$ and v^∞ satisfies (0.8).

We now proceed to show that $v^\infty = v$. First observe that $v^\infty \leqq v$ because of the definitions of v^∞ and v. Also we can construct u^* as in the previous lemma such that for all $x \in \bar{\theta}$

$$E\left\{\int_0^{T_1} e^{-t} f(y_0(x, 0, t, u^*), u^*(x, t))\, dt + e^{-T_1} v^\infty(Y_1)\right\} \leqq v^\infty(x) + \delta. \tag{3.28}$$

Apply (3.28) at $x = Y_1$ to obtain

$$v^\infty(Y_1) \geqq -\delta + E\left\{\int_0^{T_2 - T_1} e^{-t} f(y_0(Y_1, 0, t, u^*), u^*(Y_1, t))\, dt + e^{-(T_2 - T_1)} v^\infty(Y_2) \,\middle|\, Y_1, T_1\right\}.$$

The above inequality and (3.28) yields

$$E\left\{\sum_{n=0}^{1} \int_{T_n}^{T_{n+1}} e^{-t} f(y_0(Y_n, T_n, t, u^*), u^*(Y_n, t - T_n))\, dt + e^{-T_2} v^\infty(Y_2)\right\} \leqq v^\infty(x) + \delta E \sum_{n=0}^{1} e^{-T_n}. \tag{3.29}$$

Iterate this procedure to obtain

$$J(x, u^*) \leqq v^\infty(x) + \delta E\left(\sum_{n=0}^{\infty} e^{-T_n}\right) \leqq v^\infty(x) + \delta(1 - \gamma)^{-1}$$

where γ is as in (3.27). □

Acknowledgments. This paper comprises a part of the author's thesis written under the direction of Professor W. H. Fleming at Brown University. The author would like to express thanks to Professor W. H. Fleming for suggesting the problem, helpful conversations, good advice and careful reading of an earlier manuscript, which led to substantial changes.

REFERENCES

[1] 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.

[2] M. G. CRANDALL AND P.-L. LIONS, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1-42.

[3] 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, Knowles and Lewis, eds., North-Holland, Amsterdam, 1983.

[4] M. H. A. DAVIS, *Piecewise deterministic Markov processes: a general class of non-diffusion stochastic models*, J. Roy. Statist. Soc. (B), to appear.

[5] S. D. DESKMUKH AND S. R. PLISKA, *Optimal consumption and exploration of non-renewable resources under uncertainty*, Econometrica, 48 (1980), pp. 177-200.

[6] W. H. FLEMING AND R. RISHEL, *Deterministic and Stochastic Optimal Control*, Springer, Berlin, 1975.

[7] P.-L. LIONS, *Generalized Solutions of Hamilton-Jacobi Equations*, Research Notes in Mathematics 69, Pitman, London, 1982.

[8] P.-L. LIONS, *Optimal control and viscosity solutions*, Proc. Rome meeting, 1984, to appear in Springer Lecture Notes in Mathematics.
[9] S. R. PLISKA, *On a functional differential equation that arises in a Markov control problem*, J. Differential Equations, 28 (1978), pp. 390–405.
[10] H. M. SONER, *Optimal control of a one-dimensional storage process*, Appl. Math. Optim., 13 (1985), pp. 175–191.
[11] ———, *Optimal control with state space constraint*, I, this Journal, 24 (1986), pp. 552–561.
[12] D. VERMES, *Optimal control of piecewise deterministic processes*, Stochastics, to appear.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 6, November 1986

003

LIPSCHITZIAN SOLUTIONS OF PERTURBED NONLINEAR PROGRAMMING PROBLEMS*

B. CORNET† AND J.-PH. VIAL†

Abstract. We prove that if a second order sufficient condition and a constraint regularity assumption hold, then for sufficiently small perturbations of the constraints and the objective function, the set of local minimizers reduces to a singleton. Moreover, the minimizer and the associated multipliers are Lipschitzian functions of the parameter.

Key words. stability, nonlinear programming, weak convexity

AMS(MOS) subject classifications. 90C31, 90C30

1. Introduction. This paper deals with the stability of solutions and multipliers of nonlinear programming problems when the data are subjected to small perturbations. In order to formulate the problem, we introduce an open subset U of $\mathbb{R}^n$, a metric space P, functions f and g from $U \times P$ to $\mathbb{R}$ and $\mathbb{R}^m$ and a nonempty closed subset Q of $\mathbb{R}^m$. The problem of interest is then:

$$P(\alpha) \qquad \begin{array}{ll} \text{minimize} & f(x, \alpha), \\ \text{subject to} & g(x, \alpha) \in Q, \qquad x \in U, \end{array}$$

where x is the variable in which the minimization is done and α a perturbation parameter which belongs to P and which remains fixed in the minimization problem.

We are interested in the behavior of local minimizers of $P(\alpha)$ when the parameter α varies. Our main result can be informally stated as follows. Under a set of assumptions dealing with (i) the smoothness of the functions f and g, (ii) the regularity of the constraints at $(\bar{x}, \bar{\alpha})$, (iii) the weak convexity of the set Q and (iv) a strong sufficient second-order condition at $(\bar{x}, \bar{\alpha})$, it is shown that, for small perturbations of the parameter $\bar{\alpha}$, the solution $\bar{x}$ of $P(\bar{\alpha})$ persists and is in Lipschitzian dependence with respect to the parameter. The importance of this Lipschitz property should be appreciated in the light of recent developments of calculus for Lipschitzian mappings (Clarke (1975), Rockafellar (1981)).

The above formulation of problem $P(\alpha)$ allows us to take into account the classical nonlinear programming problem with equality and/or inequality constraints, i.e., $Q = \{0\}^{m_1} \times (-\mathbb{R}_+^{m_2})$ for nonnegative integers m_1 and m_2. The consideration of more general sets Q in $P(\alpha)$ is motivated by the following property of weakly convex sets, a class of sets introduced by Vial (1983) (see also Cornet (1981)), which includes as special cases, convex subsets of $\mathbb{R}^m$ and twice continuously differentiable submanifolds of $\mathbb{R}^m$ with or without a boundary. Let Q be a nonempty closed subset of $\mathbb{R}^m$ and let α be in $\mathbb{R}^m$; then the set of projections of α on Q, denoted $\pi(\alpha) = \{x \in Q \mid \|x - \alpha\| \leqq \|x' - \alpha\|$, for all x' in $Q\}$, clearly is the set of solutions of problem $P(\alpha)$ for well chosen mappings f and g. An important property of weakly convex sets is that the mapping π is single-valued and Lipschitzian on a neighborhood of Q. Our main theorem generalizes the known results for problems with equality and/or inequality constraints and also includes the above property of weakly convex sets.

* Received by the editors March 31, 1983, and in revised form May 1, 1985.
† CORE, 1348 Louvain-La-Neuve, Belgium.

We conclude this section by indicating the link between this paper and the rest of the literature. The standard problem with equality and/or inequality constraints has been studied by Fiacco and McCormick (1968), Robinson (1974), Fiacco (1976). The basic feature of these articles is that, under the strict complementarity slackness assumption, it is proved, using the standard implicit function theorem, that the stationary points (i.e., points that satisfy the first order necessary condition for optimality), and their associated multipliers are differentiable.

The strict complementarity slackness assumption has been removed, in the case of equality and/or inequality constraints by Robinson (1980), Kojima (1980), Jittorntrum (1984) and in the case where Q is a closed convex subset of $\mathbb{R}^m$ by Cornet-Laroque (1986) (see also Cornet-Laroque (1980), Cornet (1981)) and J.-P. Aubin (1981) when the constraints are linear. In these cases, under a somewhat stronger second order sufficient condition, a similar result is shown to hold, namely that the stationary points and associated multipliers are (locally) Lipschitzian mappings of the perturbation. The main tool for these analyses are generalizations of the standard implicit function theorem; for example Robinson (1980) proves a general implicit theorem for "generalized equations" (i.e. of variational inequalities), Cornet-Laroque (1986) use a generalization of the implicit function theorem in the case of Lipschitzian mappings due to Clarke (1976) (see also Auslender (1983)) and J.-P. Aubin (1981) a generalization of it in the case of convex processes.

Our approach in the present paper is different from the previous ones. It is direct in the sense that no implicit function theorem or generalization of it is used. We are able to show the local persistence of a local minimizer of our problem and next the Lipschitzian dependence with respect to the parameter α. The first step owes much to a result of Robinson (1982). Finally we shall mention the work of Levitin (1975) who made an analysis of the Lipschitz dependence of local minimizers, also by a direct approach. However, it contains an apparent error as is pointed out in the paper of Robinson (1982).

Our paper is organized as follows. In § 2, we recall some definitions and state the main result of the paper. We also discuss two noteworthy applications: the first deals with the projection of points on a weakly convex set, and the second deals with the standard nonlinear programming problem. The proof of the main theorem is given in § 3.

2. Statement of the main theorem and some consequences. Let us first introduce some notations and definitions. Let $x=(x_i)$, $y=(y_i)$ be in $\mathbb{R}^q$; we denote $\langle x, y\rangle = \sum_{i=1}^q x_i \cdot y_i$, the scalar product of $\mathbb{R}^q$, and $\|x\| = \langle x, x\rangle^{1/2}$ the Euclidean norm. Let A be a nonempty subset of $\mathbb{R}^q$ and let x be in $\mathbb{R}^q$; we denote $d_A(x) = \inf\{\|a-x\| \mid a \in A\}$, $B(A, \varepsilon) = \{x \in \mathbb{R}^q \mid d_A(x) < \varepsilon\}$ and $\bar{B}(A, \varepsilon) = \{x \in \mathbb{R}^q \mid d_A(x) \leqq \varepsilon\}$. Let Q be a subset of $\mathbb{R}^m$ and let x be in $\bar{Q}$; we recall the following definitions of Clarke (1975) of the tangent cone $T_Q(x)$ and the normal cone $N_Q(x)$ to Q at x,

$$T_Q(x) = \{v \in \mathbb{R}^m \mid \text{for all sequences } \{\theta_k\} \subset (0, \infty) \text{ and } \{x_k\} \subset \bar{Q} \text{ such that } \theta_k \to 0,$$
$$x_k \to x, \text{ there exists a sequence } \{v_k\} \to v \text{ such that, for all } k,\ x_k + \theta_k v_k \in \bar{Q}\},$$
$$N_Q(x) = \{\eta \in \mathbb{R}^m \mid \langle \eta, v\rangle \leqq 0, \text{ for all } v \in T_Q(x)\}.$$

DEFINITION 1. A subset Q of $\mathbb{R}^m$ is said to be weakly convex, with constant $\rho \geqq 0$, at an element y^0 in $\bar{Q}$, if there exists $\varepsilon > 0$ such that, for all y^1, y^2 in $\bar{Q} \cap B(y^0, \varepsilon)$ and for all $\lambda^2 \in N_Q(y^2) \cap \bar{B}(0, 1)$, one has

$$\langle \lambda^2, y^2 - y^1\rangle \geqq -\frac{\rho}{2}\|y^2 - y^1\|^2.$$

It is possible to give the following geometric interpretation of weakly convex sets, of constant $\rho > 0$. Let Q be such a set and let $X = \bar{Q} \cap B(y^0, \varepsilon)$ with $y_0 \in \bar{Q}$. Then for all $x \in X$ and $\eta \in N_Q(x) \cap \bar{B}(0, 1) (= N_X(x) \cap \bar{B}(0, 1))$,

$$X \cap B(x + \rho^{-1}\eta, \rho^{-1}\|\eta\|) = \varnothing.$$

For $\eta \neq 0$, one could view $B(x + \rho^{-1}\eta, \rho^{-1}\|\eta\|)$ as a "supporting ball," very much in a sense analogous to a supporting hyperplane for a convex set. In this terminology, if a set is weakly convex at y^0, one can exhibit a "supporting ball" at each point of the boundary of the set in a neighborhood of y^0. Note that the radius of the "supporting ball" is fixed in the given neighborhood. Clearly, a convex subset Q of $\mathbb{R}^m$ is weakly convex with respect to any constant $\rho \geqq 0$. We refer to Cornet (1981), Vial (1983) for other examples of weakly convex sets (such as C^2 submanifolds in $\mathbb{R}^m$ with or without a boundary) and/or for properties of weakly convex sets.

We posit the following assumptions, which describe the general framework of the paper.

Assumptions A.0

(i) U is an open subset of $\mathbb{R}^n$; P is a metric space endowed with a distance d;

(ii) the functions $f(\cdot, \cdot)$ and $g_i(\cdot, \cdot)$, $i = 1, \cdots, m$, are locally Lipschitzian from $U \times P$ to $\mathbb{R}$;

(iii) for all $\alpha \in P$, the functions $f(\cdot, \alpha)$ and $g_i(\cdot, \alpha)$, $i = 1, \cdots, m$, are twice continuously differentiable from U to $\mathbb{R}$;

(iv) the mappings $\nabla f(\cdot, \cdot)$ and $\nabla g_i(\cdot, \cdot)$, $i = 1, \cdots, m$, of first partial derivatives with respect to the first argument, are locally Lipschitzian from $U \times P$ to $\mathbb{R}^n$;

(v) the mapping $D^2 f(\cdot, \cdot)$ and $D^2 g_i(\cdot, \cdot)$ of second order derivatives with respect to the first argument, are continuous;

(vi) $m = m_1 + m_2$, where m_1 and m_2 are nonnegative integers; C is a nonempty closed subset of $\mathbb{R}^{m_2}$ and $Q = \{0\}^{m_1} \times C$ (with the convention that $Q = \{0\}^m$ if $m_2 = 0$ and $Q = C$ if $m_1 = 0$).

We consider the following perturbed nonlinear programming problem:

$$P(\alpha) \qquad \begin{array}{ll} \text{minimize} & g(x, \alpha), \\ \text{subject to} & g(x, \alpha) \in Q, \\ & x \in U, \end{array}$$

where $g(x, \alpha)$ is the vector in $\mathbb{R}^m$ with coordinates $g_i(x, \alpha)$, $i = 1, \ldots, m$, x is the variable in which the minimization is done, and $\alpha \in P$ is a perturbation term which remains fixed in the minimization problem. Note that it is possible to rewrite the constraints as follows. For all (x, α) in $U \times P$, let $g_E(x, \alpha)$ (resp. $g_I(x, \alpha)$) be the vector in $\mathbb{R}^{m_1}$ (resp. $\mathbb{R}^{m_2}$) with coordinates $g_i(x, \alpha)$, $i = 1, \cdots, m_1$ (resp. $i = m_1 + 1, \cdots, m$). Then x satisfies the constraints of $P(\alpha)$ if and only if:

$$g_E(x, \alpha) = 0 \quad \text{and} \quad g_I(x, \alpha) \in C, \qquad x \in U.$$

With $P(\alpha)$, we associate the following "generalized equation":

$$\begin{aligned} &\nabla f(x, \alpha) + \sum_{i=1}^{m} \lambda_i \nabla g_i(x, \alpha) = 0, \\ &g(x, \alpha) \in Q \quad \text{and} \quad \lambda = (\lambda_i) \in N_Q(g(x, \alpha)). \end{aligned} \tag{2.1}$$

We shall be concerned with pairs $(x^0, \alpha^0) \in U \times P$ such that x^0 is a local minimizer of $P(\alpha^0)$. If we further assume that:

Assumption A.1. The gradients $\nabla g_i(x^0, \alpha^0)$, $i = 1, \cdots, m$, are linearly independent, then we shall prove later (Lemma 3.1) that there exists $\lambda^0 \in \mathbb{R}^m$ such that $(x^0, \alpha^0, \lambda^0)$ solves (2.1). In other words, (2.1) is the first order necessary condition associated with $P(\alpha)$. For such a triplet we posit the following two assumptions:

Assumption A.2. Q is weakly convex with constant $\rho \geqq 0$ at $g(x^0, \alpha^0)$. (Note that it would be equivalent to replace the above statement by "C is weakly convex with constant $\rho \geqq 0$ at $g_I(x^0, \alpha^0)$.")

Assumption A.3. There exist real numbers $a \geqq 0$ and $c > 0$ such that, for all $h \in \mathbb{R}^n$, one has

$$\left\langle \left[D^2 f(x^0, \alpha^0) + \sum_{i=1}^{m} \lambda_i^0 D^2 g_i(x^0, \alpha^0) \right] h, h \right\rangle + a\langle \nabla f(x^0, \alpha^0), h\rangle^2 + a \sum_{i=1}^{m_1} \langle \nabla g_i(x^0, \alpha^0), h\rangle^2 \geqq c\|h\|^2.$$

Note that the index in the first sum runs from 1 to $m = m_1 + m_2$ and from 1 to m_1 in the second. (A.3) clearly implies the more familiar assumption:

Assumption A.3′. For all $h \in \mathbb{R}^n$, $h \neq 0$, such that $\langle \nabla f(x^0, \alpha^0), h\rangle = 0$ and $\langle \nabla g_i(x^0, \alpha^0), h\rangle = 0$, $i = 1, \cdots, m_1$, one has

$$\left\langle \left[D^2 f(x^0, \alpha^0) + \sum_{i=1}^{m} \lambda_i^0 D^2 g_i(x^0, \alpha^0) \right] h, h \right\rangle > 0.$$

It is an immediate consequence of a lemma of Debreu (1952) that (A.3′) implies (A.3). Thus the two assumptions are equivalent.

We can now state the main theorem.

THEOREM 2.1. *Assume* (A.0) *and let* $(x^0, \alpha^0, \lambda^0) \in U \times P \times \mathbb{R}^m$ *satisfy* (A.1), (A.2), (A.3). *Further, assume that the constants* $\rho \geqq 0$ *and* $c > 0$ *satisfy*

Assumption A.4. $c > \lambda_I^0 \|Dg_I(x^0, \alpha^0)\|^2$, *where* $\lambda_I^0 = (\lambda_{m_1}^0 + 1, \cdots, \lambda_m^0)$.

Then, if $(x^0, \alpha^0, \lambda^0)$ *satisfies condition* (2.1), *there exist neighborhoods* U' *of* x^0 *in* U, V' *of* α^0 *in* P *and mappings* $x(\cdot): V' \to U'$, $\lambda(\cdot): V' \to \mathbb{R}^m$ *such that*:

(i) $x(\cdot)$ *and* $\lambda(\cdot)$ *are Lipschitzian*;

(ii) $x(\alpha^0) = x^0$ *and* $\lambda(\alpha^0) = \lambda^0$;

(iii) *for all* α *in* V', $x(\alpha)$ *is the unique minimizer of* $P(\alpha)$ *in* U' *and* $\lambda(\alpha)$ *is the unique Kuhn-Tucker multiplier associated with* $x(\alpha)$ (*i.e.*, $(x(\alpha), \lambda(\alpha))$ *satisfies condition* (2.1)).

The proof of Theorem 2.1 is given in the next section.

Remark 1. Assumption (A.1) cannot be relaxed in the case of equality and/or inequality constraints (i.e., when $Q = \{0\}^{m_1} \times (-\mathbb{R}_+^{m_2})$) by only assuming the Mangasarian-Fromovitz's constraint qualification (see Robinson (1980)).

Remark 2. If C is convex, then (A.4) is trivially satisfied (since convex sets in $\mathbb{R}^{m_2}$ are weakly convex with constant $\rho = 0$). It is worth pointing out that in the case of equality and/or inequality constraints, (A.3) is stronger than the classical sufficient second-order condition of Fiacco and McCormick (1968). However, Theorem 2.1 does not hold if one replaces (A.3) by the classical second order condition as it has been shown by Robinson (1980).

We conclude this section by discussing two noteworthy applications of Theorem 2.1. The first one deals with the projection mapping on a weakly convex subset of $\mathbb{R}^n$.

Let Q be a nonempty closed subset of $\mathbb{R}^n$. For a fixed element α in $\mathbb{R}^n$, we consider the following minimization problem:

$$R(\alpha)\qquad \begin{aligned} &\text{minimize} && \tfrac{1}{2}\|x-\alpha\|^2,\\ &\text{subject to} && x\in Q, \end{aligned}$$

and we denote by $\pi(\alpha)$ the set of its solutions. Any element of $\pi(\alpha)$ is called a projection of α on Q. The next proposition gives some properties of the (multi-valued) mapping $\alpha\to\pi(\alpha)$ when Q is assumed to be weakly convex.

COROLLARY 2.2. *Let $(x^0,\alpha^0)\in\mathbb{R}^n\times\mathbb{R}^n$ be such that x^0 is a local minimizer of $R(\alpha^0)$. Assume that Q is weakly convex with constant $\rho\geqq 0$ at x^0 and that $1>\rho\|x^0-\alpha^0\|$. Then, there exist neighborhoods U' of x^0, V' of α^0 and a Lipschitzian mapping $x(\cdot):V'\to U'$ such that $x(\alpha^0)=x^0$ and, for all $\alpha\in V'$, $x(\alpha)$ is the unique minimizer of $R(\alpha)$ in U'.*

Proof. It is a trivial matter to check that (A.1) and (A.3) are satisfied with $c=1$ and that the Kuhn-Tucker multiplier λ^0 associated with x^0 satisfies $\lambda^0=-(x^0-\alpha^0)$. Thus (A.4) reduces to $1=c>\rho\|x^0-\alpha^0\|$. Hence the result.

Remark 3. When $m_1=0$, we give here an example showing that the inequality in (A.4) is the best possible. Let $Q=\{x\in\mathbb{R}^m\,|\,\|x\|\geqq 1\}$; clearly, Q is weakly convex at every element x in Q, with constant $\rho=1$. If $\alpha^0\neq 0$, let $\mu=\max\{1,\|\alpha^0\|^{-1}\}$, then $x^0=\mu\alpha^0$ is the unique minimizer of $R(\alpha^0)$. Since the hypotheses of Corollary 2.2 are satisfied at x^0, the conclusion of the corollary holds. However, if $\alpha^0=0$, any x^0 such that $\|x^0\|=1$ is a minimizer of $R(\alpha^0)$. Obviously, $1=\rho\|x^0-\alpha^0\|$; hence Assumption (A.4) is violated and one easily sees directly that the conclusion of Corollary 2.2 cannot hold.

The second application deals with standard nonlinear programming. Assume A.0 and assume furthermore that $Q=\{0\}^{m_1}\times(-\mathbb{R}_+^{m_2})$ (i.e., in (vi) of (A.0), $C=-\mathbb{R}_+^{m_2}$). We consider the standard perturbed nonlinear programming problem:

$$S(\alpha)\qquad \begin{aligned} &\text{minimize} && f(x,\alpha),\\ &\text{subject to} && g_i(x,\alpha)=0,\ i=1,\cdots,m_1,\\ &&& g_i(x,\alpha)\leqq 0,\ i=m_1+1,\cdots,m,\\ &&& x\in U. \end{aligned}$$

With $S(\alpha)$, we associate the first order necessary conditions:

$$\begin{aligned} &\nabla f(x,\alpha)+\sum_{i=1}^{m_1+m_2}\lambda_i\nabla g_i(x,\alpha)=0,\\ &g_i(x,\alpha)=0, \qquad i=1,\cdots,m_1,\\ &g_i(x,\alpha)\leqq 0,\quad \lambda_i\geqq 0,\quad \lambda_i g_i(x,\alpha)=0, \qquad i=m_1+1,\cdots,m_1+m_2. \end{aligned}\tag{2.2}$$

In the case of standard nonlinear programming, Assumptions (A.2) and (A.4) are satisfied, since $Q=\{0\}^{m_1}\times(-\mathbb{R}_+)^{m_2}$ is convex. We give now a consequence of Theorem 2.1, where the Assumptions (A.1) and (A.3) are weakened. First, let us introduce the following notation: for $(x,\alpha)\in U\times P$, satisfying the constraints of $S(\alpha)$, we let $I(x,\alpha)=\{i\in\{m_1+1,\cdots,m\}|g_i(x,\alpha)=0\}$ be the set of active inequality constraints.

COROLLARY 2.3. *Assume* (A.0) *and let $(x^0,\alpha^0,\lambda^0)\in U\times P\times\mathbb{R}^m$ be such that*:

(C.1) *the vectors $\nabla g_i(x^0,\alpha^0)$, $i\in\{1,\cdots,m_1\}\cup I(x^0,\alpha^0)$, are linearly independent*;

(C.2) *for all $h\in\mathbb{R}^n$, $h\neq 0$, satisfying $\langle\nabla f(x^0,\alpha^0),h\rangle=0$ and $\langle\nabla g_i(x^0,\alpha^0),h\rangle=0$, $i\in\{1,\cdots,m_1\}\cup\{i\in I(x^0,\alpha^0)|\lambda_i^0>0\}$, then*

$$\left\langle\left[D^2f(x^0,\alpha^0)+\sum_{i=1}^{m}\lambda_i^0D^2g_i(x^0,\alpha^0)\right]h,h\right\rangle>0.$$

Then, if $(x^0, \alpha^0, \lambda^0)$ *satisfies condition* (2.2), *there exist neighborhoods* U' *of* x^0 *in* U, V' *of* α^0 *in* P *and mappings* $x(\cdot): V' \to U'$, $\lambda(\cdot): V' \to \mathbb{R}^m$ *such that*:

(i) $x(\cdot)$ *and* $\lambda(\cdot)$ *are Lipschitzian*;

(ii) $x(\alpha^0) = x^0$ *and* $\lambda(\alpha^0) = \lambda^0$;

(iii) *for all* $\alpha \in V'$, $x(\alpha)$ *is the unique minimizer of* $S(\alpha)$ *in* U' *and* $\lambda(\alpha)$ *is the unique Kuhn–Tucker multiplier associated with it* (*i.e.*, $(x(\alpha), \lambda(\alpha))$ *satisfies condition* (2.2)).

The proof of Corollary 2.3 is given in the next section.

3. Proofs. We prepare the proofs of Theorem 2.1 by several lemmas. Some of them are more or less known. However the entire proofs are given for the sake of completeness. Our first lemma says that it is sufficient to prove Theorem 2.1 with Assumption (A.3) replaced by the stronger one

Assumption A.3. bis. There exist real numbers $a \geqq 0$, $c > 0$ such that for all $h \in \mathbb{R}^n$ one has

$$\left\langle D^2 f(x^0, \alpha^0)h + \sum_{i=1}^{m} \lambda_i^0 D^2 g_i(x^0, \alpha^0)h, h \right\rangle + a \sum_{i=1}^{m_1} \langle \nabla g_i(x^0, \alpha^0), h\rangle^2 \geqq c\|h\|^2,$$

(i.e. in (A.3.bis) the term $a\langle \nabla f(x^0, \alpha^0), h\rangle^2$ which appears in (A.3) has been removed.

LEMMA 3.0. *If Theorem* 2.1 *is true when* (A.3) *is replaced by the stronger Assumption* (A.3*bis*), *then it is also true under Assumption* (A.3).

Proof. Let us assume that the weak form of Theorem 2.1 holds (i.e. with Assumption (A.3) replaced by (A.3bis)). We show that Theorem 2.1 also holds. Let $(x^0, \alpha^0, \lambda^0) \in U \times P \times \mathbb{R}^m$ satisfy Assumptions (A.0), (A.1), (A.2), (A.3), (A.4) together with the necessary conditions (2.1) associated with the problem:

$$P(\alpha) \quad \begin{aligned} &\text{minimize } f(x, \alpha) \\ &\text{subject to } g(x, \alpha) \in Q, \\ &\qquad x \in U. \end{aligned}$$

We now associate to all $\alpha \in P$ the following modified problem:

$$\tilde{P}(\alpha) \quad \begin{aligned} &\text{minimize } v \\ &\text{subject to } f(x, \alpha) - v = 0, \\ &\qquad g(x, \alpha) \in Q, \\ &\qquad (v, x) \in \mathbb{R} \times U. \end{aligned}$$

Clearly (v, x) is a solution of $\tilde{\mathcal{P}}(\alpha)$ if and only if x is a solution of $\mathcal{P}(\alpha)$ and $v = f(x, \alpha)$. Moreover at this solution $[(v, x), (\mu, \lambda)]$ satisfies the necessary conditions (2.1) associated with $\tilde{\mathcal{P}}(\alpha)$ if and only if $\mu = 1$, (x, λ) satisfies the necessary conditions (2.1) associated with $P(\alpha)$ and $v = f(x, \alpha)$.

Hence $((v^0, x^0), (1, \lambda^0))$, with $v^0 = f(x^0, \alpha^0)$ satisfies the necessary conditions (2.1) associated with $\tilde{\mathcal{P}}(\alpha^0)$. From the fact that $(x^0, \alpha^0, \lambda^0)$ satisfies Assumptions (A.0), (A.1), (A.2), (A.3) and (A.4) for problem $\tilde{\mathcal{P}}(\alpha^0)$ one deduces that $((v^0, x^0), (1, \lambda^0))$ satisfies Assumptions (A.0), (A.1), (A.2), (A.3bis) and (A.4) for problem $\tilde{\mathcal{P}}(\alpha^0)$. Applying the weak form of Theorem 2.1 to $((v^0, x^0), (1, \lambda^0))$ and using the above equivalence property between $\mathcal{P}(\alpha)$ and $\tilde{\mathcal{P}}(\alpha)$ one deduces the end of the proof of Lemma 3.0.

In the sequel we shall assume that $(x^0, \alpha^0, \lambda^0)$ satisfies (A.0), (A.1), (A.2), (A.3bis) and (A.4).

LEMMA 3.1. *There exist neighborhoods U_1 of x^0, $U_1 \subset U$, V_1 of α^0 such that, for all $\alpha \in V_1$, if $x \in U_1$ is a local minimizer of $P(\alpha)$, then there exists $\lambda = (\lambda_i) \in \mathbb{R}^m$ such that (x, λ) satisfies the first order necessary conditions at α, i.e.,*

$$\nabla f(x, \alpha) + \sum_{i=1}^{m} \lambda_i \nabla g_i(x, \alpha) = 0,$$

$$g(x, \alpha) \in Q \quad \text{and} \quad \lambda \in N_Q(g(x, \alpha)).$$

Proof. Since by (A.1) the gradients $\nabla g_i(x^0, \alpha^0)$, $i \in \{1, \cdots, m\}$ are independent from (A.0), there exist neighborhoods U_1 of x^0, $U_1 \subset U$, V_1 of α^0 such that, for all $(x, \alpha) \in U_1 \times V_1$, the vectors $\nabla g_i(x, \alpha)$, $i \in \{1, \cdots, m\}$, are independent. In the sequel of this proof, since α is fixed, there is no ambiguity denoting $f(x, \alpha), \nabla f(x, \alpha), \cdots$, simply by $f(x), \nabla f(x), \cdots$.

Let $X = \{x \in U_1 | g(x) \in Q\}$. If x is a local minimizer of $P(\alpha)$, from Clarke (1975), for all $v \in T_x(x)$ one has $\langle -\nabla f(x), v \rangle \leqq 0$. Hence from Rockafellar (1970, Cor. 16.3.2), it is sufficient to prove that the following inclusion holds:

$$\{u \in \mathbb{R}^n | Dg(x)u \in T_Q(g(x))\} \subset T_X(x).$$

Indeed, let $u \in \mathbb{R}^n$ be such that $Dg(x)u \in T_Q(g(x))$ and let $\{\theta^q\} \subset (0, \infty)$, $\{x^q\} \subset X$ be sequences such that $\theta^q \to 0$ and $x^q \to x$. From the definition of the tangent cone it is sufficient to show that there exists a sequence $\{u^q\} \subset \mathbb{R}^n$, such that $u^q \to u$ and, for all q, $x^q + \theta^q u^q \in X$. Let $v = Dg(x)u$, then $v \in T_Q(g(x))$. Since, for all q, $g(x^q) \in Q$, and $g(x^q) \to g(x)$, there exists a sequence $\{v^q\} \subset \mathbb{R}^m$ such that, for all q, $g(x^q) + \theta^q v^q \in Q$ and $v^q \to v$. We can choose vectors $b_{m+1}, \cdots, b_n$ in $\mathbb{R}^n$ so that the vectors $\nabla g_1(x), \cdots, \nabla g_m(x)$, $b_{m+1}, \cdots, b_n$, form a basis in $\mathbb{R}^n$, and we define the mapping $G: U_1 \to \mathbb{R}^n$ by $G(y) = (g_1(y), \cdots, g_m(y), \langle b_{m+1}, y \rangle, \cdots, \langle b_n, y \rangle)$. Clearly, G is continuously differentiable, and the derivative $DG(x)$ is nonsingular. Hence, by the inverse mapping theorem, there exists q_0 such that, for $q \geqq q_0$, there exists $\hat{x}^q \in U_1$ satisfying:

$$g(\hat{x}^q) = g(x^q) + \theta^q v^q,$$

$$\langle b_i, \hat{x}^q \rangle = \langle b_i, x^q \rangle + \theta^q \langle b_i, u \rangle \qquad (i = m+1, \cdots, n),$$

and such that $\hat{x}^q \to x$. For $q \geqq q_0$, let $u^q = (\hat{x}^q - x^q)/\theta^q$. Recall that, for all q, $g(x^q) + \theta^q v^q \in Q$, hence, for $q \geqq q_0$, $g(\hat{x}^q) = g(x^q) + \theta^q v^q \in Q$; thus $\{\hat{x}^q\} \subset X$. Consequently, for $q \geqq q_0$, $x^q + \theta^q u^q = \hat{x}^q \in X$. To end the proof of the lemma, it suffices to show that $u^q \to u$. Indeed, from Taylor's theorem, for $q \geqq q_0$, one has $g(\hat{x}^q) - g(x^q) = [\int_Q^1 Dg(x^q + t(\hat{x}^q - x^q))\, dt](\hat{x}^q - x^q)$. Dividing by $\theta^q > 0$, one gets $v^q = [\int_0^1 Dg(x^q + t(\hat{x}^q - x^q))\, dt]u^q$ and one easily deduces that $\lim_{q\to\infty} Dg(x)u^q = \lim_{q\to\infty} v^q$. Recall that $\lim_{q\to\infty} v^q = v = Dg(x)u$; hence, for all $i \leqq m$, $\lim_{q\to\infty} \langle \nabla g_i(x), u^q \rangle = \langle \nabla g_i(x), u \rangle$. Furthermore, for all $i \geqq m+1$, $\langle b_i, u^q \rangle = \langle b_i, (\hat{x}^q - x^q)/\theta^q \rangle = \langle b_i, u \rangle$. Since the vectors $\{\nabla g_1(x), \cdots, \nabla g_m(x), b_{m+1}, \cdots, b_n\}$ are independent, one deduces that $u^q \to u$. This ends the proof of the lemma.

LEMMA 3.2. *For all $\varepsilon > 0$ and all $c' \in (0, c)$, there exist positive real numbers k_1, k_2, a positive real number δ (independent of ε) and neighborhoods U_2 of x^0, $U_2 \subset U$, V_2 of α^0, such that the two following properties are satisfied.*

(a) *For all (x, α), (y, β), (x^1, α^1), (x^2, α^2) in $U_2 \times V_2$ such that $g(x^1, \alpha^1) \in Q$, $g(x^2, \alpha^2) \in Q$, for all $\lambda \in B(\lambda^0, \delta)$ one has:*

$$\left\langle \left[D^2 f(x, \alpha) + \sum_{i=1}^{m} \lambda_i D^2 g_i(y, \beta) \right](x^2 - x^1), (x^2 - x^1) \right\rangle$$
$$\geqq c' \|x^2 - x^1\|^2 - k_1 \|x^2 - x^1\| d(\alpha^2, \alpha^1) - k_1 d(\alpha^2, \alpha^1)^2.$$

(b) *For all* (x^1, α^1), (x^2, α^2) *in* $U_2 \times V_2$, *such that* $g(x^2, \alpha^2) \in Q$, $g(x^1, \alpha^1) \in Q$, *for all* $\lambda^2 \in N_Q(g(x^2, \alpha^2))$ *one has*:

$$\langle \lambda^2, g(x^2, \alpha^2) - g(x^1, \alpha^1) \rangle$$

$$\geqq \|\lambda_I^2\| \cdot \left[-\frac{\rho}{2} [\|Dg_I(x^0, \alpha^0)\| + \varepsilon]^2 \cdot \|x^2 - x^1\|^2 - k_2 \|x^2 - x^1\| \, d(\alpha^2, \alpha^1) - k_2 d(\alpha^2, \alpha^1)^2 \right].$$

Proof. We first claim that, for all $\varepsilon > 0$, there exist a positive real number k' and neighborhoods U' of x^0, $U' \subset U$, V' of α^0 such that, for all (x^1, α^1), (x^2, α^2) in $U' \times V'$, for all $i \in \{1, \cdots, m\}$ one has:

$$|g_i(x^2, \alpha^2) - g_i(x^1, \alpha^1) - \langle \nabla g_i(x^0, \alpha^0), (x^2 - x^1) \rangle| \leqq \varepsilon \|x^2 - x^1\| + k' d(\alpha^2, \alpha^1). \tag{3.1}$$

Indeed, for all $i \in \{1, \cdots, m\}$, from (A.0), for all $\varepsilon > 0$, there exist open neighborhoods U' of x^0, $U' \subset U$, V' of α^0, such that U' is convex, and, for all $(x, \alpha) \in U' \times V'$, $\|\nabla g_i(x, \alpha) - \nabla g_i(x^0, \alpha^0)\| < \varepsilon$. Furthermore, without any loss of generality, we can assume that there exists a positive real number k' such that g_i is Lipschitzian of constant k' on $U' \times V'$. Hence, from Taylor's theorem,

$$\begin{aligned}
&|g_i(x^2, \alpha^2) - g_i(x^1, \alpha^1) - \nabla g_i(x^0, \alpha^0)(x^2 - x^1)| \\
&\quad \leqq \left| g_i(x^2, \alpha^2) - g_i(x^1, \alpha^2) - \left\langle \left[\int_0^1 \nabla g_i(x^1 + t(x^2 - x^1), \alpha^2) \, dt \right], x^2 - x^1 \right\rangle \right| \\
&\qquad + |g_i(x^1, \alpha^2) - g_i(x^1, \alpha^1)| \\
&\qquad + \int_0^1 \|\nabla g_i(x^1 + t(x^2 - x^1), \alpha^2) - \nabla g_i(x^0, \alpha^0)\| \, dt \cdot \|x^2 - x^1\| \\
&\quad \leqq 0 + k' d(\alpha^2, \alpha^1) + \varepsilon \|x^2 - x^1\|.
\end{aligned}$$

(a) For all $c' \in (0, c)$, let $c'' \in (c', c)$. From (A.3.bis) and the continuity of the mappings $D^2 f(\cdot, \cdot)$ and $D^2 g_i(\cdot, \cdot)$ $(i = 1, \cdots, m)$, there exist neighborhoods U'' of x^0, $U'' \subset U$, V'' of α^0 and a positive real number δ such that, for all (x, α), (y, β) in $U'' \times V''$, for all $\lambda \in B(\lambda^0, \delta)$ and all $h \in \mathbb{R}^n$, one has:

$$\left\langle \left[D^2 f(x, \alpha) + \sum_{i=1}^{m} \lambda_i D^2 g_i(y, \beta) \right] h, h \right\rangle \geqq c'' \|h\|^2 - a \sum_{i=1}^{m_1} \langle \nabla g_i(x^0, \alpha^0), h \rangle^2.$$

Let $\varepsilon > 0$ be such that $c'' - a \cdot m_1 \cdot \varepsilon^2 \geqq c'$, and let U', V' be the neighborhoods associated with ε in the above claim (3.1). Let $U_2 = U' \cap U''$ and $V_2 = V' \cap V''$. Recall that, for all $(x, \alpha) \in U \times V$ such that $g(x, \alpha) \in Q$, from (A.0.vi), for $i \in \{1, \cdots, m_1\}$, $g_i(x, \alpha) = 0$. Hence the end of the proof follows easily from (3.1) and the above inequality.

(b) From the weak convexity Assumption (A.2) and the continuity of the functions g_i $(i = 1, \cdots, m)$, there exist open neighborhoods U'' of x^0, $U'' \subset U$, and V'' of α^0 such that, for all (x^1, α^1), (x^2, α^2) in $U'' \times V''$ such that $g(x^k, \alpha^k) \in Q$ $(k = 1, 2)$, and for all $\mu^2 \in N_Q(g(x^2, \alpha^2)) \cap \bar{B}(0, 1)$, one has:

$$\langle \mu^2, g(x^2, \alpha^2) - g(x^1, \alpha^1) \rangle \geqq -\frac{\rho}{2} \|g(x^2, \alpha^2) - g(x^1, \alpha^1)\|^2.$$

Take any $\varepsilon > 0$ and let U', V' be the neighborhoods associated with ε in the above claim (3.1). Let $U_2 = U' \cap U''$ and $V = V' \cap V''$. From (3.1) and the above inequality, there exists k_2 such that

$$\langle \mu^2, g(x^2, \alpha^2) - g(x^1, \alpha^1) \rangle$$

$$\geqq -\frac{\rho}{2} [\|Dg_I(x^0, \alpha^0)\| + \varepsilon]^2 \cdot \|x^2 - x^1\|^2 - k_2 \|x^2 - x^1\| \cdot d(\alpha^2, \alpha^1) - k_2 d(\alpha^2, \alpha^1)^2.$$

Now letting $\lambda^2 \in N_Q(g(x^2, \alpha^2))$, we let $\mu^2 = (0, \lambda_I^2/\|\lambda_I^2\|)$ if $\lambda_I^2 \neq 0$ and $\mu^2 = 0$ if $\lambda_I^2 = 0$. Since $Q = \{0\} \times C$, $N_Q(g(x^2, \alpha^2)) = \mathbb{R}^{m_1} \times N_C(g_I(x^2, \alpha^2))$ and since $N_C(g_I(x^2, \alpha^2))$ is a cone, one deduces that $\mu^2 \in N_Q(g(x^2, \alpha^2)) \cap \bar{B}(0, 1)$. Applying the above inequality to μ^2 and noticing that

$$\langle \mu^2, g(x^2, \alpha^2) - g(x^1, \alpha^1)\rangle = (1/\|\lambda_I^2\|)\langle \lambda^2, g(x^2, \alpha^2) - g(x^1, \alpha^1)\rangle,$$

yields the inequality of Lemma 3.2(b).

LEMMA 3.3. *Let us suppose that* (x^0, λ^0) *satisfies the first order necessary condition at* α^0. *Then there exist positive real numbers* r *and* b *such that* $\bar{B}(x^0, r) \subset U$ *and, for all* $x \in \bar{B}(x^0, r)$ *satisfying* $g(x, \alpha^0) \in Q$, *one has* $f(x, \alpha^0) \geqq f(x^0, \alpha^0) + b\|x - x^0\|^2$.

Proof. For every $\lambda = (\lambda_i) \in \mathbb{R}^m$, let $L(\cdot, \lambda): U \to \mathbb{R}$ be the function defined by $L(x, \lambda) = f(x, \alpha^0) + \sum_{i=1}^m \lambda_i g_i(x, \alpha^0)$. From Taylor's theorem, one has:

$$L(x, \lambda^0) = L(x^0, \lambda^0) + \langle \nabla L(x^0, \lambda^0),\ x - x^0\rangle + \int_0^1 (1-t)\langle D^2 L(x^0 + t(x - x^0), \lambda^0)(x - x^0), (x - x^0)\rangle\, dt.$$

Let $\rho > 0$ be the constant of weak convexity defined by (A.2). From (A.4), $c > \rho\|\lambda_I^0\| \cdot \|Dg_I(x^0, \alpha^0)\|^2$. Hence, there exist $c' \in (0, c)$ and $\varepsilon > 0$ such that, if $b = c'/2 - \rho/2\|\lambda_I^0\|[\|Dg_I(x^0, \alpha^0)\| + \varepsilon]^2$, then $b > 0$. Let U_2, V_2 be the neighborhoods of x^0 and α^0 associated with c' and ε in Lemma 3.2, and let r be a positive real number such that $\bar{B}(x^0, r) \subset U_2$. Since (x^0, α^0) satisfies the first order necessary condition (2.1) at α^0, one deduces that $\nabla L(x^0, \lambda^0) = 0$. From Lemma 3.2(a) (taking $\alpha^1 = \alpha^2 = \alpha^0$), for all $x \in \bar{B}(x^0, r)$ one deduces that:

$$L(x, \lambda^0) - L(x^0, \lambda^0) \geqq \int_0^1 (1-t)c'\|x - x^0\|^2\, dt \geqq \frac{c'}{2}\|x - x^0\|^2,$$

and, from the definition of L, one gets:

$$f(x, \alpha^0) - f(x^0, \alpha^0) \geqq \langle \lambda^0, g(x^0, \alpha^0) - g(x, \alpha^0)\rangle + (c'/2)\|x - x^0\|^2.$$

By Lemma 3.2(b) (taking $\alpha^1 = \alpha^2 = \alpha^0$), for all $x \in \bar{B}(x^0, r) \subset U_2$, one gets:

$$\langle \lambda^0, g(x^0, \alpha^0) - g(x, \alpha^0)\rangle \geqq -(\rho/2)\|\lambda_I^0\|[\|Dg_I(x^0, \alpha^0)\| + \varepsilon]^2.$$

Hence, from the two above inequalities and the definition of b, for all $x \in \bar{B}(x^0, r)$, one has $f(x, \alpha^0) - f(x^0, \alpha^0) \geqq b\|x - x^0\|^2$. This completes the proof.

LEMMA 3.4. *Let* $(x^0, \lambda^0) \in U \times \mathbb{R}^m$ *satisfy the first order necessary condition* (2.1) *at* α^0. *Then, for every neighborhood* U' *of* x^0, $U' \subset U$, *there exist* $r' > 0$ *and a neighborhood* V' *of* α^0 *such that* $B(x^0, r') \subset U'$ *and, for all* $\alpha \in V'$, *there exists a minimizer* $x(\alpha)$ *of* $P(\alpha)$ *in* $B(x^0, r')$ *(i.e.,* $x(\alpha) \in B(x^0, r')$, $g(x(\alpha), \alpha) \in Q$ *and for all* $x \in B(r^0, r)$, $g(x, \alpha) \in Q$ *one has* $f(x(\alpha), \alpha) \leqq f(x, \alpha)$*).*

The above lemma is related to a previous result of Robinson (1982), proved when Q is a closed convex cone.

Proof. Let U' be a neighborhood of x^0, $U' \subset U$, and let b, r be the positive real numbers defined by Lemma 3.3. There exists a positive real number $r' < r$ and a neighborhood V_1' of α^0 such that $\bar{B}(x^0, r') \subset U'$ and f is k-Lipschitzian on $\bar{B}(x^0, r') \times V_1'$.

For all $\alpha \in P$, let $\Gamma(\alpha) = \{x \in \bar{B}(x^0, r') | g(x, \alpha) \in Q\}$. We claim that there exists a neighborhood V_2' of α^0 such that, for all α in V_2', there exists $x(\alpha) \in \Gamma(\alpha)$ satisfying:

$$f(x(\alpha), \alpha) \leqq f(x, \alpha) \quad \text{for all } x \in \Gamma(\alpha).$$

Since f is continuous and, for all α in P, $\Gamma(\alpha)$ is a compact subset of $\mathbb{R}^n$, it is sufficient to prove that, for α in a neighborhood of α^0, $\Gamma(\alpha)$ is nonempty. Indeed, by (A.1) the vectors $\nabla g_i(x^0, \alpha^0)$, $i \in \{1, \cdots, m\}$, are independent. Hence, by the implicit function theorem (Schwartz (1967)), there exists a neighborhood V_2' of α^0 and a continuous mapping $\varphi : V_2' \to \bar{B}(x^0, r')$ such that $\varphi(\alpha^0) = x^0$ and, for all $\alpha \in V_2'$, one has $g(\varphi(\alpha), \alpha) = g(x^0, \alpha^0) \in Q$. This ends the proof of the claim.

We now claim that there exists a neighborhood V' of α^0, $V' \subset V_2'$, such that, for all $\alpha \in V'$, the element $x(\alpha)$ defined before satisfies $\|x(\alpha) - x^0\| < r'$. Clearly, the proof of the claim will end the proof of the lemma. Recall that f is k-Lipschitzian on $\bar{B}(x^0, r') \times V_1'$. Let $V_3' = V_1' \cap B(\alpha^0, r'^2 b/16k)$ and let $\eta > 0$, $\eta < \min\{r'/2, r'^2 b/16k\}$. Then, for all $x^1, x^2 \in B(x^0, r')$ satisfying $\|x^2 - x^1\| < \eta$, for all $\alpha \in V_3'$, one has:

$$f(x^1, \alpha) - f(x^2, \alpha^0) \leqq k[\eta + d(\alpha, \alpha^0)] < r'^2 b/8.$$

Furthermore, the multivalued mapping $\alpha \to \Gamma(\alpha) = \{x \in \bar{B}(x^0, r') | g(x, \alpha) \in Q\}$, from V_2' to $\mathbb{R}^n$, is upper semicontinuous, with compact values. Hence, there exists a neighborhood V_4' of α^0, $V_4' \subset V_2'$, such that, for all $\alpha \in V_4'$, $\Gamma(\alpha) \subset B(\Gamma(\alpha^0), \eta)$ and (from the continuity of φ) $\varphi(\alpha) \in B(x^0, \eta)$.

Let $V' = V_1' \cap V_2' \cap V_3' \cap V_4'$. We now show that V' satisfies the conclusion of the lemma. Recall that, for all $\alpha \in V'$, $\varphi(\alpha) \in \Gamma(\alpha)$; thus, from the definition of $x(\alpha)$, one deduces that:

$$f(\varphi(\alpha), \alpha) \geqq f(x(\alpha), \alpha).$$

On the other hand, for all $\alpha \in V'$, $x(\alpha) \in \Gamma(\alpha) \subset B(\Gamma(\alpha^0), \eta)$. Hence, there exists $y^0 \in \Gamma(\alpha^0)$ such that $\|y^0 - x(\alpha)\| < \eta$. By Lemma 3.3, one has:

$$f(y^0, \alpha^0) \geqq f(x^0, \alpha^0) + b\|y^0 - x^0\|^2.$$

Summing up the two above inequalities, for all $\alpha \in V'$, one gets:

$$b\|y^0 - x^0\|^2 \leqq f(\varphi(\alpha), \alpha) - f(x^0, \alpha^0) + f(y^0, \alpha^0) - f(x(\alpha), \alpha).$$

Let us recall that, for $\alpha \in V'$, $\|\varphi(\alpha) - x^0\| < \eta$ and $\|y^0 - x(\alpha)\| < \eta$. Hence, from the above inequality and the inequality defining η, one gets $b\|y^0 - x^0\|^2 \leqq r'^2 b/8 + r'^2 b/8$. Thus, $\|y^0 - x^0\| < r'/2$, and $\|x(\alpha) - x^0\| \leqq \|x(\alpha) - y^0\| + \|y^0 - x^0\| \leqq \eta + r'/2 < r'$. This ends the proof of the claim and the proof of the lemma.

In the following, we denote by $Dg(x, \alpha)^*$ the $n \times m$ matrix whose columns are $\nabla g_i(x, \alpha)$ $(i = 1, \cdots, m)$, i.e., the transpose of the $m \times n$ matrix $Dg(x, \alpha)$.

LEMMA 3.5. *Let ε be a positive real number. There exist a positive real number k_3 and neighborhoods U_3 of x^0, $U_3 \subset U$, V_3 of α^0 such that, for all $(x, \alpha) \in U_3 \times V_3$, the matrix $Dg(x, \alpha) \circ Dg(x, \alpha)^*$ is nonsingular and the mapping $\varphi : U_3 \times V_3 \to \mathbb{R}^m$ defined by*

$$\varphi(x, \alpha) = [Dg(x, \alpha) \circ Dg(x, \alpha)^*]^{-1} \circ Dg(x, \alpha) \nabla f(x, \alpha),$$

satisfies the following properties:

$$\|\varphi(x^2, \alpha^2) - \varphi(x^1, \alpha^1)\| \leqq k_3[\|x^2 - x^1\| + d(\alpha^2, \alpha^1)] \quad \textit{for all } (x^1, \alpha^1), (x^2, \alpha^2) \textit{ in } U_3 \times V_3;$$

$$\|\varphi(x, \alpha) - \varphi(x^0, \alpha^0)\| \leqq \varepsilon, \quad \textit{for all } (x, \alpha) \in U_3 \times V_3.$$

Proof. The proof is a straightforward consequence of the independence Assumption (A.1), using the fact that the mappings $\nabla f(\cdot, \cdot)$, $\nabla g_i(\cdot, \cdot)$ $(i = 1, \cdots, m)$ are locally Lipschitzian and that the mappings $A \to (A \circ A^*)^{-1} \circ A$, defined on the set of $m \times n$ matrices of maximal rank, is infinitely differentiable, and hence locally Lipschitzian.

Proof of Theorem 2.1. Let $c>0$, $\rho>0$ be the constants defined by (A.2) and (A.3). From (A.4), $c>\rho\|\lambda_I^0\|\cdot\|Dg_I(x^0,\alpha^0)\|^2$. Hence, there exist $c'\in(0,c)$ and $\varepsilon>0$ such that

$$c'>\rho[\|\lambda_I^0\|+\varepsilon][\|Dg_I(x^0,\alpha^0)\|+\varepsilon]^2. \tag{3.2}$$

Let $\delta>0$ be the constant defined by Lemma 3.2; without any loss of generality, we can suppose that $\varepsilon<\delta$. From (A.0), there exist a positive real number k and neighborhoods U_0 of x^0, $U_0\subset U$, V_0 of α^0 such that all the mappings $g(\cdot,\cdot)$ $Dg(\cdot,\cdot)$ and $\nabla f(\cdot,\cdot)$ are k-Lipschitzian on $U_0\times V_0$. Furthermore, let U_1, V_1 (resp. U_2, V_2 and U_3, V_3) be the neighborhoods of x^0 and α^0 associated, in Lemma 3.1 (resp. Lemma 3.2 and Lemma 3.5) with the constants c' and ε as defined above.

Now, by Lemma 3.4, we associate with $U'=U_0\cap U_1\cap U_2\cap U_3$ a positive real number r' and an open neighborhood $\tilde{V}$ of α^0. We take $V'=\tilde{V}\cap V_0\cap V_1\cap V_2\cap V_3$. Let α^1, α^2 be two elements in V'. By Lemma 3.4, there exists a minimizer of $P(\alpha^1)$ (resp. $P(\alpha^2)$) in $B(x^0,r')$ not necessarily unique, that we denote by x^1 (resp. x^2). By Lemma 3.1, let $\lambda^1\in\mathbb{R}^m$ (resp. $\lambda^2\in\mathbb{R}^m$) be the Kuhn-Tucker multiplier associated with x^1 (resp. x^2), i.e., such that (x^h,λ^h) $(h=1,2)$ satisfies the first order necessary condition at α^h:

$$-\nabla f(x^h,\alpha^h)=Dg(x^h,\alpha^h)^*\lambda^h,$$

$$g(x^h,\alpha^h)\in Q \quad\text{and}\quad \lambda^h\in N_Q(g(x^h,\alpha^h)).$$

To end the proof of the theorem, it suffices to show that $\|x^2-x^1\|\leqq Kd(\alpha^2,\alpha^1)$ and $\|\lambda^2-\lambda^1\|\leqq Kd(\alpha^2,\alpha^1)$, where K is a positive real number independent of the choice of α^1, α^2 in V'.

From the first part of the first order necessary condition and Lemma 3.5, for $h=1,2$, the matrix $Dg(x^h,\alpha^h)\circ Dg(x^h,\alpha^h)^*$ is invertible and one deduces that:

$$\lambda^h=-[Dg(x^h,\alpha^h)\circ Dg(x^h,\alpha^h)^*]^{-1}Dg(x^h,\alpha^h)\nabla f(x^h,\alpha^h)(=\varphi(x^h,\alpha^h)),$$

$$\|\lambda^2-\lambda^1\|\leqq k_3[\|x^2-x^1\|+d(\alpha^2,\alpha^1)], \tag{3.3}$$

$$\|\lambda_I^h-\lambda_I^0\|\leqq\lambda^h-\lambda^0\|\leqq\varepsilon<\delta. \tag{3.4}$$

From (3.4) and from Lemma 3.2(b), one deduces that:

$$\begin{aligned}[\|\lambda_I^0\|+\varepsilon]^{-1}\cdot\langle\lambda^2-\lambda^1, g(x^2,\alpha^2)-g(x^1,\alpha^1)\rangle&\\ \geqq-\rho[\|Dg_I(x^0,\alpha^0)\|+\varepsilon]^2\|x^2-x^1\|^2&\\ -2k_2\|x^2-x^1\|d(\alpha^2,\alpha^1)-2k_2d(\alpha^2,\alpha^1)^2.&\end{aligned} \tag{3.5}$$

Furthermore,

$$\langle\lambda^2-\lambda^1, g(x^2,\alpha^2)-g(x^1,\alpha^1)\rangle=A+B+C, \tag{3.6}$$

where

$$A=\langle\lambda^2-\lambda^1, g(x^2,\alpha^2)-g(x^2,\alpha^1)\rangle,$$

$$B=\langle-\lambda^2, g(x^1,\alpha^1)-g(x^2,\alpha^1)-g(x^1,\alpha^2)+g(x^2,\alpha^2)\rangle,$$

$$C=\langle-\lambda^1, g(x^2,\alpha^1)-g(x^1,\alpha^1)\rangle+\langle-\lambda^2, g(x^1,\alpha^2)-g(x^2,\alpha^2)\rangle,$$

and we consider successively each one of the three terms. From (3.3), using the Cauchy-Schwarz inequality and the fact that g is k-Lipschitzian on $U'\times V'$, one gets:

$$A\leqq k\,k_3[\|x^2-x^1\|+d(\alpha^2,\alpha^1)]\cdot d(\alpha^2,\alpha^1). \tag{3.7}$$

By (3.4) the multiplier λ^2 is bounded by $\|\lambda^0\|+\varepsilon$; hence, from Taylor's theorem, using the Cauchy-Schwarz inequality, one gets:

$$B \leqq [\|\lambda^0\|+\varepsilon]\|x^2-x^1\| \left\| \int_0^1 Dg(x^2+t(x^1-x^2), \alpha^1)\, dt - \int_0^1 Dg(x^1+u(x^2-x^1), \alpha^2)\, du \right\|.$$

Upon performing the change of variable $t = 1-u$ in the second integral, and using the fact that the derivative $Dg(\cdot,\cdot)$ is k-Lipschitzian on $U' \times V'$, one gets:

$$\begin{aligned} &B \leqq [\|\lambda^0\|+\varepsilon]\|x^2-x^1\| \cdot \int_0^1 \|Dg(x^2+t(x^1-x^2), \alpha^1) - Dg(x^2+t(x^1-x^2), \alpha^2)\|\, dt, \\ &B \leqq k[\|\lambda^0\|+\varepsilon]\|x^2-x^1\| d(\alpha^2, \alpha^1). \end{aligned} \tag{3.8}$$

From Taylor's theorem, one has:

$$\begin{aligned} C = {} & \langle -\lambda^1, Dg(x^1, \alpha^1)(x^2-x^1)\rangle + \langle -\lambda^2, Dg(x^2, \alpha^2)(x^1-x^2)\rangle \\ & - \left\langle \left[\int_0^1 (1-t) \sum_{i=1}^m \lambda_i^1 D^2 g_i(x^1+t(x^2-x^1), \alpha^1)\, dt \right](x^2-x^1), (x^2-x^1) \right\rangle \\ & - \left\langle \left[\int_0^1 (1-t) \sum_{i=1}^m \lambda_i^2 D^2 g_i(x^2+t(x^1-x^2), \alpha^2)\, dt \right](x^2-x^1), (x^2-x^1) \right\rangle. \end{aligned}$$

From the first order necessary condition, Taylor's theorem, and using the fact that $\nabla f(\cdot,\cdot)$ is k-Lipschitzian on $U' \times V'$, one gets:

$$\begin{aligned} &\langle -\lambda^1, Dg(x^1, \alpha^1)(x^2-x^1)\rangle + \langle -\lambda^2, Dg(x^2, \alpha^2)(x^1-x^2)\rangle \\ &\quad = -\langle \nabla f(x^2, \alpha^2) - \nabla f(x^1, \alpha^1), x^2-x^1\rangle \\ &\quad = -\langle \nabla f(x^2, \alpha^2) - \nabla f(x^2, \alpha^1), x^2-x^1\rangle - \langle \nabla f(x^2, \alpha^1) - \nabla f(x^1, \alpha^1), x^2-x^1\rangle \\ &\quad \leqq kd(\alpha^2, \alpha^1)\|x^2-x^1\| \\ &\qquad - \left\langle \left[\int_0^1 D^2 f(x^1+t(x^2-x^1), \alpha^1)\, dt \right](x^2-x^1), (x^2-x^1) \right\rangle. \end{aligned}$$

One easily gets (by performing the change of variable $t = 1-u$) that:

$$\int_0^1 tD^2 f(x^1+t(x^2-x^1), \alpha^1)\, dt = \int_0^1 (1-u) D^2 f(x^2+u(x^1-x^2), \alpha^1)\, du.$$

Hence

$$\begin{aligned} \int_0^1 D^2 f(x^1+t(x^2-x^1), \alpha^1)\, dt = {} & \int_0^1 (1-t) D^2 f(x^1+t(x^2-x^1), \alpha^1)\, dt \\ & + \int_0^1 (1-u) D^2 f(x^2+u(x^1-x^2), \alpha^1)\, du. \end{aligned}$$

From (3.4), $\lambda^h \in \bar{B}(\lambda^0, \delta)$ $(h = 1, 2)$. Hence, from Lemma 3.2(a) and the above equalities and inequalities

$$\begin{aligned} &C \leqq kd(\alpha^2, \alpha^1)\|x^2-x^1\| \\ &\qquad -2 \int_0^1 (1-t)[c'\|x^2-x^1\|^2 - k_1\|x^2-x^1\| d(\alpha^2, \alpha^1) - k_1 d(\alpha^2, \alpha^1)^2]\, dt, \\ &C \leqq -c'\|x^2-x^1\|^2 + (k+k_1)\|x^2-x^1\| d(\alpha^2, \alpha^1) + k_1 d(\alpha^2, \alpha^1)^2. \end{aligned} \tag{3.9}$$

Let $b = c' - \rho[\|\lambda_I^0\| + \varepsilon][\|Dg_I(x^0, \alpha^0)\| + \varepsilon]^2$; then, from (3.2) one has $b > 0$. From (3.5), (3.6), (3.7), (3.8), (3.9), there exists a positive real number k' such that:

$$b\|x^2 - x^1\|^2 \leqq k' d(\alpha^2, \alpha^1)[\|x^2 - x^1\| + d(\alpha^2, \alpha^1)].$$

Let $K = \max\{2, (2k')/b\}$. Then, from the above inequality, one easily deduces that $\|x^2 - x^1\| \leqq K d(\alpha^2, \alpha^1)$, and, from (3.3), $\|\lambda^2 - \lambda^1\| \leqq (k_3 + Kk_3) d(\alpha^2, \alpha^1)$. This ends the proof of the theorem.

We now give the proof of Corollary 2.3.

Proof of Corollary 2.3. Let $(x^0, \alpha^0, \lambda^0)$ satisfy the assumptions of Corollary 2.3 and let:

$$I_+(x^0, \alpha^0) = \{i \in I(x^0, \alpha^0) | \lambda_i^0 > 0\}, \qquad I_0(x^0, \alpha^0) = \{i \in I(x^0, \alpha^0) | \lambda_i^0 = 0\},$$
$$J(x^0, \alpha^0) = \{i \in \{m_1 + 1, \cdots, m\} | g_i(x^0, \alpha^0) < 0\}.$$

For $\alpha \in P$, we consider the following problem:

$$\tilde{S}(\alpha) \quad \begin{aligned} &\text{minimize } f(x, \alpha) \\ &\text{subject to } g_i(x, \alpha) = 0,\ i \in \{1, \cdots, m_1\} \cup I_+(x^0, \alpha^0), \\ &\qquad g_i(x, \alpha) \leqq 0,\ i \in I_0(x^0, \alpha^0), \\ &\qquad x \in U. \end{aligned}$$

Clearly $\tilde{S}(\alpha)$ is a problem of type $P(\alpha)$ with $Q = \{0\}^p x (-\mathbb{R}_+)^q$, where $p = \text{card } I_+(x^0, \alpha^0) + m_1$ and $q = \text{card } I_0(x^0, \alpha^0)$. If we let $\tilde{\lambda}^0$ be the vector in $\mathbb{R}^{p+q}$ with coordinates $\tilde{\lambda}_i^0 = \lambda_i^0$, for $i \in \{1, \cdots, m_1\} \cup I(x^0, \alpha^0)$, then $(x^0, \alpha^0, \tilde{\lambda}^0)$ satisfies the assumptions of Theorem 2.1. Furthermore, $(x^0, \alpha^0, \tilde{\lambda}^0)$ satisfies the necessary condition (2.2) associated with $\tilde{S}(\alpha^0)$, since $(x^0, \alpha^0, \lambda^0)$ satisfies the necessary condition (2.2) associated with $S(\alpha^0)$. Consequently, from Theorem 2.1, there exist open neighborhoods U_1 of x^0 in U, V_1 of α^0 in P and Lipschitzian mappings $x(\cdot): V_1 \to U_1$ and $\tilde{\lambda}(\cdot): V_1 \to \mathbb{R}^{p+q}$ such that $x(\alpha^0) = x^0$, $\tilde{\lambda}(\alpha^0) = \tilde{\lambda}^0$ and, for all $\alpha \in V_1$, the pair $(x(\alpha), \tilde{\lambda}(\alpha))$ satisfies the necessary condition (2.2) associated with problem $\tilde{S}(\alpha)$. We let now $\lambda(\cdot): V_1 \to \mathbb{R}^m$ be the mapping defined by $\lambda_i(\alpha) = \tilde{\lambda}_i(\alpha)$ for $i \in \{1, \cdots, m_1\} \cup I(x^0, \alpha^0)$ and $\lambda_i(\alpha) = 0$ for $i \in J(x^0, \alpha^0)$. Since, for all $i \in I_+(x^0, \alpha^0)$, $\lambda_i(\alpha^0) = \lambda_i^0 > 0$ and, for all $i \in J(x^0, \alpha^0)$, $g_i(x^0, \alpha^0) = g_i(x(\alpha^0), \alpha^0) < 0$, from the continuity of the mappings $\lambda(\cdot)$, $x(\cdot)$ and $g_i(\cdot, \cdot)$, there exists an open neighborhood V_2 of α^0 in V_1 such that, for all α in V_2, for all $i \in I_+(x^0, \alpha^0)$, $\lambda_i(\alpha) > 0$ and, for all $i \in J(x^0, \alpha^0)$, $g_i(x(\alpha), \alpha) < 0$. Consequently, one easily checks that, for all $\alpha \in V_2$, $(x(\alpha), \lambda(\alpha))$ satisfies the necessary condition (2.2) associated with problem $S(\alpha)$.

It now remains to show that there exist neighborhoods V' of α^0 in V_2 and U' of x^0 in U_1 such that, for all $\alpha \in V'$ and all $x \in U'$, $x \neq x(\alpha)$, satisfying $g_i(x, \alpha) = 0$ for $i \in \{1, \cdots, m_1\}$ and, $g_i(x, \alpha) \leqq 0$, for $i \in \{m_1 + 1, \cdots, m\}$, then one has $f(x, \alpha) > f(x(\alpha), \alpha)$. We prove this assertion by contraposition. Assume that there exist sequences $\{x^k\} \subset U'$ and $\{\alpha^k\} \subset V'$ such that $\{x^k\} \to x^0$, $\{\alpha^k\} \to \alpha^0$ and, for all k, $x^k \neq x(\alpha^k)$, $g_i(x^k, \alpha^k) = 0$ for $i \in \{1, \cdots, m_1\}$, $g_i(x^k, \alpha^k) \leqq 0$ for $i \in \{m_1 + 1, \cdots, m\}$ and $f(x^k, \alpha^k) \leqq f(x(\alpha^k), \alpha^k)$. Without any loss of generality, one can assume that $(x^k - x(\alpha^k))/\|x^k - x(\alpha^k)\|$ converges to some element h in $\mathbb{R}^n$ such that $\|h\| = 1$.

From the mean value theorem and the continuity of $\nabla f(\cdot, \cdot)$ and $\nabla g_i(\cdot, \cdot)$ at (x^0, α^0) one has, for $i = 1, \cdots, m$,

$$\begin{aligned} \langle \nabla f(x^0, \alpha^0), h \rangle &= \lim_{k \to \infty} [f(x^k, \alpha^k) - f(x(\alpha^k), \alpha^k)]/\|x^k - x(\alpha^k)\|, \\ \langle \nabla g_i(x^0, \alpha^0), h \rangle &= \lim_{k \to \infty} [g_i(x^k, \alpha^k) - g_i(x(\alpha^k), \alpha^k)]/\|x^k - x(\alpha^k)\|. \end{aligned} \tag{3.10}$$

From the very definition of the sequences $\{x^k\}$ and $\{\alpha^k\}$ and recalling that from the first part of the proof, $g_i(x(\alpha^k), \alpha^k) = 0$, $i = \{1, \cdots, m_1\} \cup I_+(x^0, \alpha^0)$, one deduces from (3.10) that

$$
\begin{aligned}
&\langle \nabla f(x^0, \alpha^0), h\rangle \leqq 0,\\
&\langle \nabla g_i(x^0, \alpha^0), h\rangle = 0, \qquad i = 1, \cdots, m_1,\\
&\langle \nabla g_i(x^0, \alpha^0), h\rangle \leqq 0, \qquad i \in I_+(x^0, \alpha^0).
\end{aligned} \tag{3.11}
$$

Since (x^0, λ^0) satisfies the necessary condition (2.2) associated with $S(\alpha^0)$, one gets

$$
\langle \nabla f(x^0, \alpha^0), h\rangle + \sum_{i \in \{1, \cdots, m_1\} \cup I_+(x^0, \alpha^0)} \lambda_i^0 \langle \nabla g_i(x^0, \alpha^0), h\rangle = 0. \tag{3.12}
$$

From (3.11) and (3.12), since $\lambda_i^0 > 0$, for $i \in I_+(x^0, \alpha^0)$, it follows that the inequalities in (3.11) are in fact equalities. Hence, by (C.2) one gets

$$
\left\langle D^2 f(x^0, \alpha^0)h + \sum_{i=1}^{m} \lambda_i^0 D^2 g_i(x^0, \alpha^0)h, h \right\rangle > 0. \tag{3.13}
$$

We end the proof of the corollary by contradicting (3.13). Denote $L_k(x) = f(x, \alpha^k) + \sum_{i=1}^m \lambda_i(\alpha^k) g_i(x, \alpha^k)$. From Taylor's theorem and from the continuity of $D^2 f(\cdot, \cdot)$ and $D^2 g_i(\cdot, \cdot)$ at (x^0, α^0):

$$
\begin{aligned}
&\frac{1}{2}\left\langle D^2 f(x^0, \alpha^0)h + \sum_{i=1}^{m} \lambda_i^0 D^2 g_i(x^0, \alpha^0)h, h \right\rangle\\
&= \lim_{k \to \infty} \int_0^1 (1-t)\left\langle D^2 L_k[x(\alpha^k) + t(x^k - x(\alpha^k))] \frac{(x^k - x(\alpha^k))}{\|x^k - x(\alpha^k)\|}, \frac{x^k - x(\alpha^k)}{\|x^k - x(\alpha^k)\|} \right\rangle dt\\
&= \lim_{k \to \infty} [L_k(x^k) - L_k(x(\alpha^k)) - \langle \nabla L_k(x(\alpha^k)), x^k - x(\alpha^k)\rangle] / \|x^k - x(\alpha^k)\|^2,
\end{aligned}
$$

where

$$
h = \lim \frac{x^k - x(\alpha^k)}{\|x^k - x(\alpha^k)\|}.
$$

Since $(x(\alpha^k), \lambda(\alpha^k))$ satisfies the necessary conditions (2.2) associated with $S(\alpha^k)$, then $\nabla L_k(x(\alpha^k)) = 0$, $\lambda_i(\alpha^k) g_i(x(\alpha^k), \alpha^k) = 0$, for all $i \in \{1, \cdots, m\}$ and $\lambda_i(\alpha^k) \geqq 0$, for $i \in \{m_1 + 1, \cdots, m\}$. Consequently, $L_k(x(\alpha^k)) = f(x(\alpha^k))$ and $L_k(x^k) = f(x^k) + \sum_{i=1}^m \lambda_i(\alpha^k) g_i(x^k, \alpha^k) \leqq f(x^k)$. Hence, for all k,

$$
L_k(x^k) - L_k(x(\alpha^k)) - \langle \nabla L_k(x(\alpha^k)), x^k - x(\alpha^k)\rangle \leqq 0.
$$

Dividing the above inequality by $\|x^k - x(\alpha^k)\|^2$, passing to the limit, when $k \to \infty$, from the above equalities one deduces that:

$$
\left\langle D^2 f(x^0, \alpha^0)h + \sum_{i=1}^{m} \lambda_i^0 D^2 g_i(x^0, \alpha^0)h, h \right\rangle \leqq 0
$$

which contradicts (3.13). This ends the proof of the corollary.

REFERENCES

[1] J. P. AUBIN, *Lipschitz behaviour of solutions to convex minimization problems*, Working Paper, IASA, 1981.

[2] A. AUSLENDER, *Theorem of constant rank for Lipschitzian maps and applications in optimization theory*, in Mathematical Programming with Data Perturbations, Vol. 2, A. V. Fiacco, ed., Marcel Dekker, New York, 1983.

[3] F. CLARKE, *Generalized gradients and applications*, Trans. Amer. Math. Soc., 205 (1975), pp. 247-262.
[4] ———, *On the inverse function theorem*, Pacific J. Math., 64 (1976), pp. 97-102.
[5] B. CORNET AND G. LAROQUE, *Lipschitz properties of constrained demand functions and constrained maximizers*, INSEE Working Paper 8005, 1980.
[6] B. CORNET, *Contributions à la théorie des mécanismes dynamiques d'allocation des ressources*, thèse d'état, Université Paris IX Dauphine, 1981.
[7] B. CORNET AND G. LAROQUE, *Lipschitz properties of solutions in mathematical programming*, J. Optim. Theory Appl., 1986, to appear.
[8] G. DEBREU, *Definite and semi-definite quadratic forms*, Econometrica, 20 (1952), pp. 295-299.
[9] A. V. FIACCO AND G. P. MCCORMICK, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley, New York, 1968.
[10] A. V. FIACCO, *Sensitivity analysis for nonlinear programming using penalty methods*, Math. Programming, 10 (1976), pp. 287-311.
[11] A. V. FIACCO AND W. P. HUTZLER, *Basic results in the development of sensitivity and stability analysis in nonlinear programming*, Comput. Oper. Res., 9, 1 (1982), pp. 9-28.
[12] J. B. HIRIART-URRUTY, *Tangent cones, generalized gradients and mathematical programming in Banach spaces*, Math. Oper. Res., 4, 1 (1979), pp. 79-97.
[13] K. JITTORNTRUM, *Solution point differentiability without strict complementarity in nonlinear programming*, Math. Programming, 21 (1984), pp. 127-138.
[14] M. KOJIMA, *Strongly stable stationary solutions in nonlinear programs*, in Analysis and Computation of Fixed Points, S. M. Robinson, ed., Academic Press, New York, 1980, pp. 93-138.
[15] E. S. LEVITIN, *On the local perturbation theory of a problem of mathematical programming in a Banach space*, Soviet Math. Dokl., 15 (1975), pp. 603-608.
[16] S. M. ROBINSON, *Perturbed Kuhn-Tucker points and rates of convergence for a class of nonlinear programming algorithms*, Math. Programming, 7 (1974), pp. 1-16.
[17] ———, *Strongly regular generalized equations*, Math. Oper. Res., 5, 1 (1980), pp. 43-62.
[18] ———, *Generalized equations and their solution, Part* II: *Applications to nonlinear programming*, in Optimality and Stability in Mathematical Programming, M. Guignard, ed., Mathematical Programming Study, 19 (1982).
[19] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton Univ. Press, Princeton, NJ, 1970.
[20] ———, *The Theory of Subgradients and its Applications to Problems of Optimization, Convex and Nonconvex Functions*, Helderman-Verlag, Berlin, 1981.
[21] L. SCHWARTZ, *Analyse mathématique*, Cours de l'Ecole Polytechnique, Hermann, Paris, 1967.
[22] J. P. VIAL, *Strong and weak convexity of sets and functions*, Math. Oper. Res., 8 (1983), pp. 231-259.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 6, November 1986

004

ANALYSIS OF THE PERIODIC LYAPUNOV AND RICCATI EQUATIONS VIA CANONICAL DECOMPOSITION*

SERGIO BITTANTI†, PATRIZIO COLANERI‡ AND GUIDO GUARDABASSI†

Abstract. A new proof is given of the following result: a periodic Riccati equation admits a unique positive semidefinite periodic solution and the solution is stabilizing if and only if the underlying system is stabilizable and detectable. The proof hinges on the decomposition of the Riccati equation induced by the system canonical decomposition. The solution structure is therefore naturally pointed out.

Key words. periodic Riccati equation, periodic Lyapunov equation, periodic solutions, canonical decomposition

AMS(MOS) subject classifications. 93B10, 93G50, 93G60, 93G15, 43E30

1. Introduction. As is well known, the necessary and sufficient condition for the existence of a positive semidefinite solution of the time-invariant Riccati equation is stated in terms of stabilizability and detectability of the underlying (time-invariant) linear system. This fundamental fact was first proved by Wonham via quasi-linearization techniques [1]. An important alternative proof, based on the analysis of the associated Hamiltonian matrix, is due to Kucera [2]. The possibility of extending the above result to the periodically time-varying case, in order to solve many important periodic estimation and control problems, was considered by Hewer in [3] and by Kano and Nishimura in [4]. Precisely, Hewer adopted the quasi-linearization approach while Kucera's line of reasoning was followed by Kano and Nishimura. However, some preliminary yet basic results of [3] turned out not to be technically sound [5]. In particular, the very notions of stabilizability and detectability for linear periodic systems, introduced by Hewer in [3] and set then by Kano and Nishimura at the basis of their development, were not clearly related either to their original definitions or to the asymptotic stability of appropriate parts of the Kalman canonical decomposition (if any) of the periodic system at hand. These various issues have only recently found a satisfactory clarification [6]-[8], so that, in retrospective, Kano and Nishimura's paper can now be looked at as a valid extension to the periodic case of the basic Wonham-Kucera theorem recalled at the beginning of this introduction.

The main purpose of this paper is to provide, for the periodically time-varying case, a totally new and alternative proof of the Wonham-Kucera theorem. Our proof is based on the canonical decomposition of the periodic Riccati equation induced by the Kalman canonical decomposition of the underlying periodic dynamical system. Besides admitting an easy yet fairly deep system-theoretic interpretation, this proof does not require any simplifying assumption on the eigenvalues of the monodromy matrix of the associated Hamiltonian system. Last but not least, our approach, when specialized to the time-invariant case, does obviously provide a third and novel proof of the Wonham-Kucera theorem.

The paper is organized as follows. Preliminarily, in § 2 we study the action on the periodic Riccati equation of a state-space nonsingular transformation performed on

* Received by the editors November 14, 1984; accepted for publication (in revised form) August 23, 1985. This work was supported by the Centro di Teoria dei Sistemi of the CNR (Milano) and by M.P.I.

† Dipartimento di Elettronica, Politecnico di Milano, 20133 Milano, Italy.

‡ Centro di Teoria dei Sistemi of the CNR, c/o Dipartimento di Elettronica, Politecnico di Milano, 20133 Milano, Italy.

the underlying periodic dynamical system. In particular, the underlying dynamical system can, without any loss of generality, be assumed to be in the Kalman canonical form. A corresponding decomposition of the periodic Riccati equation is then obtained. The existence of a periodic solution for a periodic Lyapunov differential equation is discussed in § 3. The so-obtained results, together with the canonical decomposition of the periodic Riccati equation introduced in § 2, are used in § 4 to analyse the structure of periodic solutions to the periodic Riccati equation and to prove the main theorem.

2. Decomposition of the periodic matrix Riccati equation. Consider the linear system:

$$\dot{x}(t) = A(t)x(t) + B(t)u(t), \tag{1a}$$

$$y(t) = C(t)x(t) \tag{1b}$$

where $A: R \to R^{n\times n}$, $B: R \to R^{n\times m}$, $C: R \to R^{p\times n}$ are continuous functions. We assume that system (1) is T-periodic, namely:

$$A(t+T) = A(t), \quad B(t+T) = B(t), \quad C(t+T) = C(t) \quad \forall t.$$

We shall denote by $\Phi_A(t, \tau)$ the transition matrix (generated by A) of system (1).

According to most of the literature on linear periodic systems [9]-[11], $\Phi_A(T, 0)$ is referred to in the sequel as the *monodromy matrix* (of system (1)), while its eigenvalues are said to be the *characteristic multipliers* of A. System (1) is asymptotically stable if and only if the characteristic multipliers of A lie inside the open unit disk. In such a case, we also say, for short, that A (instead of system (1)) is asymptotically stable.

Associated with (1), let

$$-\dot{P}(t) = A(t)'P(t) + P(t)A(t) + C(t)'C(t) - P(t)B(t)B(t)'P(t) \tag{2}$$

be a T-periodic matrix Riccati differential equation the periodic solutions of which are the principal object of the present paper.

The first part of this section is devoted to showing the action on (2) of a nonsingular, continuously differentiable, T-periodic state-space transformation of (1). Specifically let

$$\hat{x}(t) = S(t)x(t) \tag{3}$$

where S is continuously differentiable, $S(t)$ is nonsingular and equal to $S(t+T)$, for all t. It is well known that (3) carries system (1) into

$$\dot{\hat{x}}(t) = \hat{A}(t)\hat{x}(t) + \hat{B}(t)u(t), \qquad y(t) = \hat{C}(t)\hat{x}(t) \tag{4}$$

with

$$\hat{A}(t) = S(t)A(t)S(t)^{-1} + \dot{S}(t)S(t)^{-1},$$
$$\hat{B}(t) = S(t)B(t),$$
$$\hat{C}(t) = C(t)S(t)^{-1}.$$

Being A, B, C continuous and S continuously differentiable, $\hat{A}$, $\hat{B}$, $\hat{C}$ are also continuous. Moreover, the periodicity of A, B, C and S entails the periodicity of $\hat{A}$, $\hat{B}$, $\hat{C}$. Therefore, the Riccati differential equation

$$-\dot{\hat{P}}(t) = \hat{A}(t)'\hat{P}(t) + \hat{P}(t)\hat{A}(t) + \hat{C}(t)'\hat{C}(t) - \hat{P}(t)\hat{B}(t)\hat{B}(t)'\hat{P}(t), \tag{5}$$

associated with system (4), is T-periodic as well. Now, the question is what are the relationships between the T-periodic solutions of (5) and the T-periodic solutions of (2). The answer is given by the following theorem, the proof of which is omitted, as the theorem is straightforward.

THEOREM 1. *$\hat{P}$ is a T-periodic solution to* (5) *if and only if $P = S'\hat{P}S$ is a T-periodic solution to* (2).

COROLLARY 1. *Since S is nonsingular, $P = S'\hat{P}S$ establishes a bijective correspondence between the positive semidefinite T-periodic solutions of* (2) *and the positive semidefinite T-periodic solution of* (5).

In view of Corollary 1, the existence of positive semi-definite T-periodic solutions of (2) can be investigated by looking at the equivalent S-transformed equation (5). In particular, reference can usefully be made to any state transformation S which brings system (1) into a canonical form. In [12] it is proven that, for any periodic system with continuous coefficients, a periodic state-space transformation does exist yielding a system in the Kalman canonical form.

In conclusion, there is no loss of generality in assuming that system (1) is in canonical form, i.e. that matrices A, B and C conform to the following zero-nonzero pattern:

$$A(t) = \begin{bmatrix} A_0(t) & 0 \\ \tilde{A}_0(t) & \bar{A}_0(t) \end{bmatrix} = \left[\begin{array}{cc:cc} A_{11}(t) & A_{12}(t) & 0 & 0 \\ 0 & A_{22}(t) & 0 & 0 \\ \hdashline A_{31}(t) & A_{32}(t) & A_{33}(t) & A_{34}(t) \\ 0 & A_{42}(t) & 0 & A_{44}(t) \end{array}\right], \tag{6a}$$

$$B(t) = \begin{bmatrix} B_0(t) \\ \bar{B}_0(t) \end{bmatrix} = \left[\begin{array}{c} B_{11}(t) \\ 0 \\ \hdashline B_{33}(t) \\ 0 \end{array}\right], \tag{6b}$$

$$C(t) = [C_0(t) \quad 0] = [C_{11}(t) \quad C_{22}(t) \quad 0 \quad 0]. \tag{6c}$$

Of course, the pair (A_0, C_0) is completely observable while the pair (A_r, B_r), where

$$A_r(t) = \begin{bmatrix} A_{11}(t) & 0 \\ A_{31}(t) & A_{33}(t) \end{bmatrix}, \qquad B_r(t) = \begin{bmatrix} B_{11}(t) \\ B_{33}(t) \end{bmatrix},$$

is completely reachable. For more information on the structural properties of periodic systems, the interested reader is referred to [7].

In the following we analyze the partition of the generic symmetric solution of (2) induced by the Kalman canonical structure of (1). With reference to the system decomposition into the observable and nonobservable parts, any symmetric solution of (2) can be given the form:

$$P(t) = \begin{bmatrix} P_0(t) & \tilde{P}_0(t) \\ \tilde{P}_0(t)' & \bar{P}_0(t) \end{bmatrix}, \tag{7}$$

where $P_0(t)$ and $A_0(t)$ ($\bar{P}_0(t)$ and $\bar{A}_0(t)$) are square matrices of the same dimensions. The matrix $P_0(t)$ will henceforth be referred to as the *observability submatrix* of $P(t)$. Furthermore, referring to (6), let

$$P(t) = [P_{ij}(t)]_{i,j=1,2,3,4}, \qquad P_{ij}(t) = P_{ji}(t)' \tag{8}$$

where, for any $i = 1, 2, 3, 4$, $P_{ii}(t)$ and $A_{ii}(t)$ are square matrices of the same dimensions. Then it should be apparent that

$$P_0(t) = \begin{bmatrix} P_{11}(t) & P_{12}(t) \\ P_{12}(t)' & P_{22}(t) \end{bmatrix}, \qquad \bar{P}_0(t) = \begin{bmatrix} P_{33}(t) & P_{34}(t) \\ P_{34}(t)' & P_{44}(t) \end{bmatrix}, \tag{9a}$$

$$\tilde{P}_0(t) = \begin{bmatrix} P_{13}(t) & P_{14}(t) \\ P_{23}(t) & P_{24}(t) \end{bmatrix}. \tag{9b}$$

In view of this *two-level canonical decomposition* of P, a corresponding two-level decomposition of (2) can be obtained.

The coarser decomposition is as follows:

$$\begin{aligned}
-\dot{P}_0(t) &= P_0(t)A_0(t) + \tilde{P}_0(t)\tilde{A}_0(t) + A_0(t)'P_0(t) + \tilde{A}_0(t)'\tilde{P}_0(t)' + C_0(t)'C_0(t) \\
&\quad - P_0(t)B_0(t)B_0(t)'P_0(t) - P_0(t)B_0(t)\bar{B}_0(t)'\tilde{P}_0(t)' \\
&\quad - \tilde{P}_0(t)\bar{B}_0(t)B_0(t)'P_0(t) - \tilde{P}_0(t)\bar{B}_0(t)\bar{B}_0(t)'\tilde{P}_0(t)', \qquad (10a)\\
-\dot{\tilde{P}}_0(t) &= \tilde{P}_0(t)\bar{A}_0(t) + A_0(t)'\tilde{P}_0(t) + \tilde{A}_0(t)'\bar{P}_0(t) \\
&\quad - P_0(t)B_0(t)B_0(t)'\tilde{P}_0(t) - P_0(t)B_0(t)\bar{B}_0(t)'\bar{P}_0(t)' \\
&\quad - P_0(t)\bar{B}_0(t)B_0(t)'\tilde{P}_0(t) - \tilde{P}_0(t)\bar{B}_0(t)\bar{B}_0(t)'\bar{P}_0(t), \qquad (10b)\\
-\dot{\bar{P}}_0(t) &= \bar{P}_0(t)\bar{A}_0(t) + \bar{A}_0(t)\bar{P}_0(t)' \\
&\quad - \tilde{P}_0(t)'B_0(t)B_0(t)'\tilde{P}_0(t) - \tilde{P}_0(t)'B_0(t)\bar{B}_0(t)'\bar{P}_0(t) \\
&\quad - P_0(t)\bar{B}_0(t)B_0(t)'\tilde{P}_0(t) - \bar{P}_0(t)\bar{B}_0(t)\bar{B}_0(t)'\bar{P}_0(t). \qquad (10c)
\end{aligned}$$

As for the finer decomposition, we shall consider only the subset of equations which corresponds to the elements of $P_0(t)$; namely:

$$\begin{aligned}
-\dot{P}_{11}(t) &= P_{11}(t)A_{11}(t) + A_{11}(t)'P_{11}(t) + P_{13}(t)A_{31}(t) \\
&\quad + A_{31}(t)'P_{13}(t)' + C_{11}(t)'C_{11}(t) \\
&\quad - P_{11}(t)B_{11}(t)B_{11}(t)'P_{11}(t) - P_{11}(t)B_{11}(t)B_{33}(t)'P_{13}(t)' \\
&\quad - P_{13}(t)B_{33}(t)B_{11}(t)'P_{11}(t) - P_{13}(t)B_{33}(t)B_{33}(t)'P_{13}(t)', \qquad (11a)\\
-\dot{P}_{12}(t) &= P_{11}(t)A_{12}(t) + P_{12}(t)A_{22}(t) + A_{11}(t)'P_{12}(t) + P_{13}(t)A_{32}(t) \\
&\quad + P_{14}(t)A_{42}(t) + A_{31}(t)'P_{23}(t)' + C_{11}(t)'C_{22}(t) \\
&\quad - P_{11}(t)B_{11}(t)B_{11}(t)'P_{12}(t) - P_{11}(t)B_{11}(t)B_{33}(t)'P_{23}(t)' \\
&\quad - P_{13}(t)B_{33}(t)B_{11}(t)'P_{12}(t) - P_{13}(t)B_{33}(t)B_{33}(t)'P_{23}(t)', \qquad (11b)\\
-\dot{P}_{22}(t) &= P_{12}(t)'A_{12}(t) + P_{22}(t)A_{22}(t) + A_{12}(t)'P_{12}(t) + A_{22}(t)'P_{22}(t) \\
&\quad + P_{23}(t)A_{32}(t) + P_{24}(t)A_{42}(t) + A_{32}(t)'P_{23}(t) \\
&\quad + A_{42}(t)'P_{24}(t) + C_{22}(t)'C_{22}(t) \\
&\quad - P_{12}(t)'B_{11}(t)B_{11}(t)'P_{12}(t) - P_{12}(t)'B_{11}(t)B_{33}(t)'P_{23}(t)' \\
&\quad - P_{23}(t)B_{33}(t)B_{11}(t)'P_{12}(t) - P_{23}(t)B_{33}(t)B_{33}(t)'P_{23}(t)'. \qquad (11c)
\end{aligned}$$

3. Periodic solutions of the periodic Sylvester matrix differential equation. This section deals with the T-periodic solutions of a T-periodic matrix differential equation of the form

$$-\dot{Q}(t) = M(t)Q(t) + Q(t)N(t)' + W(t) \qquad (12)$$

where $M: R \to R^{h\times h}$, $N: R \to R^{k\times k}$ and $W: R \to R^{h\times k}$ are given continuous and T-periodic functions. When $h = k$ and $M = N$, (12) is a Lyapunov differential equation. Besides its independent interest, the result stated below as Theorem 2 plays a crucial role in the next section, where the existence of periodic solutions to the periodic matrix Riccati differential equation are thoroughly investigated.

THEOREM 2. *If M and N are asymptotically stable, then*

i) *equation* (12) *admits a unique T-periodic solution*;

ii) *under the additional assumption $h = k$, $M = N$, $W(t) = W(t)' \leqq 0$ for all t and $W(t) \neq 0$ for some t,* (12) *does not admit any positive semidefinite T-periodic solution.*

Proof of (i). For any matrix H, define vec (H) as the vector obtained by composing in a single column all columns of H taken in their natural ordering; furthermore, let $q(t) = \text{vec}\,(Q(t))$, $w(t) = \text{vec}\,(W(t))$.

It is well known [13] that (12) can be rewritten as follows:

$$-\dot{q}(t) = (I_k \otimes M(t) + N(t) \otimes I_h) q(t) + W(t), \tag{13}$$

where $\otimes$ denotes the Kronecker product and I_j is the $j \times j$ identity matrix. First, we show that $\Phi_N(t, 0) \otimes \phi_M(t, 0)$ is the transition matrix of system (13). In fact, $\Phi_N(0, 0) \otimes \phi_M(0, 0) = I_k \otimes I_h = I_{hk}$ and, also in view of the "mixed product rule" [13, p. 24],

$$\begin{aligned}
\frac{d}{dt}(\Phi_N(t, 0) \otimes \Phi_M(t, 0)) &= \Phi_N(t, 0) \otimes \dot{\Phi}_M(t, 0) + \dot{\Phi}_N(t, 0) \otimes \Phi_M(t, 0) \\
&= \phi_N(t, 0) \otimes (M(t)\Phi_M(t, 0)) + (N(t)\Phi_N(t)) \otimes \Phi_M(t, 0) \\
&= (I_k \otimes M(t))(\Phi_N(t, 0) \otimes \Phi_M(t, 0)) \\
&\quad + (N(t) \otimes I_h)(\Phi_N(t, 0) \otimes \Phi_M(t, 0)) \\
&= (I_k \otimes M(t) + N(t) \otimes I_h)(\Phi_N(t, 0) \otimes \Phi_M(t, 0)).
\end{aligned}$$

As for the monodromy matrix of system (13), recall that the eigenvalues of $\Phi_M(T, 0)$ and $\phi_N(T, 0)$ lie, by assumption, in the open unit disk. Since the eigenvalues of the Kronecker product are given by the product of any pair of eigenvalues of the two factors, the conclusion can readily be drawn that system (13) is asymptotically stable. Hence, the T-periodic solution forced by W exists and is unique [10].

Proof of (ii). The line of reasoning adopted here is similar to the one developed in [14] to extend the so-called Lyapunov Lemma to the periodically time-varying case.

For any $V: R \to R^{h \times h}$ such that $W(t) = -V(t)V(t)'$, for all t, consider the Kalman canonical decomposition of the pair (N, V) into the reachable and unreachable parts:

$$N(t) = \begin{bmatrix} N_r(t) & \tilde{N}_r(t) \\ 0 & \bar{N}_r(t) \end{bmatrix}, \qquad V(t) = \begin{bmatrix} V_r(t) \\ 0 \end{bmatrix}.$$

Any possible solution of (12) can of course be partitioned accordingly, i.e.:

$$Q(t) = \begin{bmatrix} Q_r(t) & \tilde{Q}_r(t) \\ \tilde{Q}_r(t)' & \bar{Q}_r(t) \end{bmatrix}.$$

Thus (12) splits into:

$$\begin{aligned}
-\dot{Q}_r(t) &= N_r(t)Q_r(t) + Q_r(t)N_r(t)' - V_r(t)V_r(t)' + \tilde{N}_r(t)\tilde{Q}_r(t) + \tilde{Q}_r(t)\tilde{N}_r(t)', \\
-\dot{\tilde{Q}}_r(t) &= N_r(t)\tilde{Q}_r(t) + \tilde{Q}_r(t)\bar{N}_r(t)' + \tilde{N}_r(t)\bar{Q}_r(t), \\
-\dot{\bar{Q}}_r(t) &= \bar{N}_r(t)\bar{Q}_r(t) + \bar{Q}_r(t)\bar{N}_r(t)'.
\end{aligned}$$

By applying part (i) of this theorem to the last equation and recalling that N (whereby N_r and $\bar{N}_r$) is by assumption asymptotically stable, the conclusion can be drawn that the unique positive semidefinite T-periodic solution is $\bar{Q}_r = 0$. The same argument applied to the second equation leads to recognizing that the unique positive semidefinite T-periodic solution of (12), if any, must be of the form:

$$Q(t) = \begin{bmatrix} Q_r(t) & 0 \\ 0 & 0 \end{bmatrix},$$

where $Q_r(t)$ is a positive semidefinite T-periodic solution of

$$-\dot{Q}_r(t) = N_r(t)Q_r(t) + Q_r(t)N_r(t)' - V_r(t)V_r(t)'.$$

The remainder of the proof is meant to show that assuming the existence of such a solution leads to an unavoidable contradiction. In fact, let

$$G = \int_0^T \Phi_{N_r}(t,0)\,V(t)\,V(t)'\Phi_{N_r}(t,0)'\,dt.$$

By a well-known resolution formula for the Lyapunov equation [15], one gets

$$G = \Phi_{N_r}(T,0)\,Q_r(0)\Phi_{N_r}(T,0)' - Q_r(0).$$

Let now λ be an eigenvalue and z an associated nonzero eigenvector of $\Phi_{N_r}(T,0)$. Denoting by a star the conjugate transpose, elementary manipulations yield $z^*Gz = (|\lambda|^2-1)z^*Q_r(0)z$.

Since, by assumption, $|\lambda|<1$ (N_r is asymptotically stable) and a positive semidefinite Q_r exists, the corresponding value of the right-hand side must be nonpositive. As for the left-hand side, recall that the pair (N_r, V_r) is completely reachable (or, equivalently, controllable) only if $V_r(t)'\Phi_{N_r}(t,0)'z = 0$ a.e. in $[0, T]$ implies $z = 0$ [7].

Since $z^*Gz = \int_0^T \| V_r(t)'\Phi_{N_r}(t,0)'z\|^2\,dt$, by complete reachability the conclusion must be drawn that $z^*Gz > 0$. The contradiction completes the proof of (ii).

4. Periodic solutions of the periodic matrix Riccati equation. In this section, necessary and sufficient conditions for the existence of a unique positive semidefinite T-periodic solution of (2) are obtained in terms of stabilizability and detectability of system (1). It has recently been proved [12], that several different yet equivalent characterizations exist of stabilizability and detectability for linear periodic systems. Exactly as in the time-invariant case [1] one of the most interesting characterizations refers to the Kalman canonical decomposition.

Precisely [12], a T-periodic continuous matrix function $K: R \to R^{m\times n}$ such that $A+BK$ is asymptotically stable exists if and only if the uncontrollable part of system (1) is asymptotically stable (*stabilizability*).

Dually, a T-periodic continuous matrix function $L: R \to R^{n\times p}$ such that $A-LC$ is asymptotically stable exists (whereby an asymptotic periodic state-reconstructor can actually be designed [12]) if and only if the unobservable part of system (1) is asymptotically stable (*detectability*). The main results of this section are visualized in Fig. 1. They can be stated as follows.

THEOREM 3. *If the unreachable and observable part of system* (1) *is asymptotically stable, then the Riccati differential equation* (2) *admits a positive semidefinite T-periodic solution* $P: R \to R^{n\times n}$. *If, in addition, system* (1) *is in Kalman canonical form, then, only the observability sub-matrix* P_0 *of this solution is different from zero and* $P_0(t)$ *is positive definite for all t.*

THEOREM 4. *The periodic Riccati differential equation* (2) *admits a unique positive semidefinite periodic solution* $P: R \to R^{n\times n}$ *and* $A-BB'P$ *is asymptotically stable if and only if system* (1) *is stabilizable and detectable.*

As is shown in Fig. 1, two preliminary lemmas (Lemmas 2 and 3) are needed, in addition to Theorem 3, to prove Theorem 4. On the other hand, to prove Lemma 2, one more lemma (Lemma 1) is needed which in turn stands on one of the Shayman "Inertia Theorems" [16]. The proof of Theorem 3 appeals to the same Inertia Theorem mentioned above, to Theorem 2 and to a third result, proved in [17]. The Shayman Inertia Theorem used in the sequel is reported below, without proof, for easy reference.

INERTIA THEOREM. *If system* (1) *is completely observable and the matrix Riccati differential equation* (2) *admits a symmetric T-periodic solution* $P: R \to R^{n\times n}$, *then, for each t,* $P(t)$ *is nonsingular and the number of positive eigenvalues of* $P(t)$ *is equal to the number of characteristic multipliers of* $A-BB'P$ *lying in the open unit disk.*

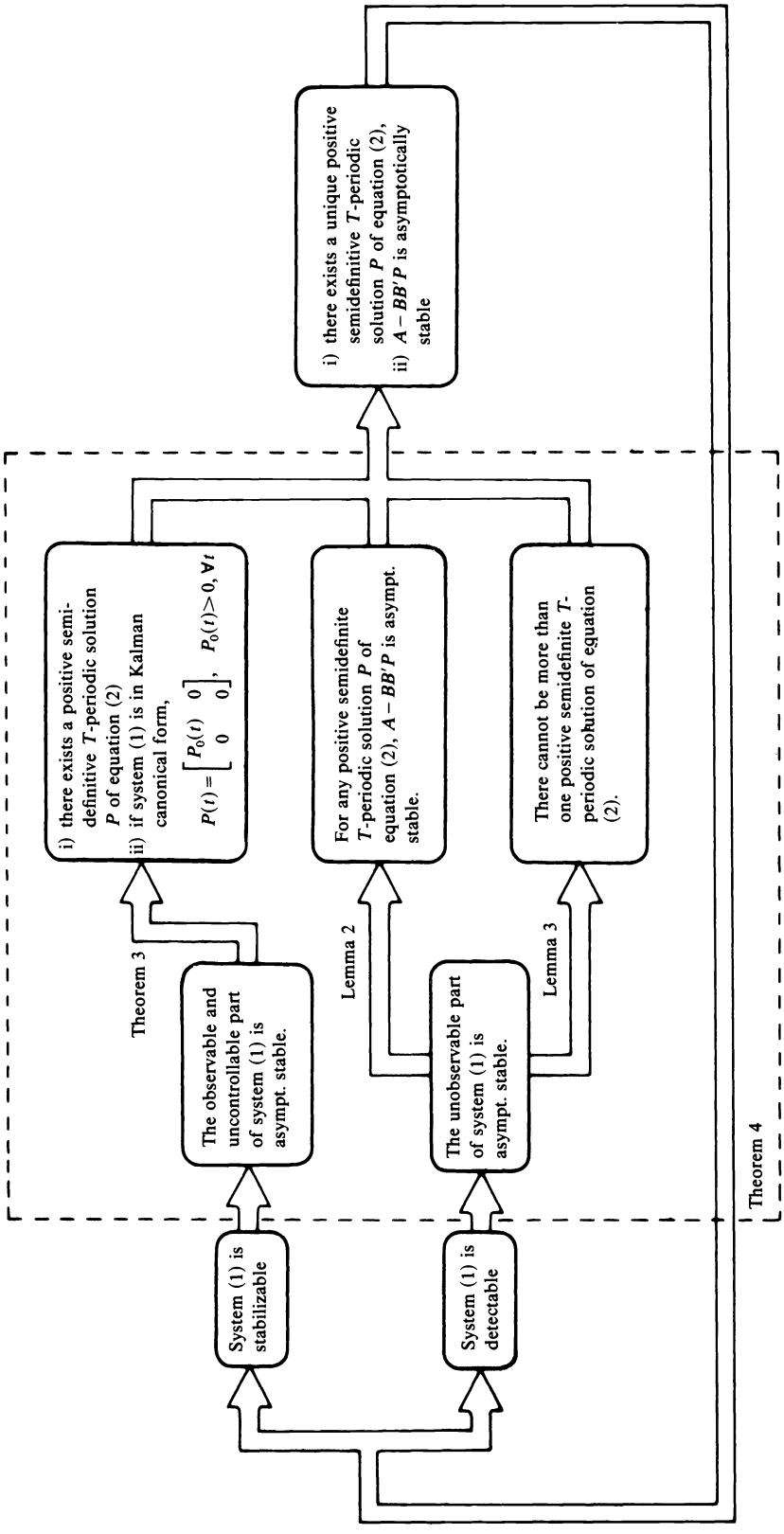

System (1) is stabilizable
System (1) is detectable
The observable and uncontrollable part of system (1) is asympt. stable.
The unobservable part of system (1) is asympt. stable.
Theorem 3
Lemma 2
Lemma 3
Theorem 4
i) there exists a positive semi-definitive T-periodic solution P of equation (2)
ii) if system (1) is in Kalman canonical form,
$P(t) = \begin{bmatrix} P_0(t) & 0 \\ 0 & 0 \end{bmatrix}$, $P_0(t) > 0, \forall t$
For any positive semidefinite T-periodic solution P of equation (2), $A - BB'P$ is asympt. stable.
There cannot be more than one positive semidefinite T-periodic solution of equation (2).
i) there exists a unique positive semidefinitive T-periodic solution P of equation (2),
ii) $A - BB'P$ is asymptotically stable

FIG. 1

BCG THEOREM. *If system* (1) *is completely reachable and completely observable, then*

i) *there exists a positive definite T-periodic solution P of equation* (2);

ii) *no other positive T-periodic semidefinite solution of equation* (2) *may exist different from P*;

iii) $A - BB'P$ *is asymptotically stable.*

A proof of the BCG theorem can be found in [17]. It consists of two distinct theorems, corresponding to points (i) and (ii), (iii), respectively, of the statement above. Since the proof of (i) [17, Thm. 1] is somewhat sketchy, a more complete proof is given in the Appendix.

LEMMA 1. *If system* (1) *is detectable and in Kalman canonical form, then the only nonzero element of the canonical decomposition* (7) *of any positive semidefinite T-periodic solution of equation* (2) *is* $P_0(t)$. *Furthermore* $P_0(t)$ *is positive definite, for all t.*

Proof. Referring to the canonical decomposition (7) of any possible T-periodic solution P of (2) and letting

$$\begin{aligned}
F(t) &= \tilde{P}_0(t)\tilde{A}_0(t) + \tilde{A}_0(t)'\tilde{P}_0(t)' - P_0(t)B_0(t)\bar{B}_0(t)\tilde{P}_0(t)' \\
&\quad - \tilde{P}_0(t)\bar{B}_0(t)B_0(t)'P_0(t) - \tilde{P}_0(t)\bar{B}_0(t)\bar{B}_0(t)'\tilde{P}_0(t)', \\
G(t) &= \tilde{P}_0(t)\bar{A}_0(t) + A_0(t)'\tilde{P}_0(t) + \tilde{A}_0(t)'\bar{P}_0(t) \\
&\quad - P_0(t)B_0(t)B_0(t)'\tilde{P}_0(t) - P_0(t)B_0(t)\bar{B}_0(t)'\bar{P}_0(t)' \\
&\quad - \tilde{P}_0(t)\bar{B}_0(t)B_0(t)'\tilde{P}_0(t) - \tilde{P}_0(t)\bar{B}_0(t)\bar{B}_0(t)'P_0(t), \\
H(t) &= B_0(t)'\tilde{P}_0(t) + \bar{B}_0(t)\bar{P}_0(t),
\end{aligned}$$

the corresponding decomposition (10) becomes

$$\begin{aligned}
-\dot{P}_0(t) &= P_0(t)A_0(t) + A_0(t)'P_0(t) + C_0(t)'C_0(t) \\
&\quad - P_0(t)B_0(t)B_0(t)'P_0(t) + F(t),
\end{aligned} \tag{14a}$$

$$-\dot{\tilde{P}}_0(t) = G(t), \tag{14b}$$

$$-\dot{\bar{P}}_0(t) = \bar{P}_0(t)\bar{A}_0(t) + \bar{A}_0(t)'\bar{P}_0(t) - H(t)'H(t). \tag{14c}$$

Thanks to the detectability assumption, $\bar{A}_0$ is asymptotically stable. Hence, by Theorem 2(ii), should H be not identically zero, (14c) would not admit any positive semidefinite T-periodic solution. Therefore, all positive semidefinite T-periodic solution of (2) must be such as to make $H = 0$.

In view of Theorem 2(i), the only positive T-periodic solution of (14c) is then $\bar{P}_0 = 0$, whereby $P(t) \geqq 0$ for all t implies $\tilde{P}_0 = 0$. Finally, $\bar{P}_0 = 0$ and $\tilde{P}_0 = 0$ imply $F = 0$ so that (14a) formally conforms to (2). In view of the Inertia Theorem, the complete observability of (A_0, C_0) implies that all positive semidefinite T-periodic solutions of (14a), being nonsingular, are in fact positive definite, for all t.

LEMMA 2. *Assume that system* (1) *is detectable and P is a positive semidefinite T-periodic solution of* (2); *then* $A - BB'P$ *is asymptotically stable.*

Proof. Assuming, without loss of generality, that system (1) is in Kalman canonical form; any positive semidefinite T-periodic solution P of equation (2) must, by Lemma 1, be such that

$$A(t) - B(t)B(t)'P(t) = \begin{bmatrix} A_0(t) - B_0(t)B_0(t)'P_0(t) & 0 \\ ? & \bar{A}_0(t) \end{bmatrix}$$

where ? denotes a block we do not consider specifically.

Detectability of (1) implies that $\bar{A}_0$ is asymptotically stable. On the other hand, the Inertia Theorem applied to (14a) ensures that $A_0 - B_0 B_0' P_0$ is asymptotically stable. Hence the lemma is proved.

LEMMA 3. *If system* (1) *is detectable, there cannot be more than one positive semidefinite T-periodic solution of* (2).

Proof. By contradiction, suppose that (2) admits two positive semidefinite T-periodic solutions $P^{(1)}$ and $P^{(2)}$. Let

$$P^{(12)}(t) = P^{(1)}(t) - P^{(2)}(t),$$
$$A^{(i)}(t) = A(t) - B(t)B(t)'P^{(i)}(t), \qquad i = 1, 2.$$

Obviously, $P^{(12)}$ is T-periodic and satisfies the equation:

$$-\dot{P}^{(12)}(t) = A^{(1)}(t)'P^{(12)}(t) + P^{(12)}(t)A^{(2)}(t). \tag{15}$$

By Lemma 2, $A^{(1)}$ and $A^{(2)}$ are asymptotically stable. Hence, Theorem 2(i), applied to (15), leads to concluding that the only T-periodic solution is $P^{(12)} = 0$, namely $P^{(1)} = P^{(2)}$.

Proof of Theorem 3. Without loss of generality, we can assume, since the very beginning, that system (1) is in Kalman canonical form. It is easy to check, then, that ($\bar{P}_0 = 0$, $\tilde{P}_0 = 0$) is a T-periodic solution of equations (10b) and (10c).

As for equation (10a), setting $\bar{P}_0 = 0$ and $\tilde{P}_0 = 0$ yields

$$-\dot{P}_0(t) = P_0(t)A_0(t) + A_0(t)'P_0(t) + C_0(t)'C_0(t) - P_0(t)B_0(t)B_0(t)'P_0(t). \tag{16}$$

Consider, then, the finer decomposition (11) of equation (16) induced by the Kalman canonical form of system (1):

$$\begin{aligned} -\dot{P}_{11}(t) &= P_{11}(t)A_{11}(t) + A_{11}(t)'P_{11}(t) + C_{11}(t)'C_{11}(t) \\ &\quad - P_{11}(t)B_{11}(t)B_{11}(t)'P_{11}(t), \end{aligned} \tag{17a}$$

$$\begin{aligned} -\dot{P}_{12}(t) &= P_{12}(t)A_{22}(t) + [A_{11}(t) - B_{11}(t)B_{11}(t)'P_{11}(t)]'P_{12}(t) \\ &\quad + P_{11}(t)A_{12}(t) + C_{11}(t)'C_{22}(t), \end{aligned} \tag{17b}$$

$$\begin{aligned} -\dot{P}_{22}(t) &= P_{22}(t)A_{22}(t) + A_{22}(t)'P_{22}(t) + P_{12}(t)'A_{12}(t) \\ &\quad + A_{12}(t)'P_{12}(t) + C_{22}(t)'C_{22}(t) - P_{12}(t)'B_{11}(t)B_{11}(t)'P_{12}(t). \end{aligned} \tag{17c}$$

Since (A_{11}, B_{11}) is completely controllable and (A_{11}, C_{11}) completely observable, the Riccati equation (17a)

(i) admits a unique positive semidefinite solution P_{11},

(ii) $P_{11}(t)$ is in fact positive definite, for all t, and $A_{11} - B_{11}B_{11}'P_{11}$ is asymptotically stable.

Substituting into (17b) and letting

$$M(t) = [A_{11}(t) - B_{11}(t)B_{11}(t)'P_{11}(t)]',$$
$$N(t) = A_{22}(t)',$$
$$W(t) = P_{11}(t)A_{12}(t) + C_{11}(t)'C_{22}(t),$$

(17b) takes on the form of (12).

Since M has already been recognized to be asymptotically stable and the asymptotic stability of the unreachable and observable part of system (1) implies the asymptotic stability of N, by Theorem 2(i), (17b) admits a (unique) T-periodic solution P_{12}.

Substituting back into (17c), and now letting

$$M(t) = N(t) = A_{22}(t)',$$

$$W(t) = P_{12}(t)'A_{12}(t) + A_{12}(t)'P_{12}(t) + C_{22}(t)'C_{22}(t) - P_{12}(t)'B_{11}(t)B_{11}(t)'P_{12}(t),$$

Theorem 2(i) enables us to conclude that (17c) as well admits a (unique) T-periodic solution.

To complete the proof, we have to show that

$$P_0(t) = \begin{bmatrix} P_{11}(t) & P_{12}(t) \\ P_{12}(t)' & P_{22}(t) \end{bmatrix}$$

is positive semidefinite, for all t. To this purpose consider

$$A_0(t) - B_0(t)B_0(t)'P_0(t) = \begin{bmatrix} A_{11}(t) - B_{11}(t)B_{11}(t)'P_{11}(t) & A_{12} - B_{11}(t)B_{11}(t)'P_{12}(t) \\ 0 & A_{22}(t) \end{bmatrix}.$$

Since both $A_{11} - B_{11}B_{11}'P_{11}$ and A_{22} are asymptotically stable, all the characteristic multipliers of $A_0 - B_0B_0'P_0$ lie in the open unit disk.

The Inertia Theorem applied to the Riccati equation (16) leads then to the conclusion that $P_0(t)$ is positive definite, for all t.

Proof of Theorem 4. Again we assume, without any loss of generality, that system (1) is in Kalman canonical form.

Sufficiency. Stabilizability and detectability of system (1) imply that all the assumptions of Lemma 2, Lemma 3 and Theorem 3 are verified (Fig. 1). Hence, it should be apparent that existence follows from Theorem 3, uniqueness from Lemma 3 and "feedback stabilization" from Lemma 2.

Necessity. Conversely, the existence of a stabilizing feed back control law based on a unique positive semidefinite T-periodic solution P of equation (2) obviously implies that system (1) is stabilizable.

To complete the proof, we only need to show that system (1) is detectable. To this end, observe first that stabilizability of system (1) implies, in particular, the asymptotic stability of its observable and unreachable part. Hence, by Theorem 3, only the observability submatrix P_0 of the unique positive semidefinite solution P of equation (2) can be different from zero. Furthermore, $P_0(t)$ is positive definite for all t. It is then easy to see that

$$A - BB'P = \begin{bmatrix} ? & 0 \\ ? & \bar{A}_0 \end{bmatrix}.$$

Therefore, the asymptotic stability of $A - BB'P$ implies the asymptotic stability of $\bar{A}_0$, namely the detectability of system (1).

5. Concluding remarks. In this paper, a new proof of the (periodic) Wonham-Kucera theorem has been presented. The proof is entirely based on the canonical

decomposition of the periodic Riccati equation induced by the Kalman canonical decomposition of the underlying periodic system. The novel approach enables gaining a deep insight into the structure of the T-periodic solutions of the periodic Riccati equation. The extent to which the knowledge of the solution structure can be exploited from a computational point of view in finding the unique positive semidefinite periodic solution of a given periodic Riccati equation is an interesting open issue worthy of further investigation.

Appendix. In this appendix we prove that if system (1) is completely reachable, then there exists a positive semidefinite periodic solution of (2). In fact, consider the following optimal periodic control problem.

Minimize

$$J=\frac{1}{2}\int_t^\tau \{\|y(\sigma)\|^2+\|u(\sigma)\|^2\}\,d\sigma \tag{A.1}$$

subject to

$$\dot{x}(\sigma)=A(\sigma)x(\sigma)+B(\sigma)u(\sigma), \tag{A.2}$$

$$y(\sigma)=C(\sigma)x(\sigma), \tag{A.3}$$

$$x(t)=x_t. \tag{A.4}$$

Let $\Delta(t, \tau, D)$ be the solution of (2) at time t such that $\Delta(\tau, \tau, D)=D$.

It is claimed that the limit

$$\bar{P}(t)=\lim_{\tau\to\infty}\Delta(t, \tau, 0) \tag{A.5}$$

exists and is a T-periodic positive semidefinite solution of (2). Indeed, it is well known that the optimal value J^0 of the performance index (A.1) is given by

$$J^0=\tfrac{1}{2}x_t'\Delta(t, \tau, 0)x_t. \tag{A.6}$$

From (A.6) it follows that $\Delta(t, \cdot, 0)$ is a monotonically nondecreasing matrix function. Furthermore, since $(A(\cdot), B(\cdot))$ is completely controllable, there exists a bounded control function $\tilde{u}(\cdot)$ and a time point $\tilde{\tau}>t$ such that

$$\tilde{u}(\sigma)=0 \quad \forall\sigma>\tilde{\tau},$$

$$\tilde{x}(\sigma)=0 \quad \forall\sigma\geqq\tilde{\tau},$$

where $\tilde{x}(\cdot)$ is the solution of (A.2)-(A.4) associated with control function $\tilde{u}(\cdot)$.

Hence,

$$\tilde{J}=\frac{1}{2}\int_t^\infty \{\|C(\sigma)\tilde{x}(\sigma)\|^2+\|\tilde{u}(\sigma)\|^2\}\,d\sigma<\infty.$$

Since $J^0\leqq\tilde{J}$, $\Delta(t, \cdot, 0)$ is bounded from above. Thus the conclusion is drawn that limit (A.5) does exists. Note that, as $\Delta(t, \tau, 0)$ is positive semidefinite for any t and τ, $\bar{P}(t)$ is also positive semidefinite. Furthermore, the time periodicity of (2) implies that

$$\Delta(t+T, \tau+T, 0)=\Delta(t, \tau, 0).$$

Consequently,

$$\bar{P}(t+T)=\lim_{\tau\to\infty}\Delta(t+T, \tau, 0)=\lim_{\tau\to\infty}\Delta(t, \tau, 0)=\bar{P}(t).$$

Finally, note that

$$\Delta(t, \tau, 0) = \Delta(t, \bar{\tau}, \Delta(\bar{\tau}, \tau, 0))$$

for any $\bar{\tau} \in [t, \tau]$. Since $\Delta(t, \tau, D)$ is differentiable with respect to D (see e.g. [9, p. 11]), then

$$\bar{P}(t) = \lim_{\tau \to \infty} \Delta(t, \tau, 0) = \Delta(t, \bar{\tau}, \lim_{\tau \to \infty} \Delta(\bar{\tau}, \tau, 0)) = \Delta(t, \bar{\tau}, \bar{P}(\bar{\tau}))$$

for all t and $\bar{\tau} \geqq t$. Hence, $\bar{P}(t)$ is a solution of (2).

REFERENCES

[1] W. M. WONHAM, *On a matrix Riccati equation of stochastic control*, this Journal, 6 (1968), pp. 681-697.

[2] V. KUCERA, *A contribution to matrix quadratic equations*, IEEE Trans. Automat. Control, AC-17 (1972), pp. 344, 347.

[3] G. A. HEWER, *Periodicity, detectability and the matrix Riccati equation*, this Journal, 13 (1975), pp. 1235-1251.

[4] H. KANO AND T. NISHIMURA, *Periodic solutions of matrix Riccati equations with detectability and stabilizability*, Internat. J. Control, 29 (1979), pp. 471-487.

[5] S. BITTANTI, G. GUARDABASSI, C. MAFFEZZONI AND L. SILVERMAN, *Periodic systems: controllability and the matrix Riccati equation*, this Journal, 16 (1978), pp. 37-40.

[6] S. BITTANTI, P. BOLZERN, P. COLANERI AND G. GUARDABASSI, *H and K-controllability of linear periodic systems*, 22nd IEEE Conference Decision and Control, San Antonio, TX, 1983, pp. 1376-1379.

[7] S. BITTANTI, P. COLANERI AND G. GUARDABASSI, *H-Controllability and observability of linear periodic systems*, this Journal, 22 (1984), pp. 889-893.

[8] S. BITTANTI AND P. BOLZERN, *Four equivalent notions of stabilizability of periodic linear systems*, American Control Conference, San Diego, CA, 1984, pp. 1321-1323.

[9] A. HALANAY, *Differential Equations*, Academic Press, New York, 1966.

[10] V. A. YAKUBOVICH AND V. M. STARZHINSKII, *Linear Differential Equations with Periodic Coefficients*, John Wiley, New York, 1972.

[11] M. ROSEAU, *Equations differentielles*, Masson, Paris, 1976.

[12] S. BITTANTI AND P. BOLZERN, *Stabilizability and detectability of linear periodic systems*, Systems Control Lett., 6 (1985), pp. 141-145.

[13] A. GRAHAM, *Kronecker Products and Matrix Calculus with Applications*, John Wiley, New York, 1981.

[14] S. BITTANTI, P. BOLZERN AND P. COLANERI, *The extended periodic Lyapunov lemma*, Automatica, 21 (1985), pp. 603-605.

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

[16] M. A. SHAYMAN, *Inertia theorems for the periodic Lyapunov equation and periodic Riccati equation*, Systems Control Lett., 4 (1984), pp. 27-32.

[17] S. BITTANTI, P. COLANERI AND G. GUARDABASSI, *Periodic solutions of periodic Riccati equations*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 665-668.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 6, November 1986

005

ALMOST DISTURBANCE DECOUPLING WITH BOUNDED PEAKING*

HARRY L. TRENTELMAN†

Abstract. This paper is concerned with a generalization of the almost disturbance decoupling problem by state feedback. Apart from approximate decoupling from the external disturbances to a first to-be-controlled output, we require a second output to be uniformly bounded with respect to the accuracy of decoupling. The problem is studied using the geometric approach to linear systems. We introduce some new almost controlled invariant subspaces and study their geometric structure. Necessary and sufficient conditions for the solvability of the above problem are formulated in terms of these controlled invariant subspaces. A conceptual algorithm is introduced to calculate the feedback laws needed to achieve the design purpose.

Key words. almost disturbance decoupling, almost invariant subspaces, linear systems, geometric approach, high gain feedback, output stabilization

AMS(MOS) subject classifications. G3-B28, G3-B50, G3-C05, G3-C15, G3-C35, G3-C45, G3-C60

1. Introduction. In this paper we are concerned with the problem of almost disturbance decoupling by state feedback as introduced by Willems [20]. This problem deals with the situation in which we cannot achieve exact decoupling from the external disturbances to an exogenous output channel as, for example, in [22], but only *approximate* decoupling with any desired degree of accuracy. In general, the feedback gain necessary to achieve this will increase as the desired degree of accuracy increases. It may then happen however that some of the state variables tend to peak excessively. It is of considerable practical interest to know when it is possible to achieve disturbance decoupling within any desired degree of accuracy, while this peaking phenomenon will not occur.

The system that we will be considering in this paper is given by the equations

$$\begin{aligned} \dot{x} &= Ax + Bu + Gd, \\ z_1 &= H_1 x, \\ z_2 &= H_2 x, \end{aligned} \tag{1.1}$$

where the control $u(t)$, the state $x(t)$, the disturbance $d(t)$ and the outputs $z_1(t)$ and $z_2(t)$ are real vectors of finite dimensions. We will assume that the vector $z_2(t)$ is an *enlargement* of $z_1(t)$, i.e., there is a matrix M such that $H_1 = MH_2$. If for any positive real number ε a feedback matrix F_ε can be chosen such that in the closed loop system with zero initial condition, for all disturbances $d(\cdot)$ in the unit ball of $L_p[0, \infty)$ we have

$$\|z_1\|_{L_p} \leqq \varepsilon \tag{1.2}$$

then we say that for the system under consideration the L_p-almost disturbance decoupling problem from d to z_1 is solvable. After choosing F_ε to achieve this approximate decoupling, the output $z_2(t)$ of course depends on ε and, for certain disturbances $d(\cdot)$, it may then happen that $\|z_2\|_{L_p} \to \infty$ as $\varepsilon \to 0$, i.e., as the accuracy of decoupling increases.

* Received by the editors April 28, 1983, and in revised form June 1, 1985.

† Department of Mathematics and Computing Science, Eindhoven University of Technology, 5600 MB Eindhoven, the Netherlands. This research was supported by the Netherlands Organization for the Advancement of Pure Science Research (Z.W.O.).

As an example, consider the system (1.1) with

$$A=\begin{pmatrix}0 & 1 & 0\\ 0 & 0 & 1\\ 0 & 0 & 0\end{pmatrix},\qquad B=\begin{pmatrix}0\\ 0\\ 1\end{pmatrix},\qquad G=\begin{pmatrix}1\\ 0\\ 0\end{pmatrix},$$

$$H_1=(1\quad 0\quad 0),\qquad H_2=\begin{pmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{pmatrix}.$$

Define a feedback matrix F_ε by

$$F_\varepsilon := \left(-\frac{27}{\varepsilon^3},\quad -\frac{27}{\varepsilon^2},\quad -\frac{9}{\varepsilon}\right).$$

It can then be verified that the impulse response from the disturbance d to z_1 is given by

$$W_{1,\varepsilon}(t) := H_1 e^{(A+BF_\varepsilon)t}G = e^{-3t/\varepsilon}\left(1+\frac{3}{\varepsilon}t+\frac{9}{2\varepsilon^2}t^2\right)$$

and that $\| W_{1,\varepsilon}\|_{L_1} = \varepsilon$. Hence, for any $1 \leqq p \leqq \infty$, the above feedback matrix F_ε achieves L_p-amost disturbance decoupling from d to z_1. On the other hand, however, the impulse response from d to z_2 is calculated to be

$$W_{2,\varepsilon}(t) := H_2 e^{(A+BF_\varepsilon)t}G = e^{-3t/\varepsilon}\begin{pmatrix}1+\frac{3}{\varepsilon}t+\frac{9}{2\varepsilon^2}t^2\\ -\frac{27}{2\varepsilon^3}t^2\\ -\frac{27}{\varepsilon^3}t+\frac{81}{2\varepsilon^4}t^2\end{pmatrix}$$

and it can be verified that $\| W_{2,\varepsilon}\|_{L_1} = O(1/\varepsilon) \to \infty$ as $\varepsilon \to 0$, i.e., we have obtained almost disturbance decoupling from d to z_1 at the cost of highly undesired peaking behaviour of the output $z_2(t)$.

The question which we ask in this paper is this: *When is it possible to choose F_ε such that simultaneously* (1.2) *holds and there exists a constant C (independent of ε) such that for all disturbances $d(\cdot)$ in the unit ball of $L_p[0,\infty)$ we have*

$$\|z_2\|_{L_p} \leqq C \tag{1.3}$$

for all ε? That is, the output $z_2(t)$ is bounded uniformly as ε tends to zero. If this behaviour is achieved, we say that we have *L_p-bounded peaking* from d to z_2. Problems of this kind have been considered before. Francis and Glover [3] considered a bounded peaking problem in the context of cheap control. More recently, Kimura [9] found conditions that guarantee bounded peaking in the context of perfect regulation. We will study the above problem using the by now well known concepts of almost controlled invariant and almost controllability subspace [19], [20]. We will also use the approach of frequency domain description of geometric concepts as initiated in Hautus [5].

The outline of this paper is as follows. In § 2 we will introduce some notational conventions used in this paper and state some preliminary results and background.

Section 3 contains a description of the main problem we will be concerned with in this paper. In § 4 we will introduce the disturbance decoupling problem with output stability. This problem is an extension of the (exact) disturbance decoupling problem

as treated in [22]. Its solution will be needed to solve our main problem, but is also important in its own right. In § 5 we will derive a necessary condition for the solvability of $(\text{ADDPBP})_p$. This condition will be in the form of a subspace inclusion involving an almost controlled invariant subspace. Section 6 contains an investigation of the geometric structure of the almost controlled invariant subspace that was introduced in § 5. In § 7 these structural results will be used to prove that for certain classes of systems the subspace inclusion derived in § 5 in fact constitutes a necessary and sufficient condition for solvability of $(\text{ADDPBP})_p$. The sequences of state feedback maps that achieve the design purpose will be constructed explicitly. Section 8 contains some corollaries of our main result and some extensions. In § 9 a numerical example is worked out to illustrate the computational feasibility of our theory. Finally, the paper closes with some concluding remarks in § 10. Several technical details of proofs in this paper are deferred to Appendices A, B and C.

2. Preliminaries and background. In this section we will introduce some notation used in this paper and review some relevant facts on controlled invariant and almost controlled invariant subspaces. Also some basic facts on the convergence of subspaces will be given.

2.1. In this paper the following notation will be used: If $\mathscr{X}$ is a normed vector space, we will write $\|\cdot\|$ for the norm on $\mathscr{X}$. If $l:[0,\infty)\to\mathscr{X}$ is a measurable function, then we will denote

$$\|l\|_{L_p} := \begin{cases} \left(\int_0^\infty \|l(t)\|^p\,dt\right)^{1/p} & \text{if } 1 \leqq p < \infty, \\ \operatorname{ess\,sup}_{t\geqq 0} \|l(t)\| & \text{if } p = \infty. \end{cases}$$

If $\|l\|_{L_p} < \infty$, we will say that $l \in L_p[0,\infty)$. If M is a square matrix then $\sigma(M)$ will denote its spectrum. If Λ_1 and Λ_2 are sets of complex numbers then $\Lambda_1 \cup \Lambda_2$ will denote their disjoint union. For any positive integer n, we will denote $\underline{n} := \{1, 2, \cdots, n\}$.

Consider the system (1.1). Let $u(t) \in \mathscr{U} := \mathbb{R}^m$, $x(t) \in \mathscr{X} := \mathbb{R}^n$, $d(t) \in \mathscr{D} := \mathbb{R}^q$, $z_1(t) \in \mathscr{Z}_1 := \mathbb{R}^{p_1}$ and $z_2(t) \in \mathscr{Z}_2 := \mathbb{R}^{p_2}$. Let A, B, G, H_1 and H_2 be real matrices of appropriate dimensions. We will write $\mathscr{K}_i := \ker H_i$ $(i = 1, 2)$, $\mathscr{B} := \operatorname{im} B$ and $A_F := A + BF$. The reachable subspace will be denoted by $\langle A \mid \mathscr{B}\rangle := \mathscr{B} + A\mathscr{B} + \cdots + A^{n-1}\mathscr{B}$. A collection of subspaces $\mathscr{B}_1, \mathscr{B}_2, \cdots, \mathscr{B}_k$ will be called a chain in $\mathscr{B}$ if $\mathscr{B} \supset \mathscr{B}_1 \supset \mathscr{B}_2 \supset \cdots \supset \mathscr{B}_k$. If $0 \neq b \in \mathscr{B}$ we will denote $\mathscr{b} := \operatorname{span} b$.

If $\mathscr{V} \subset \mathscr{X}$ is A_F-invariant, the restriction of A_F to $\mathscr{V}$ will be denoted by $A_F \mid \mathscr{V}$. We will write $A_F \mid \mathscr{X}/\mathscr{V}$ or $\bar{A}_F$ for the quotient map induced by A_F on the factor space $\mathscr{X}/\mathscr{V}$ (see [22]). If $\mathscr{V}$ and $\mathscr{W}$ are both A_F-invariant and $\mathscr{W} \subset \mathscr{V}$, then $A_F \mid \mathscr{V}/\mathscr{W}$ will denote the map induced by $A_F \mid \mathscr{V}$ on the factor space $\mathscr{V}/\mathscr{W}$. We define the canonical projection $P: \mathscr{X} \to \mathscr{X}/\mathscr{V}$ by $Px := x + \mathscr{V}$. If $\bar{B} := PB$, then $(\bar{A}_F, \bar{B})$ will be called the system induced in $\mathscr{X}/\mathscr{V}$. If $H: \mathscr{X} \to \mathscr{Z}$ is a linear map and $\mathscr{V} \subset \ker H$, then $\bar{H}: \mathscr{X}/\mathscr{V} \to \mathscr{Z}$ is defined by $\bar{H}P = H$. A distribution $f \in \mathscr{D}'_+$ (i.e., the space of finite-dimensional valued distributions with support on $[0,\infty)$) will be called a Bohl distribution if there exist vectors f_i and matrices K, L, M such that $f = \sum_{i=0}^{N} f_i \delta^{(i)} + f_{-1}$. Here $f_{-1}(t) := Ke^{Lt}M$, $\delta^{(0)}$ denotes Dirac's delta and $\delta^{(i)}$ its ith distributional derivative. f will be called regular if $f_i = 0$ $(i = 0, \cdots, N)$ and impulsive if $f_{-1} \equiv 0$.

2.2. We will now review some basic facts from geometric control theory. If $\mathscr{K} \subset \mathscr{X}$ is a subspace, then $\mathscr{V}^*(\mathscr{K})$ will denote the largest (A, B)- or controlled invariant subspace in $\mathscr{K}$ and $\mathscr{R}^*(\mathscr{K})$ will denote the largest controllability subspace in $\mathscr{K}$ [22].

If $\mathbb{C}_g$ is a symmetric subset of the complex plane $\mathbb{C}$ (i.e., $\lambda \in \mathbb{C}_g \Leftrightarrow \bar{\lambda} \in \mathbb{C}_g$ and $\mathbb{C}_g$ contains at least one point of the real axis), then $\mathcal{V}_g^*(\mathcal{K})$ will denote the largest stabilizability subspace in $\mathcal{K}$ ([5] or [11]).

A subspace $\mathcal{V}_a \subset \mathcal{X}$ will be called almost controlled invariant if for all $x_0 \in \mathcal{V}_a$ and for all $\varepsilon > 0$ there is a state trajectory $x_\varepsilon(\cdot)$ such that $x_\varepsilon(0) = x_0$ and $d(\mathcal{V}_a, x_\varepsilon(t)) \leq \varepsilon$ for all t. A subspace $\mathcal{R}_a \subset \mathcal{X}$ will be called an almost controllability subspace if for all $x_0, x_1 \in \mathcal{R}_a$ there is a $T > 0$ such that for all $\varepsilon > 0$ there is a state trajectory $x_\varepsilon(\cdot)$ such that $x_\varepsilon(0) = x_0$, $x_\varepsilon(T) = x_1$ and $d(\mathcal{R}_a, x_\varepsilon(t)) \leqq \varepsilon$ for all t. Basic facts on these classes of subspaces can be found in [19] or [20] (see also [17]). A subspace $\mathcal{V}_a \subset \mathcal{X}$ is almost controlled invariant if and only if $\mathcal{V}_a = \mathcal{V} + \mathcal{R}_a$, where $\mathcal{V}$ is controlled invariant and $\mathcal{R}_a$ is an almost controllability subspace. A subspace $\mathcal{R}_a$ is an almost controllability subspace if and only if there is a map $F: \mathcal{X} \to \mathcal{U}$ and a chain $\{\mathcal{B}_i\}_{i=1}^k$ in $\mathcal{B}$ such that $\mathcal{R}_a = \mathcal{B}_1 + A_F\mathcal{B}_2 + \cdots + A_F^{k-1}\mathcal{B}_k$. We will say that $\mathcal{R}_a$ is a singly generated almost controllability subspace if there is a map $F: \mathcal{X} \to \mathcal{U}$, a vector $b \in \mathcal{B}$ and an integer $k > 0$ such that $\mathcal{R}_a = \mathcal{b} \oplus A_F\mathcal{b} \oplus \cdots \oplus A_F^{k-1}\mathcal{b}$.

Again, for $\mathcal{K} \subset \mathcal{X}$, $\mathcal{V}_a^*(\mathcal{K})$ will denote the largest almost controlled invariant subspace in $\mathcal{K}$ and $\mathcal{R}_a^*(\mathcal{K})$ the largest almost controllability subspace in $\mathcal{K}$. We will denote $R_b^*(\mathcal{K}) := \mathcal{B} + A\mathcal{R}_a^*(\mathcal{K})$ and $\mathcal{V}_b^*(\mathcal{K}) := \mathcal{V}^*(\mathcal{K}) + \mathcal{R}_b^*(\mathcal{K})$. The subspace $\mathcal{V}_b^*(\mathcal{K})$ plays an important role in the problem of almost disturbance decoupling. In fact, in [20] the following result was obtained:

PROPOSITION 2.1. *Consider the system $\dot{x} = Ax + Bu$, $z = Hx$. Then for all $\varepsilon > 0$ there exists a map $F_\varepsilon: \mathcal{X} \to \mathcal{U}$ such that $\|H \exp[t(A + BF_\varepsilon)]G\|_{L_1} \leqq \varepsilon$ if and only if* $\operatorname{im} G \subset \mathcal{V}_b^*(\ker H)$.

Let $\mathcal{K} := \ker H$. The space $\mathcal{V}_b^*(\mathcal{K})$ will be called the space of distributionally weakly unobservable states with respect to the output z. $\mathcal{R}_b^*(\mathcal{K})$ will be called the space of strongly controllable states with respect to the output z. For this terminology see [6].

A proof of the following result can be found in [1, Lemma 1].

LEMMA 2.2. *Let $\mathcal{K} \subset \mathcal{X}$. Then the following equalities hold:*

(i) $\mathcal{R}_b^*(\mathcal{K}) \cap \mathcal{K} = \mathcal{R}_a^*(\mathcal{K})$,

(ii) $\mathcal{R}_a^*(\mathcal{K}) \cap \mathcal{V}^*(\mathcal{K}) = \mathcal{R}^*(\mathcal{K})$,

(iii) $\mathcal{R}_b^*(\mathcal{K}) \cap \mathcal{V}^*(\mathcal{K}) = \mathcal{R}^*(\mathcal{K})$. □

This paper will sometimes deal with a new system (A, BW), obtained by deleting the part of the input matrix B lying in $\mathcal{V}^*(\mathcal{K})$. This system is obtained by taking $\tilde{\mathcal{B}} \subset \mathcal{B}$ such that $\tilde{\mathcal{B}} \oplus (\mathcal{B} \cap \mathcal{V}^*(\mathcal{K})) = \mathcal{B}$ and by letting W be a map such that $\tilde{\mathcal{B}} = \operatorname{im} BW$ (see also [1]). The supremal almost controllability subspace contained in $\mathcal{K}$ with respect to this new system (A, BW) will be denoted by $\tilde{\mathcal{R}}_a^*(\mathcal{K})$. We will correspondingly denote $\tilde{\mathcal{B}} + A\tilde{\mathcal{R}}_a^*(\mathcal{K})$ by $\tilde{\mathcal{R}}_b^*(\mathcal{K})$. The following result follows from [1, Lemma 2]:

LEMMA 2.3.

$$\mathcal{V}^*(\mathcal{K}) \cap \tilde{\mathcal{R}}_b^*(\mathcal{K}) = \{0\}. \qquad \square$$

Assume now that $\mathcal{V} \subset \mathcal{X}$ is (A, B)-invariant. Let F be such that $(A + BF)\mathcal{V} \subset \mathcal{V}$. Let $(\bar{A}_F, \bar{B})$ be the system induced in $\mathcal{X}/\mathcal{V}$ and $P: \mathcal{X} \to \mathcal{X}/\mathcal{V}$ the canonical projection. We then have the following result:

LEMMA 2.4. *If $\mathcal{R}_a$ is an almost controllability subspace with respect to (A, B), then $P\mathcal{R}_a$ is an almost controllability subspace with respect to $(\bar{A}_F, \bar{B})$.*

Proof. Let Px_0 and Px_1 be in $P\mathcal{R}_a$, with $x_0, x_1 \in \mathcal{R}_a$. There is a $T > 0$ and, for all $\varepsilon > 0$, a trajectory $x_\varepsilon(\cdot)$ such that $x_\varepsilon(0) = x_0$, $x_\varepsilon(T) = x_1$ and $d(\mathcal{R}_a, x_\varepsilon(t)) \leqq \varepsilon$ for all t. It can be seen immediately that $z_\varepsilon(t) := Px_\varepsilon(t)$ is a trajectory of the system $(\bar{A}_F, \bar{B})$. Moreover, $z_\varepsilon(0) = Px_0$, $z_\varepsilon(T) = Px_1$ and $d(P\mathcal{R}_a, z_\varepsilon(t)) = \inf_{r \in R_a} \|Pr - Px_\varepsilon(t)\| \leqq \|P\| d(\mathcal{R}_a, x_\varepsilon(t)) \leqq \varepsilon \|P\|$. □

We will also need the following proposition, which is proven in [17, Thm. 2.39] (see also [15] or [16]).

PROPOSITION 2.5. *Consider the system* $\dot{x} = Ax + Bu$. *Let* $\mathcal{R}_a$ *be an almost controllability subspace. Suppose* Λ *is a symmetric set of* $\dim \langle A \mid \mathcal{B} \rangle - \dim \mathcal{R}_a$ *complex numbers. Then there is an* (A, B)*-invariant subspace* $\mathcal{V}$ *and a map* $F: \mathcal{X} \to \mathcal{U}$ *such that* $\mathcal{V} \oplus \mathcal{R}_a = \langle A \mid \mathcal{B} \rangle$ *and* $\sigma(A_F \mid \mathcal{V}) = \Lambda$.

To conclude this section, we shall recall some facts on left-invertibility of linear systems. Again consider the system $\dot{x} = Ax + Bu$, $z = Hx$. Assume that the map B is injective. We will say that the system (A, B, H) is left-invertible if the transfer matrix $H(Is - A)^{-1}B$ is an injective rational matrix. The following result was proven in [22, Ex. 4.4] (see also [6, Thm. 3.26]).

LEMMA 2.6. (A, B, H) *is a left-invertible system if and only if* $\mathcal{R}^*(\ker H) = 0$. □

2.3. In the following, we will review some basic facts on the frequency domain approach to the geometric concepts of this paper. We will denote $\mathcal{X}[s]$ (respectively, $\mathcal{X}(s)$, $\mathcal{X}_+(s)$) for the set of all n-vectors whose components are polynomials (respectively, rational functions, strictly proper rational functions) with coefficients in $\mathbb{R}$. If $\mathcal{K} \subset \mathcal{X} = \mathbb{R}^n$, then $\mathcal{K}[s]$ (respectively, $\mathcal{K}(s)$, $\mathcal{K}_+(s)$) will denote the set of all $\xi(s) \in \mathcal{X}[s]$ (respectively, $\mathcal{X}(s)$, $\mathcal{X}_+(s)$) with the property that $\xi(s) \in \mathcal{K}$ for all s. Slightly generalizing a definition by Hautus [5], if for a given $x \in \mathcal{X}$ there are rational functions $\xi(s) \in \mathcal{X}(s)$ and $\omega(s) \in \mathcal{U}(s)$ such that $x = (Is - A)\xi(s) + B\omega(s)$ for all s, we will say that x has a (ξ, ω)*-representation.*

For a description of (almost) controlled invariant subspaces in terms of (ξ, ω)-representations, we refer to [5], [12], [13] and [17]. We shall need the following fact:

LEMMA 2.7. *Let* $\mathcal{K} \subset \mathcal{X}$. *Then we have:* $x \in \mathcal{R}_b^*(\mathcal{K})$ *if and only if* x *has a* (ξ, ω)*-representation with* $\xi(s) \in \mathcal{K}[s]$ *and* $\omega(s) \in \mathcal{U}[s]$. □

2.4. Finally, we will recall some facts on the convergence of subspaces. In this paper we will use the common notion of convergence of subspaces in the sense of Grassmanian topology. Let $\{\mathcal{V}_\varepsilon; \varepsilon > 0\}$ be a sequence of subspaces of $\mathcal{X}$ of fixed dimension. It can be proven that $\mathcal{V}_\varepsilon \to \mathcal{V} (\varepsilon \to 0)$ if and only if there is a basis $\{v_1, \cdots, v_q\}$ for $\mathcal{V}$ and there are bases $\{v_1(\varepsilon), \cdots, v_q(\varepsilon)\}$ of $\mathcal{V}_\varepsilon$ such that, for all i, $v_i(\varepsilon) \to v_i$ as $\varepsilon \to 0$ (convergence in $\mathcal{X}$). We will need the following lemma, which can be proven by standard means:

LEMMA 2.8. *Suppose* $v_1, \cdots, v_q$ *are independent vectors and* $v_i(\varepsilon) \to v_i$ *for all* i. *Then for* ε *sufficiently small,* $v_1(\varepsilon), \cdots, v_q(\varepsilon)$ *are linearly independent. Consequently, if* $\mathcal{V}_\varepsilon \to \mathcal{V}$ *and* $\mathcal{W}_\varepsilon \to \mathcal{W}$, *where* $\mathcal{V} \cap \mathcal{W} = \{0\}$, *then for* ε *sufficiently small* $\mathcal{V}_\varepsilon \cap \mathcal{W}_\varepsilon = \{0\}$ *and* $\mathcal{V}_\varepsilon \oplus \mathcal{W}_\varepsilon \to \mathcal{V} \oplus \mathcal{W}$. □

3. Mathematical problem formulation. Consider the system (1.1). We will assume that $z_2(t)$ is an *enlargement* of $z_1(t)$, that is, there is a matrix M such that $H_1 = MH_2$ or, equivalently,

$$\ker H_2 =: \mathcal{K}_2 \subset \mathcal{K}_1 := \ker H_1. \tag{3.1}$$

From now on, (3.1) will be a standing assumption. Throughout this paper we will also assume that B is injective.

Consider the following synthesis problem. Fix $1 \leqq p \leqq \infty$. We will say that *the* L_p*-almost disturbance decoupling problem with bounded peaking* $(\text{ADDPBP})_p$ is solvable if there is a constant C and for all $\varepsilon > 0$ there is a feedback map $F_\varepsilon: \mathcal{X} \to \mathcal{U}$ such that, with the feedback law $u = F_\varepsilon x$ in the closed loop system for $x(0) = 0$ for all $d \in L_p[0, \infty)$, the following inequalities hold:

$$\|z_1\|_{L_p} \leqq \varepsilon \|d\|_{L_p}, \tag{3.2}$$

$$\|z_2\|_{L_p} \leqq C\|d\|_{L_p}. \tag{3.3}$$

Note that if we take $H_1 = H_2$, we obtain the original L_p-almost disturbance decoupling problem, $(\text{ADDP})_p$, without the requirement of bounded peaking (see [20] or [17]). Another interesting special case is to take $H_2 = I$, which corresponds to the requirement of bounded peaking *of the entire state vector.*

In the present paper, necessary and sufficient geometric conditions for the solvability of the above problem will be derived for the cases $p = 1$, $p = 2$ and $p = \infty$. We will first show how the solvability of $(\text{ADDPBP})_p$ can be expressed in terms of the closed loop impulse response matrices from the disturbance d to the outputs z_1 and z_2, respectively. If $F_\varepsilon : \mathscr{X} \to \mathscr{U}$ is a state feedback map, then denote the closed loop transition matrix by

$$T_\varepsilon(t) := e^{(A+BF_\varepsilon)t} \tag{3.4}$$

and let

$$\hat{T}_\varepsilon(s) := (Is - A - BF_\varepsilon)^{-1} \tag{3.5}$$

denote its Laplace transform. We then have the following:

LEMMA 3.1. *Fix $p \in \{1, \infty\}$. Then $(\text{ADDPBP})_p$ is solvable if and only if there is a constant C and for all $\varepsilon > 0$ a feedback map $F_\varepsilon : \mathscr{X} \to \mathscr{U}$ such that $\|H_1 T_\varepsilon G\|_{L_1} \leqq \varepsilon$ and $\|H_2 T_\varepsilon G\|_{L_1} \leqq C$.*

$(\text{ADDPBP})_2$ is solvable if and only if there is a constant C and for all $\varepsilon > 0$ a feedback map F_ε such that $H_1\hat{T}_\varepsilon(s)G$ and $H_2\hat{T}_\varepsilon(s)G$ are asymptotically stable and such that $\sup_{\omega\in\mathbb{R}} \|H_1\hat{T}_\varepsilon(i\omega)G\| \leqq \varepsilon$ and $\sup_{\omega\in\mathbb{R}} \|H_2\hat{T}_\varepsilon(i\omega)G\| \leqq C$.

Proof. The proof follows immediately from the fact that for $p = 1$ and for $p = \infty$ the L_p-included norm of the closed loop convolution operator from d to z_i equals exactly the L_1-norm of its kernel, i.e., $\|H_i T_\varepsilon G\|_{L_1}$. The second statement follows from the fact that the L_2-induced norm of the convolution operator from d to z_i equals the H^∞-norm $\sup_{\omega\in\mathbb{R}} \|H_i\hat{T}_\varepsilon(i\omega)G\|$ (see, for example, [2]). □

4. Disturbance decoupling with stability constraints. Prior to considerations involving the peaking behaviour of the enlarged output z_2, we should make sure that the output z_2 is in $L_p[0, \infty)$ at all. Hence, an important part of the solution of the problem posed in § 3 is to construct the required feedback maps F_ε in such a way that, for any $d \in L_p[0, \infty)$, in the closed loop system with $x(0) = 0$ we have $z_2 \in L_p[0, \infty)$. Therefore, in this section the following variation on the well known (exact) disturbance decoupling problem [22] will be considered. Again, consider the system given by (1.1). The usual disturbance decoupling problem is concerned with the determination of a feedback map $F : \mathscr{X} \to \mathscr{U}$ such that in the closed loop system the external disturbance d does not influence a specified output z_1. We will consider the more general situation in which simultaneously we demand stability of the second, larger, output z_2.

In this section, $\mathbb{C}_g$, the stability set, will be a given subset of the complex plane $\mathbb{C}$ which is symmetric. Asymptotic stability is thus obtained by taking $\mathbb{C}_g = \{\lambda \in \mathbb{C} : \operatorname{Re}\lambda < 0\}$. A rational matrix or rational vector is called stable if all its poles are in $\mathbb{C}_g$. We will consider the following problem: (DDPOS) *the disturbance decoupling problem with output stabilization* is said to be solvable if there is a feedback map F such that $H_1(Is - A_F)^{-1}G = 0$ and $H_2(Is - A_F)^{-1}G$ is stable.

In order to be able to formulate conditions for the solvability of the above problem, introduce the following subspace:

DEFINITION 4.1. $\mathscr{V}_g(\mathscr{K}_1, \mathscr{K}_2)$ will denote the subspace of all points $x \in \mathscr{K}_1$ for which there is a (ξ, ω)-representation with $\xi(s) \in \mathscr{K}_{1,+}(s)$, $\omega(s) \in \mathscr{U}_+(s)$ and $H_2\xi(s)$ is *stable.*

Thus, interpreted in the time domain, $\mathcal{V}_g(\mathcal{K}_1, \mathcal{K}_2)$ is the subspace consisting of all points in which a regular Bohl state trajectory starts that lies entirely in $\mathcal{K}_1$. The components of this trajectory modulo $\mathcal{K}_2$ are stable. It follows immediately from the definition that $\mathcal{V}_g(\mathcal{K}_1, \mathcal{K}_2)$ is contained in $\mathcal{V}^*(\mathcal{K}_1)$. By the assumption (3.1), if a trajectory lies in $\mathcal{K}_2$, the same is true for $\mathcal{K}_1$. Consequently we also have the inclusion $\mathcal{V}^*(\mathcal{K}_2) \subset \mathcal{V}_g(\mathcal{K}_1, \mathcal{K}_2)$.

We note that Definition 4.1 is a generalization of a definition by Hautus [5]. His space S_Σ^- (see [5, p. 706]) coincides with $\mathcal{V}_g(\mathcal{K}_1, \mathcal{K}_2)$ if $\mathcal{K}_1$ is taken to be $\mathcal{X}$. The following theorem can be proven to be completely analogous to [5, Thm. 4.3]:

THEOREM 4.2.

$$\mathcal{V}_g(\mathcal{K}_1, \mathcal{K}_2) = \mathcal{V}_g^*(\mathcal{K}_1) + \mathcal{V}^*(\mathcal{K}_2). \qquad \square$$

Note that it follows from the above theorem that $\mathcal{V}_g(\mathcal{K}_1, \mathcal{K}_2)$ is controlled invariant. The next theorem provides the key step in the solution of DDPOS. The result states that what can be done in Definition 4.1 by open loop control can in fact be done by state feedback:

THEOREM 4.3. *There exists a map $F: \mathcal{X} \to \mathcal{U}$ such that*

$$A_F \mathcal{V}_g(\mathcal{K}_1, \mathcal{K}_2) \subset \mathcal{V}_g(\mathcal{K}_1, \mathcal{K}_2), \tag{4.1}$$

$$A_F \mathcal{V}^*(\mathcal{K}_2) \subset \mathcal{V}^*(\mathcal{K}_2), \tag{4.2}$$

$$\sigma(A_F | \mathcal{V}_g(\mathcal{K}_1, \mathcal{K}_2)/\mathcal{V}^*(\mathcal{K}_2)) \subset \mathbb{C}_g. \tag{4.3}$$

Proof. During this proof, denote $\mathcal{V}_g := \mathcal{V}_g(\mathcal{K}_1, \mathcal{K}_2)$. Since $\mathcal{V}^*(\mathcal{K}_2) \subset \mathcal{V}_g$ and since both spaces are controlled invariant, they are compatible (see [22, Ex. 9.1]). Hence, there is a map $F_0: \mathcal{X} \to \mathcal{U}$ such that $A_{F_0}\mathcal{V}^*(\mathcal{K}_2) \subset \mathcal{V}^*(\mathcal{K}_2)$ and $A_{F_0}\mathcal{V}_g \subset \mathcal{V}_g$. Let $\bar{\mathcal{B}} := \mathcal{B} \cap \mathcal{V}_g$ and let V be any matrix such that $\bar{\mathcal{B}} = \operatorname{im} BV$.

Consider the controllability subspace $\langle A_{F_0} | \bar{\mathcal{B}} \rangle$. By the facts that $\bar{\mathcal{B}} \subset \mathcal{V}_g$ and $A_{F_0}\mathcal{V}_g \subset \mathcal{V}_g$, this controllability subspace is contained in $\mathcal{K}_1$. Since any controllability subspace is also a stabilizability subspace, it must be contained in the largest stabilizability subspace $\mathcal{V}_g^*(\mathcal{K}_1)$ in $\mathcal{K}_1$. It then follows that $\bar{\mathcal{B}} \subset \mathcal{V}_g^*(\mathcal{K}_1)$, so

$$\mathcal{B} \cap \mathcal{V}_g^*(\mathcal{K}_1) = \bar{\mathcal{B}}. \tag{4.4}$$

Next, we claim that $\mathcal{V}_g^*(\mathcal{K}_1)$ is A_{F_0}-invariant. First, since it is (A, B)-invariant, we have $A_{F_0}\mathcal{V}_g^*(\mathcal{K}_1) \subset \mathcal{V}_g^*(\mathcal{K}_1) + \mathcal{B}$. On the other hand, $A_{F_0}\mathcal{V}_g^*(\mathcal{K}_1) \subset A_{F_0}\mathcal{V}_g \subset \mathcal{V}_g$. Hence, again using $\mathcal{V}_g^*(K_1) \subset \mathcal{V}_g$, we obtain $A_{F_0}\mathcal{V}_g^*(\mathcal{K}_1) \subset (\mathcal{V}_g^*(\mathcal{K}_1) + \mathcal{B}) \cap \mathcal{V}_g \subset \mathcal{V}_g^*(\mathcal{K}_1) + (\mathcal{B} \cap \mathcal{V}_g) = \mathcal{V}_g^*(\mathcal{K}_1)$. The last equality follows from (4.4).

Using (4.4) and [5, Prop. 2.16], we deduce that the pair $(A_{F_0} | \mathcal{V}_g^*(\mathcal{K}_1), BV)$ is stabilizable.

Let $P_1: \mathcal{V}_g \to \mathcal{V}_g/\mathcal{V}^*(\mathcal{K}_2)$ be the canonical projection. Let $(\bar{A}_{F_0}, \overline{BV})$ be the system induced in $\mathcal{V}_g/\mathcal{V}^*(\mathcal{K}_2)$. It can easily be seen, for example, by using a rank test (see [4] or [5, Thm. 2.13]), that the latter system is stabilizable. Hence, there is a map $\bar{F}_1$ on the factor space such that $\sigma(\bar{A}_{F_0} + \overline{BV}\bar{F}_1) \subset \mathbb{C}_g$. Now, let F_1 be any map on $\mathcal{V}_g$ such that $F_1 = \bar{F}_1 P_1$ and define F_1 arbitrary on a complement of $\mathcal{V}_g$. Define $F := F_0 + VF_1$. Since $F | \mathcal{V}^*(\mathcal{K}_2) = F_0 | \mathcal{V}^*(\mathcal{K}_2)$ ("$|$" denotes "restriction to"), we then have $A_F\mathcal{V}^*(\mathcal{K}_2) \subset \mathcal{V}^*(\mathcal{K}_2)$ and it can be verified that Fig. 1 commutes.

We are now in a position to prove the main result of this section.

THEOREM 4.4. DDPOS *is solvable iff* $\operatorname{im} G \subset \mathcal{V}_g(\mathcal{K}_1, \mathcal{K}_2)$.

Proof. ($\Leftarrow$) Choose F as in Theorem 4.3. Then $\langle A_F | \operatorname{im} G \rangle \subset \mathcal{K}_1$, which yields the decoupling from d to z_1.

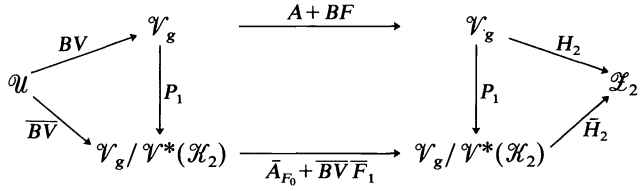


FIG. 1

Let $\bar{H}_2$ be as in the Fig. 1 and let $\bar{A}_F := A_F \mid \mathcal{V}_g(\mathcal{K}_1, \mathcal{K}_2)/\mathcal{V}^*(\mathcal{K}_2)$. Then $H_2(Is - A_F)^{-1}G = \bar{H}_2(Is - \bar{A}_F)^{-1}P_1G$, which is stable since $\sigma(\bar{A}_F) \subset \mathbb{C}_g$.

($\Rightarrow$) *If* F is such that $H_1(Is - A_F)^{-1}G = 0$ and $H_2(Is - A_F)^{-1}G$ is stable then for $d \in \mathcal{D}$ let $\xi(s) := (Is - A_F)^{-1}Gd$ and $\omega(s) := F\xi(s)$. Then clearly $Gd = (Is - A)\xi(s) + B\omega(s)$ and $H_2\xi(s)$ is stable. □

Remark 4.5. If in the above problem we take $H_1 = H_2 = H$, DDPOS reduces to the ordinary disturbance decoupling problem DDP (see [22]). In this case we have, denoting $\mathcal{K} := \ker H$, $\mathcal{V}_g(\mathcal{K}_1, \mathcal{K}_2) = \mathcal{V}_g^*(\mathcal{K}) + \mathcal{V}^*(\mathcal{K}) = \mathcal{V}^*(\mathcal{K})$. If we take $H_1 = 0$ and $H_2 = H$, we arrive at OSDP as studied in Hautus [5]. The solvability of this problem requires the existence of a state feedback F such that $H(Is - A_F)^{-1}G$ is stable. Necessary and sufficient conditions can be found by noting that $\mathcal{V}_g(\mathcal{K}_1, \mathcal{K}_2) = \mathcal{V}_g^*(\mathcal{X}) + \mathcal{V}^*(\mathcal{K})$. As also noted in [5], if we take $H_1 = 0$, $H_2 = H$ and im $G = \mathcal{X}$, the above theorem provides necessary and sufficient conditions for the solvability of the output stabilization problem, OSP.

5. A necessary geometric condition for the solvability of (ADDPBP)$_p$. In this section we shall establish a necessary condition for the solvability of (ADDPBP)$_p$. This condition will be in the form of a subspace inclusion. The proof is rather technical and some of the details are deferred to Appendix A. In the rest of this paper, the stability set will be taken to be $\mathbb{C}_g = \{\lambda \in \mathbb{C} \mid \text{Re } \lambda < 0\}$.

Consider the system $\dot{x} = Ax + Bu$, $z_1 = H_1x$, $z_2 = H_2x$ and assume that (3.1) is satisfied. The following subspace will play an important role in the sequel:

DEFINITION 5.1. $\mathcal{V}_b(\mathcal{K}_1, \mathcal{K}_2)$ will denote the subspace of all $x \in X$ that have a (ξ, ω)-representation with $\xi(\omega) \in \mathcal{K}_1(s)$, $\omega(s) \in \mathcal{U}(s)$ and $H_2\xi(s)$ is proper and stable.

Interpreted in the time domain, $\mathcal{V}_b(\mathcal{K}_1, \mathcal{K}_2)$ consists exactly of those points in $\mathcal{X}$ that can serve as an initial condition for some Bohl distributional trajectory that lies entirely in $\mathcal{K}_1$, while the vector of components of the trajectory modulo $\mathcal{K}_2$ is the sum of a stable regular Bohl function and a Dirac delta.

It follows immediately from the definition and [12, Thm. 4.1] that $\mathcal{V}_b(\mathcal{K}_1, \mathcal{K}_2)$ is contained in $\mathcal{V}_b^*(\mathcal{K}_1)$, the subspace of distributionally weakly unobservable states with respect to the output z_1. It is also immediate that $\mathcal{V}_g(\mathcal{K}_1, \mathcal{K}_2)$ is contained in $\mathcal{V}_b(\mathcal{K}_1, \mathcal{K}_2)$. We are now in a position to state the main result of this section:

THEOREM 5.2. *Fix* $p \in \{1, 2, \infty\}$. *Then the following holds*:

$$\{(\text{ADDPBP})_p \textit{ is solvable}\} \Rightarrow \{\text{im } G \subset \mathcal{V}_b(\mathcal{K}_1, \mathcal{K}_2)\}.$$

In the remainder of this section we will establish a proof of the above theorem. Again, consider the system $\dot{x} = Ax + Bu$, $z_1 = H_1x$, $z_2 = H_2x$. Assume that for $\varepsilon > 0$, $u_\varepsilon(t)$ is a regular Bohl input. Let $x_0 \in X$. Let $z_{1,\varepsilon}(t)$ and $z_{2,\varepsilon}(t)$ be the outputs corresponding to the above input and initial condition $x(0) = x_0$. Denote $\hat{z}_{i,\varepsilon}(s)$ for the corresponding Laplace transforms of $z_{i,\varepsilon}(t)$. We then have the following lemma:

LEMMA 5.3. *Suppose that either of the following conditions is satisfied*:

(i) $\|z_{1,\varepsilon}\|_{L_1} \to 0$ *as* $\varepsilon \to 0$ *and there exists a constant* C *such that* $\|z_{2,\varepsilon}\|_{L_1} \leqq C$ *for all* ε.

(ii) $\hat{z}_{1,\varepsilon}(s)$ *and* $\hat{z}_{2,\varepsilon}(s)$ *are stable for all* ε, $\sup_{\omega\in\mathbb{R}} \|\hat{z}_{1,\varepsilon}(i\omega)\| \to 0$ *as* $\varepsilon \to 0$ *and there exists a constant* C *such that* $\sup_{\omega\in\mathbb{R}} \|\hat{z}_{2,\varepsilon}(i\omega)\| \leqq C$ *for all* ε.
Then $x_0 \in \mathcal{V}_b(\mathcal{K}_1, \mathcal{K}_2)$. □

A detailed proof of Lemma 5.3 can be found in Appendix A. The idea of the proof is the following. First we note that the initial condition x_0 above has for each $\varepsilon > 0$ a (ξ, ω)-representation $x_0 = (Is - A)\xi_\varepsilon(s) + B\omega_\varepsilon(s)$. Here $\omega_\varepsilon(s)$ is the (rational) Laplace transform of $u_\varepsilon(t)$. Using the asymptotic behaviour as described by the condition (i) or (ii) above, we then analyse the limiting behaviour for $\varepsilon \to 0$ of the sequences of rational vectors $\xi_\varepsilon(s)$ and $\omega_\varepsilon(s)$. This will lead to a (ξ, ω)-representation for x_0 with the properties described in Definition 5.1. To conclude this section we apply Lemma 5.3 to obtain the following:

Proof of Theorem 5.2. Assume that $(\text{ADDPBP})_p$ is solvable. Let $x_0 \in \text{im } G$. Let F_ε be as in Lemma 3.1 and define $u_\varepsilon(t) := F_\varepsilon T_\varepsilon(t)x_0$. Then, depending on p, one of the conditions (i) or (ii) in Lemma 5.3 is satisfied. It follows that $x_0 \in \mathcal{V}_b(\mathcal{K}_1, \mathcal{K}_2)$. □

6. The geometric structure of $\mathcal{V}_b(\mathcal{K}_1, \mathcal{K}_2)$. In the sequel, it will turn out that under certain assumptions on the system (1.1) the subspace inclusion in Theorem 5.2 is also a sufficient condition for the solvability of $(\text{ADDPBP})_p$. In order to prove this and to be able to construct the required feedback maps, we need more detailed information on the geometric structure of the subspace $\mathcal{V}_b(\mathcal{K}_1, \mathcal{K}_2)$ as introduced in the previous section. In the present section, we will first show that the subspace $\mathcal{V}_b(\mathcal{K}_1, \mathcal{K}_2)$ can always be written as the sum of the subspace $\mathcal{V}_g(\mathcal{K}_1, \mathcal{K}_2)$ (see § 4) together with an almost controllability subspace depending on $\mathcal{K}_1$ and $\mathcal{K}_2$. Using this result, we will show that if either $\mathcal{R}^*(\mathcal{K}_1) = \{0\}$ or $\mathcal{K}_2 = \{0\}$, then $\mathcal{V}_b(\mathcal{K}_1, \mathcal{K}_2)$ admits a decomposition into the direct sum of $\mathcal{V}_g(\mathcal{K}_1, \mathcal{K}_2)$ together with a number of singly generated almost controllability subspaces, with a particular position with respect to the subspaces $\mathcal{K}_1$ and $\mathcal{K}_2$. The main result of this section will be the following theorem:

THEOREM 6.1. *Assume that* $\mathcal{R}^*(\mathcal{K}_1) = \{0\}$ *or that* $\mathcal{K}_2 = \{0\}$. *Then there is an integer* $m' \in \mathbb{N}$, *there are integers* $r_1, \cdots, r_{m'} \in \mathbb{N}$ *and vectors* $b_1, \cdots, b_{m'} \in \mathcal{B}$ *and there is a map* $F: \mathcal{X} \to \mathcal{U}$ *such that*

$$\mathcal{V}_b(\mathcal{K}_1, \mathcal{K}_2) = \mathcal{V}_g(\mathcal{K}_1, \mathcal{K}_2) \oplus \bigoplus_{i=1}^{m'} \mathcal{L}_i, \tag{6.1}$$

where

$$\mathcal{L}_i := \bigoplus_{j=1}^{r_i} A_F^{j-1} b_i,$$

with

$$\bigoplus_{j=1}^{r_i-1} A_F^{j-1} b_i \subset \mathcal{K}_1 \tag{6.2}$$

and

$$\bigoplus_{j=1}^{r_i-2} A_F^{j-1} b_i \subset \mathcal{K}_2. \tag{6.3}$$

If in the statement of the above theorem one of the integers r_i is such that $r_i - 1 < 1$ or $r_i - 2 < 1$, then the corresponding sums in (6.2) or (6.3) are understood to be equal to $\{0\}$. It will turn out in the proof of Theorem 6.1 that in the case that $\mathcal{R}^*(\mathcal{K}_1) = \{0\}$ the integer m' may be chosen to be equal to m $(= \dim \mathcal{B})$. In the case that $\mathcal{K}_2 = \{0\}$ it will appear that m' may be chosen to be equal to $m - \dim [\mathcal{V}^*(\mathcal{K}_1) \cap \mathcal{B}]$ and also that

in this case the integers r_i may be taken to be either 1 or 2. Since $\mathcal{V}_g(\mathcal{K}_1, \{0\}) = \mathcal{V}_g^*(\mathcal{K}_1)$ (see Theorem 4.2) the theorem thus states that $\mathcal{V}_b(\mathcal{K}_1, \{0\})$ is equal to the direct sum of $\mathcal{V}_g^*(\mathcal{K}_1)$ together with a number of singly generated almost controllability subspaces which are equal to either span $\{b_i\}$ (with $0 \neq b_i \in \mathcal{B}$) or span $\{b_i, A_F b_i\}$, with $\{b_i, A_F b_i\}$ linearly independent and $b_i \in \mathcal{K}_1 \cap \mathcal{B}$.

The result of Theorem 6.1 will be instrumental in the next section, where we will consider the sufficiency of the subspace inclusion im $G \subset \mathcal{V}_b(\mathcal{K}_1, \mathcal{K}_2)$ for solvability of $(\text{ADDPBP})_p$ and propose a "scheme" for calculation of the required feedback maps. In the remainder of the present section we will establish a proof of Theorem 6.1.

We introduce the following subspace:

DEFINITION 6.2. $\mathcal{R}_b(\mathcal{K}_1, \mathcal{K}_2)$ will denote the subspace of all $x \in \mathcal{X}$ that have a (ξ, ω)-representation with $\xi(s) \in \mathcal{K}_1[s]$, $\omega(s) \in \mathcal{U}[s]$ and $H_2\xi(s)$ is *constant* (i.e., if $\xi(s) = \sum_{i=0}^{N} x_i s^i$ then $H_2 x_i = 0$ for $i \geqq 1$).

Interpreted in the time domain, $\mathcal{R}_b(\mathcal{K}_1, \mathcal{K}_2)$ consists exactly of those points in $\mathcal{X}$ that can be driven to 0 along a purely distributional Bohl trajectory that lies entirely in $\mathcal{K}_1$, while the vector of components of this trajectory modulo $\mathcal{K}_2$ is a Dirac delta.

It follows immediately from the definition and Lemma 2.7 that every point in $\mathcal{R}_b(\mathcal{K}_1, \mathcal{K}_2)$ is strongly controllable with respect to the output z_1. Moreover, it is also immediate that every point x that is strongly controllable with respect to the output z_2, is an element of $\mathcal{R}_b(\mathcal{K}_1, \mathcal{K}_2)$. Hence, the inclusion $\mathcal{R}_b^*(\mathcal{K}_2) \subset \mathcal{R}_b(\mathcal{K}_1, \mathcal{K}_2) \subset \mathcal{R}_b^*(\mathcal{K}_1)$ holds. In fact, we have the following nice result:

THEOREM 6.3.

(i) $\mathcal{R}_b(\mathcal{K}_1, \mathcal{K}_2) = \mathcal{B} + A(\mathcal{R}_b^*(\mathcal{K}_2) \cap \mathcal{K}_1)$,

(ii) $\mathcal{V}_b(\mathcal{K}_1, \mathcal{K}_2) = \mathcal{V}_g(\mathcal{K}_1, \mathcal{K}_2) + \mathcal{R}_b(\mathcal{K}_1, \mathcal{K}_2)$.

Proof. (i) Suppose that $x = (Is - A)\xi(s) + B\omega(s)$, with $\xi(s) \in \mathcal{K}_1[s]$, $\omega(s) \in \mathcal{U}[s]$ and $H_2\xi(s)$ is constant. Let $\xi(s) = \sum_{i=0}^{N} x_i s^i$ and $\omega(s) = \sum_{i=0}^{N+1} u_i s^i$. Obviously, $\xi(s) = x_0 + s\xi_1(s)$ and $\omega(s) = u_0 + s\omega_1(s)$, where $\xi_1(s) \in \mathcal{K}_2[s]$ and $\omega_1(s) \in \mathcal{U}[s]$. Hence, $x = Bu_0 - Ax_0 + sx_0 + s^2\xi_1(s) - As\xi_1(s) + Bs\omega_1(s)$ and by equating powers it follows that

$$x = -Ax_0 + Bu_0, \tag{6.4}$$

$$-x_0 = (Is - A)\xi_1(s) + B\omega_1(s). \tag{6.5}$$

Therefore, $x_0 \in \mathcal{R}_b^*(\mathcal{K}_2)$ (see Lemma 2.7). Since also $x_0 \in \mathcal{K}_1$, we obtain $x \in \mathcal{B} + A(\mathcal{R}_b^*(\mathcal{K}_2) \cap \mathcal{K}_1)$. Conversely, if $x = Bu_0 - Ax_0$ with $x_0 \in \mathcal{R}_b^*(\mathcal{K}_2) \cap \mathcal{K}_1$, there is $\xi_1(s) \in \mathcal{K}_2[s]$ and $\omega_1(s) \in \mathcal{U}[s]$ such that $-x_0 = (Is - A)\xi_1(s) + B\omega_1(s)$. Defining then $\xi(s) := x_0 + s\xi_1(s)$ and $\omega(s) := u_0 + s\omega_1(s)$, we obtain a (ξ, ω)-representation of x with $\xi(s) \in \mathcal{K}_1[s]$, $\omega(s) \in \mathcal{U}[s]$ and $H_2\xi(s) = H_2x_0$ is constant.

(ii) Assume that $x \in \mathcal{V}_b(\mathcal{K}_1, \mathcal{K}_2)$. There is a (ξ, ω)-representation for x with $\xi(s) \in \mathcal{K}_1(s)$, $\omega(s) \in \mathcal{U}(s)$ and $H_2\xi(s)$ proper and stable. Decompose $\xi(s) = \xi_1(s) + \xi_2(s)$ and $\omega(s) = \omega_1(s) + \omega_2(s)$, where $\xi_1(s)$ and $\omega_1(s)$ are polynomial vectors and $\xi_2(s)$ and $\omega_2(s)$ are strictly proper. Obviously, $\xi_1(s) \in \mathcal{K}_1[s]$, $\xi_2(s) \in \mathcal{K}_{1,+}(s)$, $\omega_1(s) \in \mathcal{U}[s]$ and $\omega_2(s) \in \mathcal{U}_+(s)$. Moreover, $H_2\xi_1(s)$ is constant and $H_2\xi_2(s)$ is strictly proper and stable. Now, since the left-hand side of this equation is proper and the right-hand side is a polynomial vector, both sides must, in fact, be constant. Thus, there is a vector $x_1 \in \mathcal{X}$ such that $x_1 = (Is - A)\xi_1(s) + B\omega_1(s) = x - (Is - A)\xi_2(s) - B\omega_2(s)$. It follows that $x_1 \in \mathcal{R}_b(\mathcal{K}_1, \mathcal{K}_2)$ and $x - x_1 \in \mathcal{V}_g(\mathcal{K}_1, \mathcal{K}_2)$. Since $x = x_1 + (x - x_1)$, we obtain that $x \in \mathcal{V}_g(\mathcal{K}_1, \mathcal{K}_2) + \mathcal{R}_b(\mathcal{K}_1, \mathcal{K}_2)$. The converse inclusion follows immediately from the definitions. □

The importance of the above theorem is that it shows, together with Theorem 4.2, that

$$\mathcal{V}_b(\mathcal{K}_1, \mathcal{K}_2) = \mathcal{V}_g^*(\mathcal{K}_1) + \mathcal{V}^*(\mathcal{K}_2) + \mathcal{B} + A(\mathcal{R}_b^*(\mathcal{K}_2) \cap \mathcal{K}_1). \tag{6.6}$$

Thus, the space $\mathscr{V}_b(\mathscr{K}_1, \mathscr{K}_2)$ can, in principle, be calculated using existing algorithms. The stabilizability subspace and the controlled invariant subspace appearing in (6.6) can be calculated using the invariant subspace algorithm ISA [22, p. 91] and a construction as in [22, p. 114]. The almost controllability subspace $\mathscr{R}_b^*(\mathscr{K}_2)$ can be calculated using the almost controllability subspace algorithm (ACSA)' [20]. For any fixed subspace $\mathscr{K} \subset \mathscr{X}$, this algorithm is defined by

$$\mathscr{T}^{i+1}(\mathscr{K}) = \mathscr{B} + A(\mathscr{T}^i(\mathscr{K}) \cap \mathscr{K}); \ \mathscr{T}^0(\mathscr{K}) = \{0\}. \tag{6.7}$$

It is well known, see [20], that (6.7) defines a nondecreasing sequence of subspaces which reaches a limit after a finite number of iterations. Moreover, this limit equals $\mathscr{T}^n(\mathscr{K}) = \mathscr{R}_b^*(\mathscr{K})$. In the sequel, denote

$$\mathscr{F}^i(\mathscr{K}_1, \mathscr{K}_2) := \mathscr{T}^i(\mathscr{K}_2) \cap \mathscr{K}_1. \tag{6.8}$$

Using the properties of the sequence $\mathscr{T}^i(\mathscr{K})$ stated above, together with Theorem 6.3, the following result is immediate:

LEMMA 6.4. *$\mathscr{F}^i(\mathscr{K}_1, \mathscr{K}_2)$ is a nondecreasing sequence which reaches a limit after a finite number of iterations. This limit equals $\mathscr{F}^n(\mathscr{K}_1, \mathscr{K}_2) = R_b^*(\mathscr{K}_2) \cap \mathscr{K}_1$. Consequently,*

$$\mathscr{R}_b(\mathscr{K}_1, \mathscr{K}_2) = \mathscr{B} + A\mathscr{F}^n(\mathscr{K}_1, \mathscr{K}_2). \tag{6.9}$$

□

Other properties of the sequence $\mathscr{F}^i(\mathscr{K}_1, \mathscr{K}_2)$ are proven in Lemma B.1, Appendix B. Using these properties, we obtain the following lemma:

LEMMA 6.5. *Assume that $\mathscr{R}^*(\mathscr{K}_1) = \{0\}$. Then there is a chain $\{\mathscr{B}_i\}_{i=1}^n$ in $\mathscr{B}$ and a map $F: \mathscr{X} \to \mathscr{U}$ such that*

$$\mathscr{R}_b(\mathscr{K}_1, \mathscr{K}_2) = \mathscr{B} \oplus A_F\mathscr{B}_1 \oplus \cdots \oplus A_F^n\mathscr{B}_n, \tag{6.10}$$

$$\bigoplus_{i=1}^{n} A_F^{i-1}\mathscr{B}_i \subset \mathscr{K}_1, \tag{6.11}$$

$$\bigoplus_{i=2}^{n} A_F^{i-2}\mathscr{B}_i \subset \mathscr{K}_2, \tag{6.12}$$

$$\dim \mathscr{B}_i = \dim A_F^i\mathscr{B}_i = \dim [\mathscr{F}^i(\mathscr{K}_1, \mathscr{K}_2)/\mathscr{F}^{i-1}(\mathscr{K}_1, \mathscr{K}_2)]. \tag{6.13}$$

Proof. See Appendix B. □

We are now in a position to establish a proof of Theorem 6.1 for the case $\mathscr{R}^*(\mathscr{K}_1) = \{0\}$:

Proof of Theorem 6.1 (*Case* 1: $\mathscr{R}^*(\mathscr{K}_1) = \{0\}$). During this proof we will denote $\mathscr{R}_b(\mathscr{K}_1, \mathscr{K}_2)$ by $\mathscr{R}_b$, $\mathscr{V}_b(\mathscr{K}_1, \mathscr{K}_2)$ by $\mathscr{V}_b$ and $\mathscr{V}_g(\mathscr{K}_1, \mathscr{K}_2)$ by $\mathscr{V}_g$. According to Theorem 6.5 we have that $\mathscr{V}_b = \mathscr{V}_g + \mathscr{R}_b$. We claim that the latter sum is a direct sum. Indeed, this follows immediately from the facts that $\mathscr{V}_g \subset \mathscr{V}^*(\mathscr{K}_1)$ and $\mathscr{R}_b \subset \mathscr{R}_b^*(\mathscr{K}_1)$, while $\mathscr{V}^*(\mathscr{K}_1) \cap \mathscr{R}_b^*(\mathscr{K}_1) = \mathscr{R}^*(\mathscr{K}_1) = \{0\}$ (see Lemma 2.2). By Lemma 6.5 there is a chain $\{\mathscr{B}_i\}_{i=1}^n$ in $\mathscr{B}$ and a map F such that (6.10) to (6.13) hold. Let $\mathscr{B}_l$ be the first subspace in the chain which is not zero, i.e., $\mathscr{B}_l \neq \{0\}$ and $\mathscr{B}_j = \{0\}$ for $j = l+1, \cdots, n$. Choose a basis for $\mathscr{B}$ as follows. First choose a basis $\{b_1, \cdots, b_{d_l}\}$ for $\mathscr{B}_l$. Extend this to a basis $\{b_1, \cdots, b_{d_l}, b_{d_l+1}, \cdots, b_{d_{l-1}}\}$ for $\mathscr{B}_{l-1}$ (here, $d_i := \dim \mathscr{B}_i$). Continue this procedure until we have a basis for $\mathscr{B}$.

By the fact that $\dim \mathscr{B}_i = \dim A_F^i\mathscr{B}_i$, $\forall_i$, the following vectors form a basis for $\mathscr{R}_b$:

$$\begin{array}{l}
A_F^l b_1, \cdots, A_F^l b_{d_l}, \\
A_F^{l-1} b_1, \cdots, A_F^{l-1} b_{d_l}, A_F^{l-1} b_{d_l+1}, \cdots, A_F^{l-1} b_{d_{l-1}}, \\
\quad \vdots \qquad\qquad \vdots \qquad\quad \vdots \\
A_F b_1, \cdots, \quad A_F b_{d_l}, \quad A_F b_{d_l+1}, \cdots, \quad A_F b_{d_{l-1}}, \cdots, \cdots, A_F b_{d_1}, \\
b_1, \cdots, b_{d_l}, b_{d_l+1}, \cdots, b_{d_{l-1}}, \cdots, \cdots, b_{d_1}, \cdots, b_m.
\end{array}$$

It may immediately be verified that the above list of vectors can be rearranged to obtain m subspaces $\mathscr{L}_i := \operatorname{span}\{b_i, A_F b_i, \cdots, A_F^{r_i-1} b_i\}$, with

$$\operatorname{span}\{b_i, A_F b_i, \cdots, A_F^{r_i-2} b_i\} \subset \mathscr{K}_1$$

and

$$\operatorname{span}\{b_i, A_F b_i, \cdots, A_F^{r_i-3} b_i\} \subset \mathscr{K}_2.$$

This completes the proof of Theorem 6.1 for the case that $\mathscr{R}^*(\mathscr{K}_1) = \{0\}$. □

In the remainder of this section, we will set up a proof of Theorem 6.1, the case that $\mathscr{K}_2 = \{0\}$. In the following, let $\tilde{\mathscr{B}}$ be a subspace of $\mathscr{B}$ such that $\tilde{\mathscr{B}} \oplus [\mathscr{B} \cap \mathscr{V}^*(\mathscr{K}_1)] = \mathscr{B}$. Let W be a map such that $\tilde{\mathscr{B}} = \operatorname{im} BW$ and let $\tilde{\mathscr{R}}_b^*(\mathscr{K}_1) := \tilde{\mathscr{B}} + A\tilde{\mathscr{R}}_a^*(\mathscr{K}_1)$, where $\tilde{\mathscr{R}}_a^*(\mathscr{K}_1)$ denotes the supremal almost controllability subspace contained in $\mathscr{K}_1$ with respect to the system (A, BW) (see also Lemma 2.3). Define

$$\tilde{\mathscr{W}}(\mathscr{K}_1) := \tilde{\mathscr{B}} + A(\tilde{\mathscr{B}} \cap \mathscr{K}_1). \tag{6.14}$$

We will show that if $\mathscr{K}_2 = \{0\}$, then $\mathscr{V}_b(\mathscr{K}_1, \mathscr{K}_2)$ has a decomposition into the direct sum of $\mathscr{V}_g(\mathscr{K}_1, \mathscr{K}_2)$ (which, in that case, is equal to $\mathscr{V}_g^*(\mathscr{K}_1)$) and the subspace $\tilde{\mathscr{W}}(\mathscr{K}_1)$:

LEMMA 6.6. *Let $\mathscr{K}_1$ be a subspace of $\mathscr{X}$. Then*

$$\mathscr{V}_b(\mathscr{K}_1, \{0\}) = \mathscr{V}_g(\mathscr{K}_1, \{0\}) \oplus \tilde{\mathscr{W}}(\mathscr{K}_1). \tag{6.15}$$

Proof. In this proof, denote $\mathscr{V}_g := \mathscr{V}_g(\mathscr{K}_1, \{0\})$. Also, let $\mathscr{B}_1 := \mathscr{B} \cap \mathscr{V}^*(\mathscr{K}_1)$. Since $\mathscr{R}_b^*(\{0\}) = \mathscr{B}$, it follows from Theorem 6.3 that

$$\begin{aligned}\mathscr{V}_b(\mathscr{K}_1, \{0\}) &= \mathscr{V}_g + \mathscr{B} + A[\mathscr{B} \cap \mathscr{K}_1] \\ &= \mathscr{V}_g + \mathscr{B} + A[(\mathscr{B}_1 \oplus \tilde{\mathscr{B}}) \cap \mathscr{K}_1] \\ &= \mathscr{V}_g + \mathscr{B} + A[\mathscr{B}_1 + (\tilde{\mathscr{B}} \cap \mathscr{K}_1)].\end{aligned}$$

Now, note that $\mathscr{B}_1 \subset \mathscr{R}^*(\mathscr{K}_1)$ (see [22, Thm. 5.5]). Consequently, $A\mathscr{B}_1 \subset \mathscr{R}^*(\mathscr{K}_1) + \mathscr{B} \subset \mathscr{V}_g + \mathscr{B}$. Hence we find

$$\begin{aligned}\mathscr{V}_b(\mathscr{K}_1, \{0\}) &= \mathscr{V}_g + \mathscr{B} + A(\tilde{\mathscr{B}} \cap \mathscr{K}_1) \\ &= \mathscr{V}_g + \mathscr{B}_1 + \tilde{\mathscr{B}} + A(\tilde{\mathscr{B}} \cap \mathscr{K}_1).\end{aligned}$$

Again, by the fact that $\mathscr{B}_1 \subset \mathscr{R}^*(\mathscr{K}_1) \subset \mathscr{V}_g$, we have

$$\mathscr{V}_b(\mathscr{K}_1, \{0\}) = \mathscr{V}_g + \tilde{\mathscr{B}} + A(\tilde{\mathscr{B}} \cap \mathscr{K}_1).$$

Finally, since $\mathscr{V}_g \subset \mathscr{V}^*(\mathscr{K}_1)$ and $\tilde{\mathscr{W}}(\mathscr{K}_1) \subset \tilde{\mathscr{R}}_b^*(\mathscr{K}_1)$, it follows from Lemma 2.3 that the sum in (6.15) is direct. □

Using the above lemma we may now obtain the following proof of Theorem 6.1, the case that $\mathscr{K}_2 = \{0\}$;

Proof of Theorem 6.1 (*Case* 2: $\mathscr{K}_2 = \{0\}$). We claim that $\tilde{\mathscr{W}}(\mathscr{K}_1) = \tilde{\mathscr{B}} \oplus A(\tilde{\mathscr{B}} \cap \mathscr{K}_1)$. To prove this, assume that there is a vector $0 \neq x \in \tilde{\mathscr{B}}$ such that $x = A\bar{x}$, with $\bar{x}$ a vector in $\tilde{\mathscr{B}} \cap \mathscr{K}_1$. Define $\mathscr{V} := \operatorname{span}\{\bar{x}\}$. Since $A\mathscr{V} \subset \mathscr{V} + \mathscr{B}$, $\mathscr{V}$ is controlled invariant. Since also $\mathscr{V} \subset \mathscr{K}_1$, we find that $\mathscr{V} \subset \mathscr{V}^*(\mathscr{K}_1)$. It follows that $\bar{x} \in \mathscr{V}^*(\mathscr{K}_1) \cap \tilde{\mathscr{B}} = \{0\}$ and hence that $x = 0$. This yields a contradiction. Next, we claim that $\dim \tilde{\mathscr{B}} \cap \mathscr{K}_1 = \dim A(\tilde{\mathscr{B}} \cap \mathscr{K}_1)$. Assume the contrary. Then we may find a vector $0 \neq x \in \tilde{\mathscr{B}} \cap \mathscr{K}_1$ such that $Ax = 0$. It follows that $\operatorname{span}\{x\}$ is a controlled invariant subspace contained in $\mathscr{K}_1$ and hence that $x \in \mathscr{V}^*(\mathscr{K}_1) \cap \tilde{\mathscr{B}} = \{0\}$. Again, this is a contradiction. Now, choose a basis for $\tilde{\mathscr{W}}(\mathscr{K}_1)$ as follows: first choose a basis $b_1, \cdots, b_r$ of $\tilde{\mathscr{B}} \cap \mathscr{K}_1$. Extend this to a basis $\{b_1, \cdots, b_r, b_{r+1}, \cdots, b_{m'}\}$ of $\tilde{\mathscr{B}}$. By the above, the vectors $\{b_1, \cdots, b_r, Ab_1, \cdots, Ab_r, b_{r+1}, \cdots, b_{m'}\}$ form a basis for $\tilde{\mathscr{W}}(\mathscr{K}_1)$. These vectors can be rearranged

into one- and two-dimensional singly generated almost controllability subspaces with the properties (6.2) and (6.3). This completes the proof of Theorem 6.1. □

7. The main result. In the present section we will combine our results of the previous sections to show that if the system (1.1) is such that it satisfies at least one of the following two properties:

(7.1) the system (A, B, H_1) is left-invertible,

(7.2) the mapping H_2 is injective,

then the subspace inclusion $\text{im } G \subset \mathcal{V}_b(\mathcal{K}_1, \mathcal{K}_2)$ is both a necessary and sufficient condition for solvability of the L_p-almost disturbance decoupling problem with bounded peaking $(\text{ADDPBP})_p$ for the values $p=1$, $p=2$ and $p=\infty$.

Recall from § 5 that for these values of p the latter subspace inclusion was already shown to be a necessary condition without the extra assumptions (7.1), (7.2). Here we shall, in fact, prove that if either (7.1) or (7.2) holds then $\text{im } G \subset \mathcal{V}_b(\mathcal{K}_1, \mathcal{K}_2)$ is a sufficient condition for solvability of $(\text{ADDPBP})_p$ *for all* $1 \leqq p \leqq \infty$.

The following result is the main result of this paper:

THEOREM 7.1. *Assume that at least one of the two conditions* (7.1), (7.2) *is satisfied. Let* $p \in \{1, 2, \infty\}$. *Then* $(\text{ADDPBP})_p$ *is solvable if and only if* $\text{im } G \subset \mathcal{V}_b(\mathcal{K}_1, \mathcal{K}_2)$. □

In order to obtain a proof of the latter statement, we will prove the following:

LEMMA 7.2. *Assume that at least one of the two conditions* (7.1), (7.2) *is satisfied. Let* $T_\varepsilon(t)$ *and* $\hat{T}_\varepsilon(s)$ *be defined by* (3.4) *and* (3.5). *Then the following statements are equivalent*:

(i) *There exists a constant* C *and a sequence* $\{F_\varepsilon;\ \varepsilon > 0\}$ *such that* $\|H_1 T_\varepsilon G\|_{L_1} \to 0$ $(\varepsilon \to 0)$ *and* $\|H_2 T_\varepsilon G\|_{L_1} \leqq C\ \forall \varepsilon$.

(ii) *There exists a constant* C *and a sequence* $\{F_\varepsilon;\ \varepsilon > 0\}$ *such that, for all* ε, $H_1 \hat{T}_\varepsilon G$ *and* $H_2 \hat{T}_\varepsilon G$ *are stable and* $\sup_{\omega \in \mathbb{R}} \|H_1 \hat{T}_\varepsilon(i\omega) G\| \to 0$ $(\varepsilon \to 0)$ *and* $\sup_{\omega \in \mathbb{R}} \|H_2 \hat{T}_\varepsilon(i\omega) G\| \leqq C$, $\forall \varepsilon$.

(iii) $\text{im } G \subset \mathcal{V}_b(\mathcal{K}_1, \mathcal{K}_2)$.

Note that the implications (i)⇒(iii) and (ii)⇒(iii) follow immediately from Lemma 5.3. Also note that once we have proven the above lemma, a proof of our main result Theorem 7.1 may be obtained by combining Theorem 5.2 and Lemma 3.1. We stress that the implications (iii)⇒(i) in the above, in fact, yields sufficiency of the subspace inclusion $\text{im } G \subset \mathcal{V}_b(\mathcal{K}_1, \mathcal{K}_2)$ for solvability of $(\text{ADDPBP})_p$ *for all* $1 \leqq p \leqq \infty$.

The idea of the proof of the implication (iii)⇒(i) of Lemma 7.2 is as follows. First we note that left-invertibility of the system (A, B, H_1) is equivalent to the condition $\mathcal{R}^*(\mathcal{K}_1) = \{0\}$ (Lemma 2.6), while injectivity of the map H_2 is equivalent to $\mathcal{K}_2 = \{0\}$. Thus, under the assumptions of Lemma 7.2, $\mathcal{V}_b(\mathcal{K}_1, \mathcal{K}_2)$ may be decomposed according to (6.1), (6.2) and (6.3). Each of the singly generated almost controllability subspaces L_i appearing in this decomposition will then be approximated by sequences of controlled invariant subspaces $\{\mathcal{L}_{j\varepsilon};\ \varepsilon > 0\}$. If we then define $\mathcal{V}_\varepsilon := \mathcal{V}_g \oplus \bigoplus_{j=1}^{m'} \mathcal{L}_{j\varepsilon}$, the sequence $\{\mathcal{V}_\varepsilon;\ \varepsilon > 0\}$ will converge to $\mathcal{V}_b(\mathcal{K}_1, \mathcal{K}_2)$. In this sense, $\text{im } G$ is "almost contained" in the controlled invariant subspace $\mathcal{V}_\varepsilon$. The subspace $\mathcal{V}_\varepsilon$ in turn is "almost contained" in $\mathcal{K}_1$ (where the latter "almost" should be interpreted in the L_1-sense, see also [20]), while its distance from $\mathcal{K}_2$ is uniformly bounded with respect to ε. Using the structure of the $\mathcal{L}_{j\varepsilon}$ above, we will construct a particular sequence of feedback maps $\{F_\varepsilon;\ \varepsilon > 0\}$ such that $(A + BF_\varepsilon)\mathcal{V}_\varepsilon \subset \mathcal{V}_\varepsilon$. Finally it will be shown that this sequence has the properties required by (i) and (ii) in Lemma 7.2. To start with, we will show how a singly generated almost controllability subspace can be approximated by controlled invariant subspaces. Let $b \in \mathcal{B}$ and let $\mathcal{L} := \ell \oplus \cdots \oplus A_F^{k-1}\ell$. For $i \in \underline{k}$ and $\varepsilon > 0$,

define vectors in $\mathscr{X}$ by

$$x_1(\varepsilon) := (I + \varepsilon A_F)^{-1} b, \qquad x_i(\varepsilon) := (I + \varepsilon A_F)^{-1} A_F x_{i-1}(\varepsilon). \tag{7.3}$$

Note that the matrix inversions in the above expressions are defined for ε sufficiently small. Moreover, it can be seen immediately that $x_i(\varepsilon) \to A_F^{i-1} b (\varepsilon \to 0)$. Thus it follows from Lemma 2.8 that for ε sufficiently small, the vectors $\{x_i(\varepsilon), i \in \underline{k}\}$ are linearly independent. For each ε, define a subspace $\mathscr{L}_\varepsilon$ by

$$\mathscr{L}_\varepsilon := \operatorname{span}\{x_1(\varepsilon), \cdots, x_k(\varepsilon)\}. \tag{7.4}$$

Assume $u \in \mathscr{U}$ is such that $b = Bu$ and define a map $F_\varepsilon : \mathscr{L}_\varepsilon \to U$ by

$$F_\varepsilon x_i(\varepsilon) := -\varepsilon^{-i} u, \qquad (i \in \underline{k}). \tag{7.5}$$

The main properties of the sequences $\{\mathscr{L}_\varepsilon; \varepsilon > 0\}$ and $\{F_\varepsilon; \varepsilon > 0\}$ are summarized in the following lemma:

LEMMA 7.3. *For $i \in \underline{k}$ we have $x_i(\varepsilon) \to A_F^{i-1} b$ as $\varepsilon \to 0$. Consequently, $\mathscr{L}_\varepsilon \to \mathscr{L}$. Each $\mathscr{L}_\varepsilon$ is controlled invariant and, with F_ε defined by* (7.5), *$(A_F + BF_\varepsilon)\mathscr{L}_\varepsilon \subset \mathscr{L}_\varepsilon$. Moreover, a matrix of $(A_F + BF_\varepsilon)|\mathscr{L}_\varepsilon$ is given by*

$$M_\varepsilon := -\begin{pmatrix} \varepsilon^{-1} & \varepsilon^{-2} & \cdots & \varepsilon^{-k} \\ 0 & \varepsilon^{-1} & \ddots & \\ \vdots & \ddots & \ddots & \varepsilon^{-2} \\ 0 & \cdots & 0 & \varepsilon^{-1} \end{pmatrix}. \tag{7.6}$$

Finally, for each ε, $\mathscr{L}_\varepsilon \subset \langle A | \mathscr{B} \rangle$, the reachable subspace of (A, B).

Proof. The claim $x_i(\varepsilon) \to A_F^{i-1} b$ is immediate. Since the vectors $A_F^{i-1} b$ are a basis for $\mathscr{L}$, it follows from §§ 2 and 4 that $\mathscr{L}_\varepsilon \to \mathscr{L}$. Using (7.3) and (7.5), it may be verified by straightforward calculation that $(A_F + BF_\varepsilon)x_i(\varepsilon) = -\sum_{j=1}^{i} \varepsilon^{j-i-1} x_j(\varepsilon)$. It follows that $\mathscr{L}_\varepsilon$ is indeed $A_F + BF_\varepsilon$-invariant and that a matrix of the map restricted to $\mathscr{L}_\varepsilon$ is given by (7.6). Finally, to prove that $\mathscr{L}_\varepsilon$ is contained in the reachable subspace, make a Taylor expansion to find that $(I + \varepsilon A_F)^{-1} = \sum_{m=0}^{\infty} (-\varepsilon)^m A_F^m$. It then follows immediately that $x_1(\varepsilon) \in \langle A_F | \mathscr{B} \rangle$ for all ε. The same follows for $x_2(\varepsilon)$, $x_3(\varepsilon)$ etc.

We note that a slightly different construction leading to an approximating sequence for a singly generated controllability subspace was described in [13]. The construction described by us however exhibits an important property which will be formulated in the following lemma. The proof of this result is straightforward but rather technical and will be deferred to Appendix C.

LEMMA 7.4. *Let $\mathscr{L} := \bigoplus_{i=1}^{k} A_F^{i-1} \mathscr{b}$ be such that $\bigoplus_{i=2}^{k} A_F^{i-2} \mathscr{b} \subset \mathscr{K}_1$ and $\bigoplus_{i=3}^{k} A_F^{i-3} \mathscr{b} \subset \mathscr{K}_2$. Let $x_i(\varepsilon)$ and F_ε be as defined above. Then the following holds: there is constant C such that for all $i \in \underline{k}$:*

$$\| H_1 e^{(A_F + BF_\varepsilon)t} x_i(\varepsilon) \|_{L_1} \to 0 \quad \text{as } \varepsilon \to 0, \tag{7.7}$$

$$\| H_2 e^{(A_F + BF_\varepsilon)t} x_i(\varepsilon) \|_{L_1} \leqq C \quad \text{for all } \varepsilon. \tag{7.8}$$

□

Now, in order to complete a proof of Lemma 7.2, we need one more preliminary result. Up to now we have constructed a sequence of controlled invariant subspaces converging to a singly generated almost controllability subspace and defined a feedback map on each of these controlled invariant subspaces. By applying the decomposition theorem, Theorem 6.1, and applying the above construction to each $\mathscr{L}_j$ appearing in (6.1), we can find a sequence of controlled invariant subspaces $\mathscr{R}_\varepsilon$ converging to $\bigoplus_{i=1}^{m'} \mathscr{L}_i$. In the obvious way we can define a map F_ε on $\mathscr{R}_\varepsilon$. Now the question is, can we define F_ε appropriately on a subspace complementary to $\mathscr{R}_\varepsilon$? The next construction

theorem states that, indeed, we can. It is here that we will use the results on exact disturbance decoupling with stability constraints from § 4. In the following, $\mathcal{V}_b := \mathcal{V}_b(\mathcal{K}_1, \mathcal{K}_2)$, $\mathcal{V}_g := \mathcal{V}_g(\mathcal{K}_1, \mathcal{K}_2)$ and $\mathcal{R}_b := \mathcal{R}_b(\mathcal{K}_1, \mathcal{K}_2)$ are denoted:

THEOREM 7.5. *Consider the system* (1.1). *Let* Λ *be a symmetric set of* $\dim[(\langle A \mid \mathcal{B}\rangle + \mathcal{V}_g)/\mathcal{V}_b]$ *complex numbers. Then there is a map* $F_1 : \mathcal{X} \to \mathcal{U}$ *and a subspace* $\mathcal{Z} \subset \mathcal{X}$ *such that the following conditions are satisfied*:

$$(A + BF_1)\mathcal{V}_g \subset \mathcal{V}_g, \tag{7.9}$$

$$(A + BF_1)\mathcal{V}^*(\mathcal{K}_2) \subset \mathcal{V}^*(\mathcal{K}_2), \tag{7.10}$$

$$\sigma(A + BF_1 \mid \mathcal{V}_g/\mathcal{V}^*(\mathcal{K}_2)) \subset \mathbb{C}_g, \tag{7.11}$$

$$\mathcal{V}_b \oplus \mathcal{Z} = \mathcal{V}_g + \langle A \mid \mathcal{B}\rangle, \tag{7.12}$$

$$(A + BF_1)(\mathcal{V}_g \oplus \mathcal{Z}) \subset \mathcal{V}_g \oplus \mathcal{Z}, \tag{7.13}$$

$$\sigma(A + BF_1 \mid (\mathcal{V}_g \oplus \mathcal{Z})/\mathcal{V}_g) = \Lambda. \tag{7.14}$$

Proof. Let $F_0 : \mathcal{X} \to \mathcal{U}$ be a map that satisfies the conditions (4.1), (4.2) and (4.3) of Theorem 4.3. Let $P : \mathcal{X} \to \mathcal{X}/\mathcal{V}_g$ be the canonical projection. Let $(\bar{A}_{F_0}, \bar{B})$ be the system induced by (A_{F_0}, B) in the factor space $\mathcal{X}/\mathcal{V}_g$. Since $\mathcal{V}_b = \mathcal{V}_g + \mathcal{R}_b$ and $\ker P = \mathcal{V}_g$, we have $P\mathcal{V}_b = P\mathcal{R}_b$. By Lemma 2.4 and the fact that $\mathcal{R}_b$ is an almost controllability subspace, it follows that $P\mathcal{R}_b$ is an almost controllability subspace with respect to the system $(\bar{A}_{F_0}, \bar{B})$. By [22, Prop. 1.2], $P\langle A \mid \mathcal{B}\rangle = \langle \bar{A}_{F_0} \mid \operatorname{im} \bar{B}\rangle$. Let Λ be as above. It can easily be verified that $\#\Lambda = \dim[\langle \bar{A}_{F_0} \mid \operatorname{im} \bar{B}\rangle / P\mathcal{R}_b]$. Thus, we may apply Proposition 2.5 to find an $(\bar{A}_{F_0}, \bar{B})$-invariant subspace $\bar{\mathcal{Z}} \subset \mathcal{X}/\mathcal{V}_g$ and a map $\bar{F} : \mathcal{X}/\mathcal{V}_g \to \mathcal{U}$ such that

$$P\mathcal{R}_b \oplus \bar{\mathcal{Z}} = \langle \bar{A}_{F_0} \mid \operatorname{im} \bar{B}\rangle, \tag{7.15}$$

$$(\bar{A}_{F_0} + \bar{B}\bar{F})\bar{\mathcal{Z}} \subset \bar{\mathcal{Z}}, \tag{7.16}$$

$$\sigma(\bar{A}_{F_0} + \bar{B}\bar{F} \mid \bar{\mathcal{Z}}) = \Lambda. \tag{7.17}$$

Now let $\mathcal{Z} \subset \mathcal{X}$ be any subspace such that $P\mathcal{Z} = \bar{\mathcal{Z}}$ and $\mathcal{Z} \cap \mathcal{V}_g = \{0\}$. Define a map $F_1 : \mathcal{X} \to \mathcal{U}$ by $F_1 := F_0 + \bar{F}P$. We contend that the subspace $\mathcal{Z}$ and the map F_1 satisfy the claims of the theorem. To prove (7.9) to (7.11), note that $F_1 \mid \mathcal{V}_g = F_0 \mid \mathcal{V}_g$. The claim (7.12) can be proven as follows: From (7.15) we have $P(\mathcal{V}_b + \mathcal{Z}) = P\langle A \mid \mathcal{B}\rangle$. Hence, since $\mathcal{V}_g \subset \mathcal{V}_b$, $\mathcal{V}_b + \mathcal{Z} = \mathcal{V}_g + \langle A \mid \mathcal{B}\rangle$. Assume $x \in \mathcal{V}_b \cap \mathcal{Z}$. Then $Px \in P\mathcal{V}_b \cap \bar{\mathcal{Z}} = \{0\}$. Thus, $x \in \ker P \cap \mathcal{Z} = \mathcal{V}_g \cap \mathcal{Z} = \{0\}$. It follows that $\mathcal{V}_b + \mathcal{Z} = \mathcal{V}_b \oplus \mathcal{Z}$.

To prove (7.13), note by using (7.16) that $P(A + BF_1)(\mathcal{V}_g \oplus \mathcal{Z}) = P(A_{F_0} + B\bar{F}P)(\mathcal{V}_g \oplus \mathcal{Z}) = (\bar{A}_{F_0} + \bar{B}\bar{F}_1)\bar{\mathcal{Z}} \subset \bar{\mathcal{Z}} = P(\mathcal{V}_g \oplus \mathcal{Z})$. Finally, (7.14) follows immediately from (7.17). □

We are now in a position to complete the proof of Lemma 7.2:

Proof of Lemma 7.2. (i)⇒(ii). This follows immediately from the fact that the L_2-induced norm of a convolution operator is bounded from above by the L_1-norm of its kernel (see, for example, [2]).

(iii)⇒(i). In this part we will construct a sequence of feedback maps $\{F_\varepsilon;\ \varepsilon > 0\}$ such that, for each $x \in \mathcal{V}_b$, $\|H_1 T_\varepsilon x\|_{L_1} \to 0$ and $\|H_2 T_\varepsilon x\|_{L_1} \leqq C$ for all ε, for some constant C. The construction is divided into five steps:

1. *Decomposition.* Apply Theorem 6.1 to find a decomposition $\mathcal{V}_b = \mathcal{V}_g \oplus \bigoplus_{j=1}^{m'} \mathcal{L}_j$, with $\mathcal{L}_j = \bigoplus_{i=1}^{r_j} A_F^{i-1} b_j$, such that (6.2) and (6.3) hold.

2. *Approximation of singly generated controllability subspaces.* For each $\mathcal{L}_j$, apply the construction (7.3) to (7.6). Thus we find vectors $x_i^{(j)}(\varepsilon)$ $(i \in \underline{r}_j)$, subspaces $\mathcal{L}_{j\varepsilon} := \operatorname{span}\{x_i^{(j)}(\varepsilon);\ i \in \underline{r}_j\}$ and maps $F_{j\varepsilon} : \mathcal{L}_{j\varepsilon} \to \mathcal{U}$ such that

$$x_i^{(j)}(\varepsilon) \to A_F^{i-1} b_j; \qquad \mathcal{L}_{j\varepsilon} \to \mathcal{L}_j (\varepsilon \to 0). \tag{7.18}$$

Moreover, by applying Lemma 7.4, there are constants C_j such that

$$\|H_1 e^{(A_F+BF_{j\varepsilon})t}x_i^{(j)}(\varepsilon)\|_{L_1} \to 0 (\varepsilon \to 0), \tag{7.19}$$

$$\|H_2 e^{(A_F+BF_{j\varepsilon})t}x_i^{(j)}(\varepsilon)\|_{L_1} \leqq C \quad \text{for all } \varepsilon. \tag{7.20}$$

3. *Composition.* Since the $\mathscr{L}_j$ are independent, it follows from Lemma 2.8 that for ε sufficiently small the $\mathscr{L}_{j\varepsilon}$ ($j \in \underline{m}'$) are independent. Define $\mathscr{R}_\varepsilon := \mathscr{L}_{1\varepsilon} \oplus \cdots \oplus \mathscr{L}_{m'\varepsilon}$. It follows that $\mathscr{R}_\varepsilon \to \bigoplus_j \mathscr{L}_j$. Define now $F_\varepsilon^0 : \mathscr{R}_\varepsilon \to \mathscr{U}$ by defining $F_\varepsilon^0 | \mathscr{L}_{j\varepsilon} := (F + F_{j\varepsilon}) | \mathscr{L}_{j\varepsilon}$ ($j \in \underline{m}'$).

4. *Construction of feedback outside $\mathscr{R}_\varepsilon$.* To define a map on a complement of $\mathscr{R}_\varepsilon$, let $\Lambda \subset \mathbb{C}_g$ be a symmetric set of $\dim [(\langle A | \mathscr{B} \rangle + \mathscr{V}_g)/\mathscr{V}_b]$ complex numbers and apply the construction theorem Theorem 7.5 to find a subspace $\mathscr{Z} \subset \mathscr{X}$ and a map $F_1 : \mathscr{X} \to \mathscr{U}$ such that (7.9) to (7.14) are satisfied. In the remainder of this proof, denote $\bigoplus_{j=1}^{m'} \mathscr{L}_j$ by $\tilde{\mathscr{R}}_b$. We may then prove the following:

LEMMA 7.6. *For all ε sufficiently small the following holds*:

$$\mathscr{V}_g \oplus \tilde{\mathscr{R}}_b \oplus \mathscr{Z} = \mathscr{V}_g \oplus \mathscr{R}_\varepsilon \oplus \mathscr{Z}. \tag{7.21}$$

Proof. By Lemma 6.6, $\mathscr{V}_b \oplus \mathscr{Z} = \mathscr{V}_g \oplus \tilde{\mathscr{R}}_b \oplus \mathscr{Z}$. Since, for each ε, $\mathscr{R}_\varepsilon \subset \langle A | \mathscr{B} \rangle$ (Lemma 7.3), it follows from (7.12) that $\mathscr{R}_\varepsilon \subset \mathscr{V}_g \oplus \tilde{\mathscr{R}}_b \oplus \mathscr{Z}$. Since $\mathscr{R}_\varepsilon \to \tilde{\mathscr{R}}_b$, we obtain from Lemma 2.8 that $\mathscr{R}_\varepsilon \cap (\mathscr{V}_g \oplus \mathscr{Z}) = \{0\}$ for ε sufficiently small. The equality (7.21) now follows immediately by noting that for ε sufficiently small $\dim \mathscr{R}_\varepsilon = \dim \tilde{\mathscr{R}}_b$.

5. *Definition of the sequence* $\{F_\varepsilon; \varepsilon > 0\}$. Let $\mathscr{W}$ be an arbitrary subspace of $\mathscr{X}$ such that $\mathscr{X} = \mathscr{V}_g \oplus \mathscr{R}_\varepsilon \oplus \mathscr{Z} \oplus \mathscr{W}$. In this ($\varepsilon$-dependent) decomposition of $\mathscr{X}$ define $F_\varepsilon : \mathscr{X} \to \mathscr{U}$ by $F_\varepsilon | \mathscr{V}_g \oplus \mathscr{Z} := F_1 | \mathscr{V}_g \oplus \mathscr{Z}$, $F_\varepsilon | \mathscr{R}_\varepsilon = F_\varepsilon^0 | \mathscr{R}_\varepsilon$ and F_ε arbitrary on $\mathscr{W}$.

We contend that the sequence $\{F_\varepsilon; \varepsilon > 0\}$ defined in this way satisfies the condition (i) of Lemma 7.2. To prove this, first let $x \in \mathscr{V}_g$. Since $F_\varepsilon | \mathscr{V}_g = F_1 | \mathscr{V}_g$, we have by (7.9) and the fact that $\mathscr{V}_g \subset \mathscr{K}_1$ that $x \in \langle A_{F_\varepsilon} | \mathscr{V}_g \rangle \subset \mathscr{K}_1$ for all ε. Thus, for all ε, $H_1 T_\varepsilon(t)x = 0$ for all t. Let $\bar{A}_{F_1}$ and $\bar{H}_2$ be defined by the following commutative diagram (Fig. 2), in which P_1 is the canonical projection:

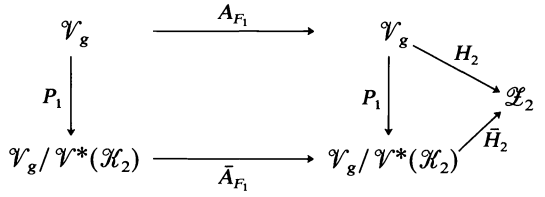


FIG. 2

We then have $H_2 T_\varepsilon(t)x = \bar{H}_2 e^{\bar{A}_{F_1}t}x$ for all ε. It follows from (7.11) that $H_2 T_\varepsilon x$ is in $L_1[0, \infty)$ with, obviously, L_1-norm independent of ε. To complete the proof it now suffices to show that for all $x \in \tilde{\mathscr{R}}_b$, $\|H_1 T_\varepsilon x\|_{L_1} \to 0$ and $\|H_2 T_\varepsilon x\|_{L_1}$ is uniformly bounded with respect to ε. Since $\tilde{\mathscr{R}}_b$ is spanned by the vectors $A_F^{s-1} b_l$, it suffices to take $x = A_F^{s-1} b_l$. By (7.21), we have $\tilde{\mathscr{R}}_b \subset \mathscr{V}_g \oplus \mathscr{R}_\varepsilon \oplus \mathscr{Z}$. Thus, there are vectors $v(\varepsilon) \in \mathscr{V}_g \oplus \mathscr{Z}$ and coefficients $\tau_{ij}(\varepsilon) \in \mathbb{R}$ such that

$$A_F^{s-1} b_l = v(\varepsilon) + \sum_{j=1}^{m'} \sum_{i=1}^{r_j} \tau_{ij}(\varepsilon) x_i^{(j)}(\varepsilon). \tag{7.22}$$

Since $x_i^{(j)}(\varepsilon) \to A_F^{i-1} b_j$, it can be proven by standard means that $v(\varepsilon) \to 0$, $\tau_{ij}(\varepsilon) \to 0$

$(i,j) \neq (s,l)$ and that $\tau_{sl}(\varepsilon) \to 1 (\varepsilon \to 0)$. Now, for $\alpha = 1, 2$ we have

$$\|H_\alpha T_\varepsilon A_F^{s-1} b_l\|_{L_1} \leqq \|H_\alpha e^{(A+BF_1)t} v(\varepsilon)\|_{L_1} + \sum_{j=1}^{m'} \sum_{i=1}^{r_j} |\tau_{ij}(\varepsilon)| \|H_\alpha e^{(A_F+BF_{j\varepsilon})t} x_i^{(j)}(\varepsilon)\|_{L_1}. \tag{7.23}$$

By (7.19), note that for $\alpha = 1$ the last term converges to 0 as $\varepsilon \to 0$. Using (7.20), it follows that for $\alpha = 2$ the last term is bounded from above, independent of ε.

Finally, we will show that for both $\alpha = 1, 2$ the first term on the right in (7.23) tends to 0 as $\varepsilon \to 0$. For this, let $\tilde{A}_{F_1}$ and $\tilde{H}_1$ be defined by the commutative diagram (Fig. 3) (P_2 is the canonical projection):

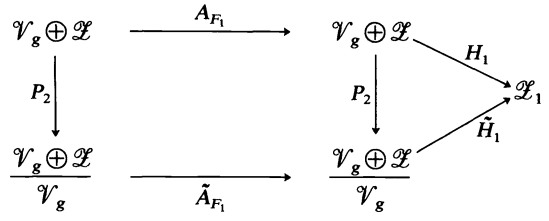


FIG. 3

By (7.14), note that $\sigma(\tilde{A}_{F_1}) = \Lambda \subset \mathbb{C}_g$. Moreover,

$$\|H_1 e^{A_{F_1} t} v(\varepsilon)\|_{L_1} = \|\tilde{H}_1 e^{\tilde{A}_{F_1} t} P_2 v(\varepsilon)\|_{L_1} \leqq \|\tilde{H}_1 e^{\tilde{A}_{F_1} t} P_2\|_{L_1} \|v(\varepsilon)\|.$$

Since $v(\varepsilon) \to 0$, this expression tends to 0 as $(\varepsilon \to 0)$.

Let $\hat{A}_{F_1}$ and $\hat{H}_2$ be defined by the commutative diagram (Fig. 4) (P_3 is the canonical projection):

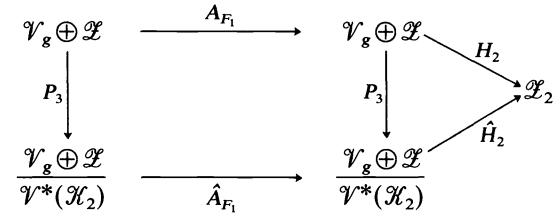


FIG. 4

It can be verified that $\sigma(\hat{A}_{F_1}) = \sigma(\tilde{A}_{F_1}) \cup \sigma(A_{F_1} | \mathscr{V}_g / \mathscr{V}^*(\mathscr{K}_2))$, which, by (7.11) and (7.14) is contained in $\mathbb{C}_g$. It follows that

$$\|H_2 e^{A_{F_1} t} v(\varepsilon)\|_{L_1} = \|\hat{H}_2 e^{\hat{A}_{F_1} t} P_3 v(\varepsilon)\|_{L_1} \leqq \|\hat{H}_2 e^{\hat{A}_{F_1} t} P_3\|_{L_1} \|v(\varepsilon)\|,$$

which again converges to 0 as $\varepsilon \to 0$. This completes the proof of Lemma 7.2. □

Remark 7.7. It is worthwhile to point out which freedom in the spectrum assignment we have in $A + BF_\varepsilon$ when we use the construction of the sequence $\{F_\varepsilon; \varepsilon > 0\}$ as in the proof of Lemma 7.2.

The lattice diagram (Fig. 5) shows the hierarchy of the relevant subspaces in combination with the freedom in the spectrum of $A + BF_\varepsilon$.

Denote $\mathscr{V}_\varepsilon := \mathscr{V}_g(\mathscr{K}_1, \mathscr{K}_2) \oplus R_\varepsilon$.

Note by Lemma 7.3 that the spectrum of the map $A + BF_\varepsilon | \mathscr{R}_\varepsilon$ consists of an eigenvalue in $-\varepsilon^{-1}$ with multiplicity equal to $\dim [\mathscr{V}_b / \mathscr{V}_g]$.

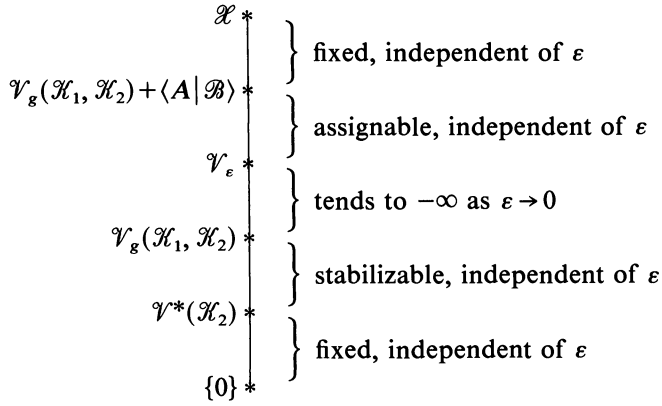


FIG. 5

8. Some special cases and extensions. In this section we will consider some special cases of the main theorem, Theorem 7.1, and spend a few words on some extensions of this result. One interesting special case of $(\text{ADDPBP})_p$ is the situation that we take $H_1 = H$ and $H_2 = I$. This corresponds to the almost disturbance decoupling problem with bounded peaking of the entire state vector. Denote $\mathscr{K} := \ker H$. Since, by Theorem 4.2, $\mathscr{V}_g(\mathscr{K}, \{0\}) = \mathscr{V}_g^*(\mathscr{K}) + \mathscr{V}^*(\{0\}) = \mathscr{V}_g^*(\mathscr{K})$ and since, by Theorem 6.3, $\mathscr{R}_b(\mathscr{K}, \{0\}) = \mathscr{B} + A(\mathscr{R}_b^*(\{0\}) \cap \mathscr{K}) = \mathscr{B} + A(\mathscr{B} \cap \mathscr{K})$, we have the following corollary of Theorem 7.1:

COROLLARY 8.1. *Fix $p \in \{1, 2, \infty\}$. Then the L_p-almost disturbance decoupling problem with bounded peaking of the entire state vector is solvable if and only if* $\operatorname{im} G \subset \mathscr{V}_g^*(\mathscr{K}) + \mathscr{B} + A(\mathscr{B} \cap \mathscr{K})$.

Next, we will spend some words on possible extensions of the results of this paper.

First we would like to point out that, while $(\text{ADDPBP})_p$ is a nontrivial extension of $(\text{ADDP})_p$, we might also consider an extension of the $L_P - L_q$ almost disturbance decoupling problem (ADDP)′, see [20] or [17]. This would lead to the following synthesis problem:

We will say that *the $L_p - L_q$ almost disturbance decoupling problem with bounded peaking* (ADDPBP)′ is solvable if there is a constant C and, for all $\varepsilon > 0$, a feedback map $F_\varepsilon : \mathscr{X} \to \mathscr{U}$ such that with the feedback law $u = F_\varepsilon x$, in the closed loop system for $x(0) = 0$ there holds, *for all* $1 \leqq p \leqq q \leqq \infty$, for all $d \in L_q[0, \infty)$

$$\|z_1\|_{L_p} \leqq \varepsilon \|d\|_{L_q}, \qquad \|z_2\|_{L_p} \leqq C \|d\|_{L_q}.$$

It may be shown that the solvability of the above problem is equivalent to the existence of a sequence of feedback maps $\{F_\varepsilon;\ \varepsilon > 0\}$ and a constant C such that for both $p = 1$ and $p = \infty$ $\|H_1 T_\varepsilon G\|_{L_p} \to 0 (\varepsilon \to 0)$ and $\|H_2 T_\varepsilon G\|_{L_p} \leqq C$ for all ε.

A theory analogous to the one above may be developed around this problem. It can be shown that, again under the assumption that either (A, B, H_1) is left-invertible or that H_2 is injective, a necessary and sufficient condition for the solvability of this problem is that

$$\operatorname{im} G \subset \mathscr{V}_g(\mathscr{K}_1, \mathscr{K}_2) + (\mathscr{R}_b^*(\mathscr{K}_2) \cap \mathscr{K}_1).$$

For more details, the reader is referred to [17].

A final extension of the results of the present paper is the situation in which we require internal stability of the closed loop system. This would lead to the following synthesis problem: We will say that *the L_p-almost disturbance decoupling problem with*

bounded peaking and strong stabilization $(\text{ADDPBPSS})_p$ is solvable if the following is true. There is a constant C and for all $\varepsilon > 0$ and all real numbers S a feedback map $F_{\varepsilon,S}: \mathscr{X} \to \mathscr{U}$ such that, with the feedback law $u = F_{\varepsilon,S}x$, in the closed loop system for $x(0) = 0$ for all $d \in L_p[0, \infty)$ the inequalities (3.2) and (3.3) hold and such that $\operatorname{Re} \sigma(A + BF_{\varepsilon,S}) \leqq S$.

Thus, we require that the spectrum of the closed loop matrix can be located to the left of any vertical line $\operatorname{Re} s = S$ in the complex plane. It may be proven that if at least one of the conditions (7.1), (7.2) hold, then for $p \in \{1, 2, \infty\}$ the latter problem is solvable if and only if (A, B) is controllable and

$$\operatorname{im} G \subset \mathscr{R}^*(\mathscr{K}_1) + \mathscr{R}_b(\mathscr{K}_1, \mathscr{K}_2). \tag{8.1}$$

We note that if (A, B, H_1) is a left-invertible system then the inclusion (8.1) becomes

$$\operatorname{im} G \subset \mathscr{B} + A[\mathscr{R}_b^*(\mathscr{K}_2) \cap \mathscr{K}_1],$$

(see Theorem 6.3). If H_2 is injective then (8.1) becomes

$$\operatorname{im} G \subset \mathscr{R}^*(\mathscr{K}_1) + \mathscr{B} + A[\mathscr{B} \cap \mathscr{K}_1].$$

Again, for details the reader is referred to [17].

9. A worked example. To illustrate the theory developed in this paper and to demonstrate its computational feasibility, in this section we will present a worked example. We will consider a linear system with two outputs and check whether $(\text{ADDPBP})_p$ is solvable for this system. Next, we will actually compute a sequence of feedback mappings that achieves our design purpose. The system that will be considered is given by $\dot{x}(t) = Ax(t) + Bu(t) + Gd(t)$, $z_1(t) = H_1x(t)$, $z_2(t) = H_2x(t)$, with

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix}, \quad H_1 = (0\ 0\ 0\ 1\ 0), \quad H_2 = I_{5\times 5}$$

and

$$G = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}.$$

Thus, $\mathscr{X} = \mathbb{R}^5$ and $\mathscr{U} = \mathbb{R}^2$. Denote $\mathscr{K}_i = \ker H_i$. The route that we will take is as follows. First, we will check whether the subspace inclusion $\operatorname{im} G \subset \mathscr{V}_b(\mathscr{K}_1, \mathscr{K}_2)$ holds to see if $(\text{ADDPBP})_p$ is solvable. It turns out that this is indeed true. After this, we will follow closely the lines of the development in § 7 and construct a required sequence $\{F_n\}$. As before, $\mathbb{C}_g = \{\lambda \in \mathbb{C} \mid \operatorname{Re} \lambda < 0\}$ and the subspaces $\mathscr{V}_g^*(\mathscr{K}_1)$ and $\mathscr{V}_g(\mathscr{K}_1, \mathscr{K}_2)$ are taken with respect to this stability set. Let the standard basis vectors in $\mathbb{R}^5$ be denoted by e_i.

Using the algorithm ISA (see [22, p. 91]) and a construction as in [22, p. 114], we calculate that $\mathscr{V}_g(\mathscr{K}_1, \mathscr{K}_2) = \mathscr{V}_g^*(\mathscr{K}_1) = \operatorname{span}\{e_1, e_2\}$ (since $\mathscr{V}^*(\mathscr{K}_2) = \{0\}$). Thus, by Theorem 4.4, DDPOS is not solvable for the above system. Since $\mathscr{K}_2 = \{0\}$, by Theorem 6.3 we should check if the subspace inclusion $\operatorname{im} G \subset \mathscr{V}_g^*(\mathscr{K}_1) + \mathscr{B} + A(\mathscr{B} \cap \mathscr{K}_1)$ holds. It may be calculated that $\mathscr{V}_g^*(\mathscr{K}_1) + \mathscr{B} + A(\mathscr{B} \cap \mathscr{K}_1) = \operatorname{span}\{e_1, e_2, e_4, e_5\}$. Since $\operatorname{im} G$ is

indeed contained in this subspace, $(\text{ADDPBP})_p$ is solvable for all $1 \leqq p \leqq \infty$. Unfortunately, $(\text{ADDPBPSS})_p$ is *not* solvable because (A, B) is an uncontrollable pair. We will now construct a required sequence of feedback mappings:

Step 1: *decomposition.* We decompose $\mathcal{V}_b = \mathcal{V}_g \oplus \tilde{\mathcal{W}}$, with $\tilde{\mathcal{W}} := \tilde{\mathcal{B}} + A(\tilde{\mathcal{B}} \cap \mathcal{K}_1)$ and $\tilde{\mathcal{B}}$ such that $\tilde{\mathcal{B}} \oplus (\mathcal{B} \cap \mathcal{V}^*(\mathcal{K}_1)) = \mathcal{B}$. Then $\tilde{\mathcal{W}} = \text{span}\{e_4, e_5\}$. Since $e_5 \in \mathcal{B}$ and $e_4 = Ae_5$, $\tilde{\mathcal{W}}$ is equal to the two-dimensional singly generated almost controllability subspace $\mathcal{b} \oplus A\mathcal{b}$, *where* $b = e_5$. Note that indeed (6.2) and (6.3) are satisfied.

Step 2: *approximation.* Approximate $\mathcal{b} \oplus A\mathcal{b}$ by (A, B)-invariant subspaces $\mathcal{R}_\varepsilon = \text{span}\{x_1(\varepsilon), x_2(\varepsilon)\}$ according to (7.3). In our case it can be calculated that $x_1(\varepsilon) = (0\,0\,-\varepsilon^2\,-\varepsilon\,1)^T$ and $x_2(\varepsilon) = Ab = (0\,0\,0\,1\,0)^T$. Note that $b = Bu$ with $u = \binom{0}{1}$. Following (7.5) for $\varepsilon > 0$ define $F_\varepsilon^0 : \mathcal{R}_\varepsilon \to \mathcal{U}$ by $F_\varepsilon^0 x_1(\varepsilon) = -\varepsilon^{-1}\binom{0}{1}$ and $F_\varepsilon^0 x_2(\varepsilon) = -\varepsilon^{-2}\binom{0}{1}$.

Step 3: *a feedback mapping outside* $\mathcal{R}_\varepsilon$. Note that for our system $\langle A \mid \mathcal{B} \rangle + \mathcal{V}_g = \mathcal{X}$. It can be verified that the conditions (7.9) to (7.14) are satisfied with $\Lambda = \{-3\}$; $\mathcal{Z} = \text{span}\{(0\,0\,1\,-3\,9)^T\}$ and $F_1 : \mathcal{Z} \to \mathcal{U}$ given by

$$F_1 = \begin{pmatrix} -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -3 \end{pmatrix}.$$

Step 4: *definition of the required sequence* $\{F_\varepsilon;\ \varepsilon > 0\}$. Note that $\mathcal{X} = \mathcal{V}_g \oplus \mathcal{R}_\varepsilon \oplus \mathcal{Z}$. In this decomposition define $F_\varepsilon | (\mathcal{V}_g \oplus \mathcal{Z}) := F_1 | (\mathcal{V}_g \oplus \mathcal{Z})$ and $F_\varepsilon | \mathcal{R}_\varepsilon := F_\varepsilon^0 | \mathcal{R}_\varepsilon$. It can be verified that the matrix of F_ε with respect to the standard bases in $\mathcal{X} = \mathcal{R}^5$ and $\mathcal{U} = \mathcal{R}^2$ is given by

$$F_\varepsilon := \begin{pmatrix} -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & f_{23}(\varepsilon) & -1/\varepsilon^2 & f_{25}(\varepsilon) \end{pmatrix},$$

where

$$f_{23}(\varepsilon) = \frac{-27\varepsilon^2 + 18\varepsilon - 3}{\varepsilon^2 + 9\varepsilon^4} \quad \text{and} \quad f_{25}(\varepsilon) = \frac{27\varepsilon^3 - 3\varepsilon - 2}{9\varepsilon^3 + \varepsilon}.$$

Finally, by valuating $A + BF_\varepsilon$ in the basis suggested by the decomposition $\mathcal{X} = \mathcal{V}_g \oplus \mathcal{R}_\varepsilon \oplus \mathcal{Z}$, we can calculate the closed loop impulse response matrices from d to z_1 and z_2, respectively:

$$W_{1\varepsilon}(t) := H_1 e^{(A+BF_\varepsilon)t} G = \left(\frac{t}{\varepsilon} + 1\right) e^{-t/\varepsilon},$$

$$W_{2\varepsilon}(t) := H_2 e^{(A+BF_\varepsilon)t} G = \begin{pmatrix} 0 \\ 0 \\ t \\ t/\varepsilon + 1 \\ -t/\varepsilon^2 \end{pmatrix} e^{-t/\varepsilon}.$$

A standard calculation shows that $\| W_{1\varepsilon} \|_{L_1} = 2\varepsilon \to 0$ and that

$$\| W_{2\varepsilon} \|_{L_1} = \int_0^\infty \| W_{2\varepsilon}(t) \|\, dt \leqq 1 + 2\varepsilon + \varepsilon^2.$$

Here, $\| \cdot \|$ denotes the Euclidean norm. From this it can be seen that indeed for every $1 \leqq p \leqq \infty$ the $L_p - L_p$ induced norm of the closed loop operator from d to z_1 tends to zero as $\varepsilon \to 0$ and that the induced norm of the operator from d to z_2 is bounded with respect to ε. Note that $\| F_\varepsilon \| \to \infty$ as $\varepsilon \to 0$.

10. Conclusions. In this paper we have developed a theory around the almost disturbance decoupling problem with bounded peaking. Necessary and sufficient conditions for the solvability of this synthesis problem were formulated in terms of a subspace inclusion involving a certain almost controlled invariant subspace. We showed that this almost controlled invariant subspace can be calculated using existing algorithms. We also provided a conceptual algorithm to calculate the sequence of feedback maps that achieve the design purpose. The calculations involved were illustrated in a numerical example.

Several questions remain to be answered. As a first direction for future research we mention the extension of the above results to the case that the system under consideration does not satisfy one of the conditions (7.1), (7.2), i.e., the system (A, B, H_1) is not left-invertible and the map H_2 is not injective. Another interesting issue would be to extend this theory to the more general situation that we allow only output feedback instead of state feedback. In this context we mention [21] and recent results in [18]. Finally, connections between this work and results on bounded peaking in the context of the nearly singular optimal control problem [3] remain to be worked out.

Appendix A. In this appendix we will establish a proof of Lemma 5.3. The proof will proceed through a series of lemmas. The first lemma is concerned with the convergence of sequence of rational functions and was proven in [7]. In the following, if $f(s)$ is a strictly proper rational function, then $\deg f$ will denote its McMillan degree. We have:

LEMMA A.1. *Let* $\{f_\varepsilon\}$ *be a sequence of strictly proper rational functions. Suppose that there exists* $r \in \mathbb{N}$ *such that* $\deg f_\varepsilon \leqq r$ *for all* ε. *Assume that* $\lim_{\varepsilon \to 0} f_\varepsilon(s)$ *exists for infinitely many* $s \in \mathbb{C}$. *Then there exists a rational function* f *such that* $f_\varepsilon(s) \to f(s)$ $(\varepsilon \to 0)$ *for all but finitely many* $s \in \mathbb{C}$. □

We can then prove the following:

LEMMA A.2. *Suppose that either the condition* (i) *or* (ii) *in Lemma* 5.3 *is satisfied. Then there are a rational vector* $\hat{z}(s)$, *proper and stable and, for* $i = 1, 2$, *subsequences* $\{\hat{z}_{i,\varepsilon'}(s)\}$ *such that* $\hat{z}_{1,\varepsilon'}(s) \to 0$ $(\varepsilon' \to 0)$ *and* $\hat{z}_{2,\varepsilon'}(s) \to \hat{z}(s)$ $(\varepsilon' \to 0)$ *for all but finitely many* s.

Proof. If the condition (i) of Lemma 5.3 holds, then for $\sigma := \operatorname{Re} s \geqq 0$:

$$\|\hat{z}_{i,\varepsilon}(s)\| \leqq \int_0^\infty e^{-\sigma t} \|z_{i,\varepsilon}(t)\| \, dt \leqq \|z_{i,\varepsilon}\|_{L_1}. \tag{A.1}$$

If the condition (ii) of Lemma 5.3 holds, then by the fact that $\hat{z}_{i,\varepsilon}(s)$ is strictly proper and has no poles in $\operatorname{Re} s \geqq 0$, applying the maximum modulus principle [8] gives, for all $\operatorname{Re} s \geqq 0$,

$$\|\hat{z}_{i,\varepsilon}(s)\| \leqq \sup_{\omega \in \mathbb{R}} \|\hat{z}_{i,\varepsilon}(i\omega)\|. \tag{A.2}$$

Hence, in both cases we have $\hat{z}_{1,\varepsilon}(s) \to 0$ $(\varepsilon \to 0)$ and $\|\hat{z}_{2,\varepsilon}(s)\| \leqq C \forall \varepsilon$. Since, for all ε, $\hat{z}_{2,\varepsilon}(s)$ is analytic in $\operatorname{Re} s > 0$ and since the sequence $\{\hat{z}_{2,\varepsilon}(s)\}$ is uniformly bounded there, by Vitali's theorem [8] there exists a function $\hat{z}(s)$, analytic in $\operatorname{Re} s > 0$, and a subsequence $\{\hat{z}_{2,\varepsilon'}(s)\}$ such that $\hat{z}_{2,\varepsilon'}(s) \to \hat{z}(s)$ $(\varepsilon' \to 0)$ *uniformly* on each compact set K in the open right half plane. Therefore, $\hat{z}_{2,\varepsilon'}(s) \to \hat{z}(s)$ $(\varepsilon' \to 0)$ pointwise in $\operatorname{Re} s > 0$. By Lemma A.1, we may assume that $\hat{z}_{2,\varepsilon'}(s) \to \hat{z}(s)$ $(\varepsilon' \to 0)$ *for all but finitely many* s and $\hat{z}(s)$ is *rational.* We contend that $\hat{z}(s)$ cannot have poles in $\operatorname{Re} s = 0$, for define $J := \{s \mid \operatorname{Re} s = 0, s \text{ is not a pole of } \hat{z}(s) \text{ and } \hat{z}_{2,\varepsilon'}(s) \to \hat{z}(s) \ (\varepsilon' \to 0)\}$. Then the complement of J in $\operatorname{Re} s = 0$ is a finite set. Suppose $s_0 \in J$. For ε' sufficiently small, $\|\hat{z}_{2,\varepsilon'}(s_0) - \hat{z}(s_0)\| \leqq 1$. Hence we have

$$\|\hat{z}(s_0)\| \leqq \|\hat{z}(s_0) - \hat{z}_{2,\varepsilon'}(s_0)\| + \|\hat{z}_{2,\varepsilon'}(s_0)\| \leqq 1 + C. \tag{A.3}$$

Therefore, $\hat{z}(s)$ is bounded on J and hence bounded on the entire imaginary axis. Also from (5.3) there follows that $\hat{z}(s)$ is *proper*. Finally, $\hat{z}_{1,\varepsilon'}(s) \to 0$ $(\varepsilon' \to 0)$ in Re $s \geqq 0$ and hence, again by Lemma A.1, for all but finitely many s. □

We will now complete the proof of Lemma 5.3. Recall that $u_\varepsilon(t)$ is a regular Bohl input and $z_{1,\varepsilon}(t) = H_1 x_\varepsilon(t)$, $z_{2,\varepsilon}(t) = H_2 x_\varepsilon(t)$, where $\dot{x}_\varepsilon = Ax_\varepsilon + Bu_\varepsilon$, $x_\varepsilon(0) = x_0$. We will prove that if either the condition (i) or (ii) in Lemma 5.3 is satisfied, then $x_0 \in \mathscr{V}_b(\mathscr{K}_1, \mathscr{K}_2)$:

Proof of Lemma 5.3. Let $F: \mathscr{X} \to \mathscr{U}$ be such that $A_F \mathscr{V}^*(\mathscr{K}_2) \subset \mathscr{V}^*(\mathscr{K}_2)$. Denote $v_\varepsilon(t) := u_\varepsilon(t) - Fx_\varepsilon(t)$. Then $\dot{x}_\varepsilon = A_F x_\varepsilon + Bv_\varepsilon$. Note that x_ε and v_ε have rational Laplace transforms. Let $P: \mathscr{X} \to \mathscr{X}/\mathscr{V}^*(\mathscr{K}_2)$ be the canonical projection and let $\overline{A_F}$ be the quotient map of A_F modulo $\mathscr{V}^*(\mathscr{K}_2)$. Let $\bar{B} := PB$ and let $\bar{H}_1$ and $\bar{H}_2$ be mappings such that $\bar{H}_1 P = H_1$ and $\bar{H}_2 P = H_2$. Decompose $\mathscr{U} = \mathscr{U}_1 \oplus \mathscr{U}_2$ with $\mathscr{U}_1 = \ker \bar{B}$ and $\mathscr{U}_2$ an arbitrary complement. Accordingly, partition $\bar{B} = (0, \bar{B}_2)$. Then $\bar{B}_2$ is injective. Let $\bar{G}(s) := \bar{H}_2(Is - \bar{A}_F)^{-1}\bar{B}_2$. Let $\bar{\mathscr{R}}^*(\bar{\mathscr{K}}_2)$ be the supremal controllability subspace in $\bar{\mathscr{K}}_2$ with respect to $(\bar{A}_F, \bar{B})$. By [22, Ex. 5.8], we have $\bar{\mathscr{R}}^*(\bar{\mathscr{K}}_2) = \{0\}$. Hence, by [22, Ex. 4.4], $\bar{G}(s)$ is a left-invertible rational matrix, with left-inverse $\bar{G}^+(s)$. Now, let $\xi_\varepsilon(s)$ and $\omega_\varepsilon(s)$ be the Laplace transforms of x_ε and v_ε respectively. Let $\bar{\xi}_\varepsilon(s) := P\xi_\varepsilon(s)$ and $\bar{x}_0 := Px_0$. Partition

$$\omega_\varepsilon(s) = \begin{pmatrix} \omega_{1,\varepsilon}(s) \\ \omega_{2,\varepsilon}(s) \end{pmatrix},$$

and conform the decomposition $\mathscr{U} = \mathscr{U}_1 \oplus \mathscr{U}_2$. The following relations then hold

$$x_0 = (Is - A_F)\xi_\varepsilon(s) - B\omega_\varepsilon(s), \tag{A.4}$$

$$\bar{x}_0 = (Is - \bar{A}_F)\bar{\xi}_\varepsilon(s) - \bar{B}_2\omega_{2,\varepsilon}(s), \tag{A.5}$$

and hence

$$\omega_{2,\varepsilon}(s) = \bar{G}^+(s)[\hat{z}_{2,\varepsilon}(s) - \bar{H}_2(Is - \bar{A}_F)^{-1}\bar{x}_0], \tag{A.6}$$

$$\bar{\xi}_\varepsilon(s) = (Is - \bar{A}_F)^{-1}(\bar{x}_0 + \bar{B}_2\omega_{2,\varepsilon}(s)). \tag{A.7}$$

Apply Lemma A.2 to obtain $\hat{z}(s)$ such that $\hat{z}_{2,\varepsilon}(s) \to \hat{z}(s)$ $(\varepsilon \to 0)$ for all but finitely many s (write ε in the subsequence for which this holds). It follows that there are rational vectors $\omega_2(s)$ and $\bar{\xi}(s)$ such that $\omega_{2,\varepsilon}(s) \to \omega_2(s)$ and $\bar{\xi}_\varepsilon(s) \to \bar{\xi}(s)$ for all but finitely many s. Define now

$$\omega(s) := \begin{pmatrix} 0 \\ \omega_2(s) \end{pmatrix}.$$

Then we have $\bar{x}_0 = (Is - \bar{A}_F)\bar{\xi}(s) - \bar{B}\omega(s)$. Moreover, since $\bar{H}_1\bar{\xi}_\varepsilon(s) = \hat{z}_{1,\varepsilon}(s) \to 0$ $(\varepsilon \to 0)$, we have $\bar{H}_1\bar{\xi}(s) = 0$. Also, since $\bar{H}_2\bar{\xi}_\varepsilon(s) = \hat{z}_{2,\varepsilon}(s) \to \hat{z}(s)$, $\bar{H}_2\bar{\xi}(s)$ is proper and stable.

Finally, let $\xi(s)$ be any rational vector such that $\bar{\xi}(s) = P\xi(s)$. Then $H_1\xi(s) = 0$, $H_2\xi(s)$ is proper and stable and, for some vector $x_1 \in \mathscr{V}^*(\mathscr{K}_2)$, $x_0 = (Is - A_F)\xi(s) - B\omega(s) + x_1$. It follows that $x_0 - x_1 \in \mathscr{V}_b(\mathscr{K}_1, \mathscr{K}_2)$. Since $\mathscr{V}^*(\mathscr{K}_2) \subset \mathscr{V}_b(\mathscr{K}_1, \mathscr{K}_2)$ we thus obtain $x_0 \in \mathscr{V}_b(\mathscr{K}_1, \mathscr{K}_2)$. This completes the proof of Lemma 5.3. □

Appendix B. In this appendix we will state and prove a result on the geometrical structure of the sequence of subspaces $\mathscr{F}^i(\mathscr{K}_1, \mathscr{K}_2)$, given by (6.8). Our result is a generalization of [19, Thm. 7.1] and deals with the representation of subspaces in terms of chains in the input space $\mathscr{B}$. Related results can be found in [14] and [10]. Further, in this appendix we will prove Lemma 6.5.

LEMMA B.1. *Given the system* $\dot{x} = Ax + Bu$ *and subspaces* $\mathscr{K}_2 \subset \mathscr{K}_1 \subset \mathscr{X}$, *let* $\mathscr{F}^i(\mathscr{K}_1, \mathscr{K}_2)$ *be defined by* (6.7) *and* (6.8). *Then the following holds: for all* $i \in \underline{n}$, *there are a chain* $\{\mathscr{B}_l\}_{l=1}^i$ *and a map* $F: \mathscr{X} \to \mathscr{U}$ *such that*

$$\mathscr{F}^i(\mathscr{K}_1, \mathscr{K}_2) = \bigoplus_{l=1}^{i} A_F^{l-1}\mathscr{B}_l, \tag{B.1}$$

$$\bigoplus_{l=2}^{i} A_F^{l-2}\mathscr{B}_l \subset \mathscr{K}_2, \tag{B.2}$$

$$\dim \mathscr{B}_l = \dim A_F^{l-1}\mathscr{B}_l = \dim [\mathscr{F}^l(\mathscr{K}_1, \mathscr{K}_2)/\mathscr{F}^{l-1}(\mathscr{K}_1, \mathscr{K}_2)] \qquad (l \in \underline{i}). \tag{B.3}$$

Remark. In the above, for consistence define $\bigoplus_{l=2}^{i} A_F^{l-2}\mathscr{B}_l = \{0\}$ if $i = 1$. In the following we will denote $\mathscr{F}^i := \mathscr{F}^i(\mathscr{K}_1, \mathscr{K}_2)$.

Proof. The proof is by induction. For $i = 1$, the lemma is trivially true: take $\mathscr{B}_1 := \mathscr{B} \cap \mathscr{K}_1$. Suppose now the lemma is true for i. Let $\{\mathscr{B}_l\}_{l=1}^i$ be a chain in $\mathscr{B}$ and F be a map such that the conditions (B.1), (B.2) and (B.3) are satisfied. We will show that $\mathscr{F}^{i+1}$ can also be represented as above. This will be done by constructing an *extra term* $\mathscr{B}_{i+1}$ for the chain $\{\mathscr{B}_l\}_{l=1}^i$ and by defining *a new feedback map* F_{new}. First, let $\mathscr{B}'_l$, $l \in \underline{i}$, be subspaces such that $\mathscr{B}'_l \oplus \mathscr{B}_l = \mathscr{B}_{l-1}$ (define $\mathscr{B}_0 := \mathscr{B}$). Using (B.1) we then have:

$$\begin{aligned}\mathscr{F}^i \cap \mathscr{K}_2 &= \left(\sum_{l=1}^{i} A_F^{l-1}(\mathscr{B}_{l+1} + \mathscr{B}'_{l+1}) + A_F^{i-1}\mathscr{B}_i\right) \cap \mathscr{K}_2 \\ &= \left(\sum_{l=1}^{i-1} A_F^{l-1}\mathscr{B}_{l+1} + \sum_{l=1}^{i-1} A_F^{l-1}\mathscr{B}'_{l+1} + A_F^{i-1}\mathscr{B}_i\right) \cap \mathscr{K}_2.\end{aligned}$$

Using the modular distributive rule [22, p. 4] and (B.2), it follows that

$$\mathscr{F}^i \cap \mathscr{K}_2 = \sum_{l=1}^{i} A_F^{l-1}\mathscr{B}_{l+1} + \left(\sum_{l=1}^{i-1} A_F^{l-1}\mathscr{B}'_{l+1} + A_F^{i-1}\mathscr{B}_i\right) \cap \mathscr{K}_2. \tag{B.4}$$

On the other hand, by (6.8),

$$\mathscr{F}^{i+1} = \mathscr{T}^{i+1}(\mathscr{K}_2) \cap \mathscr{K}_1 = (\mathscr{B} + A_F(\mathscr{T}^i(\mathscr{K}_2) \cap \mathscr{K}_2)) \cap \mathscr{K}_1. \tag{B.5}$$

Since, by the fact that $\mathscr{K}_2 \subset \mathscr{K}_1$, $\mathscr{T}^i(\mathscr{K}_2) \cap \mathscr{K}_2 = \mathscr{F}^i \cap \mathscr{K}_2$, it follows by combining (B.4) and (B.5)

$$\begin{aligned}\mathscr{F}^{i+1} &= \left(\mathscr{B} + \sum_{l=1}^{i} A_F^l\mathscr{B}_{l+1} + A_F\left[\left(\sum_{l=1}^{i-1} A_F^{l-1}\mathscr{B}'_{l+1} + A_F^{i-1}\mathscr{B}_i\right) \cap \mathscr{K}_2\right]\right) \cap \mathscr{K}_1 \\ &= (\mathscr{F}^i + \mathscr{G}) \cap \mathscr{K}_1.\end{aligned} \tag{B.6}$$

Here, we defined

$$\mathscr{G} := \mathscr{B}'_1 + A_F\left[\left(\sum_{l=1}^{i} A_F^{l-1}\mathscr{B}'_{l+1} + A_F^{i-1}\mathscr{B}_i\right) \cap \mathscr{K}_2\right].$$

Again using the modular distributive rule and $\mathscr{F}^i \subset \mathscr{K}_1$, we obtain

$$\mathscr{F}^{i+1} = \mathscr{F}^i + (\mathscr{G} \cap \mathscr{K}_1). \tag{B.7}$$

Let $\hat{\mathscr{G}} \subset \mathscr{G} \cap \mathscr{K}_1$ be a subspace such that $\mathscr{F}^{i+1} = \mathscr{F}^i \oplus \hat{\mathscr{G}}$ and let $\{v_1, \cdots, v_r\}$ be a basis for $\hat{\mathscr{G}}$. By definition of $\mathscr{G}$, each v_j can be represented as $v_j = \sum_{l=1}^{i} A_F^{l-1}b'_{l,j} + A_F^i b_j$, with

$$\sum_{l=2}^{i} A_F^{l-2}b'_{l,j} + A_F^{i-1}b_j \in \mathscr{K}_2. \tag{B.8}$$

Here, $b_j \in \mathcal{B}_i$ and $b'_{l,j} \in \mathcal{B}'_l$ ($j \in \underline{r}$, $l \in \underline{i}$). By the assumption that $\hat{\mathcal{G}} \cap \mathcal{F}^i = \{0\}$, it can be verified that for fixed $l \in \{0, \cdots, i\}$, the vectors $\{A_F^l b_1, \cdots, A_F^l b_r\}$ are linearly independent. Define now

$$\mathcal{B}_{i+1} := \text{span}\,\{b_1, \cdots, b_r\}. \tag{B.9}$$

Note that $\mathcal{B}_{i+1} \subset \mathcal{B}_i$. Also, for $l \in \underline{i}$, define vectors $x_{j,l}$ by

$$x_{j,i} := b_j,\ x_{j,l} := \sum_{k=1}^{l-1} A_F^{l-k-1} b'_{i-k+1,j} + A_F^{l-1} b_j. \tag{B.10}$$

From (B.10), for $l = 2, \cdots, i$ we have $x_{j,l} = A_F x_{j,l-1} + b'_{i-l+2,j}$ and $v_j = A_F x_{j,i} + b'_{i,j}$. Moreover, by the independency of the vectors $\{A_F^l b_j; j \in \underline{r}\}$ and by the fact that the spaces $A_F^{l-1}\mathcal{B}_l$ ($l \in \underline{i}$) are independent (the sum in (B.1) is direct), it can be shown that the vectors $\{x_{j,l}; j \in \underline{r}, l \in \underline{i}\}$ are linearly independent. Extend this system to a basis for $\mathcal{X}$. Let $u_{j,k} \in \mathcal{U}$ be such that $b'_{i-k+1,j} = Bu_{j,k}$ and define a map $F'' : \mathcal{X} \to \mathcal{U}$ by defining it in the above basis by $F'' x_{j,l} := u_{j,l}$ and F'' arbitrary on the extension. It can then be seen that for $l \in \underline{i}$

$$\begin{aligned} &\text{span}\,\{x_{1,l}, \cdots, x_{r,l}\} = (A_F + BF'')^{l-1}\mathcal{B}_{i+1}, \\ &\hat{\mathcal{G}} = \text{span}\,\{v_1, \cdots, v_r\} = (A_F + BF'')^i \mathcal{B}_{i+1} \end{aligned} \tag{B.11}$$

and, for $l = 1, \cdots, i-1$

$$BF''(A_F + BF'')^{l-1}\mathcal{B}_{i+1} \subset \mathcal{B}'_{i+1-l} \subset \mathcal{B}_1. \tag{B.12}$$

Let $\{\mathcal{B}''_l\}_{l=1}^i$ be a chain in $\mathcal{B}$ such that $\mathcal{B}_{i+1} \oplus \mathcal{B}''_l = \mathcal{B}_l$. Since $\mathcal{F}^{i+1} = \mathcal{F}^i \oplus \hat{\mathcal{G}}$, by (B.1) and (B.11) we obtain

$$\begin{aligned} \mathcal{F}^{i+1} &= \sum_{l=1}^{i} A_F^{l-1}\mathcal{B}_l + (A_F + BF'')^i \mathcal{B}_{i+1} \\ &= \sum_{l=1}^{i} A_F^{l-1}\mathcal{B}_{i+1} + (A_F + BF'')^i \mathcal{B}_{i+1} + \sum_{l=1}^{i} A_F^{l-1}\mathcal{B}''_l. \end{aligned}$$

Thus, by (B.1^):

$$\mathcal{F}^{i+1} = \sum_{l=1}^{i+1} (A_F + BF'')^{l-1}\mathcal{B}_{i+1} + \sum_{l=1}^{i} A_F^{l-1}\mathcal{B}''_l. \tag{B.13}$$

We contend that *all* sums in (B.13) are, in fact, direct sums. To prove this, assume the contrary. Then the following strict inequality must hold

$$\begin{aligned} \dim \mathcal{F}^{i+1} &< \sum_{l=1}^{i+1} \dim (A_F + BF'')^{l-1}\mathcal{B}_{i+1} + \sum_{l=1}^{i} \dim A_F^{l-1}\mathcal{B}''_l \\ &\leqq \sum_{l=1}^{i} \dim \mathcal{B}_{i+1} + \dim \hat{\mathcal{G}} + \sum_{l=1}^{i} \dim \mathcal{B}''_l = \sum_{l=1}^{i} \dim \mathcal{B}_l + \dim \hat{\mathcal{G}}, \end{aligned}$$

where the last equality follows from the fact that $\mathcal{B}_{i+1} \oplus \mathcal{B}''_l = \mathcal{B}_l$. On the other hand, however, $\dim \mathcal{F}^{i+1} = \dim \mathcal{F}^i + \dim \hat{\mathcal{G}}$, which by (B.3) and (B.1) equals $\sum_{l=1}^{i} \dim \mathcal{B}_l + \dim \hat{\mathcal{G}}$. Hence we obtain a contradiction. It follows that

$$\mathcal{F}^{i+1} = \bigoplus_{l=1}^{i+1} (A_F + BF'')^{l-1}\mathcal{B}_{i+1} \oplus \bigoplus_{l=1}^{i} A_F^{l-1}\mathcal{B}''_l. \tag{B.14}$$

Define $\mathcal{V} := \bigoplus_{l=1}^{i} (A_F + BF'')^{l-1}\mathcal{B}_{i+1}$ and $\mathcal{W} := \bigoplus_{l=1}^{i} A_F^{l-1}\mathcal{B}_l''$. By (B.14) we have that $\mathcal{V} \cap \mathcal{W} = \{0\}$. Decompose $\mathcal{X} = \mathcal{V} \oplus \mathcal{W} \oplus \mathcal{R}$, where $\mathcal{R}$ is arbitrary. In this decomposition, define $F_{\text{new}}: \mathcal{X} \to \mathcal{U}$ as follows

$$F_{\text{new}}|\mathcal{V} := (F + F'')|\mathcal{V}, \qquad F_{\text{new}}|\mathcal{W} := F|\mathcal{W} \tag{B.15}$$

and F_{new} arbitrary on $\mathcal{R}$. It can now be seen immediately from (B.14) that

$$\mathcal{F}^{i+1} = \bigoplus_{l=1}^{i+1} (A + BF_{\text{new}})^{l-1}\mathcal{B}_{i+1} \oplus \bigoplus_{l=1}^{i} (A + BF_{\text{new}})^{l-1}\mathcal{B}_l'' = \bigoplus_{l=1}^{i+1} (A + BF_{\text{new}})^{l-1}\mathcal{B}_l. \tag{B.16}$$

This already verifies (B.1). Next, we will verify (B.3).

It is claimed that for $l = 1, 2, \cdots, i+1$, $\dim (A + BF_{\text{new}})^{l-1}\mathcal{B}_l = \dim \mathcal{B}_l$. Suppose the contrary. Then we have

$$\begin{aligned} \sum_{l=1}^{i} \dim \mathcal{B}_l + \dim \hat{\mathcal{G}} = \dim \mathcal{F}^{i+1} &= \sum_{l=1}^{i+1} \dim (A + BF_{\text{new}})^{l-1}\mathcal{B}_l \\ &= \sum_{l=1}^{i} \dim (A + BF_{\text{new}})^{l-1}\mathcal{B}_l + \dim \hat{\mathcal{G}} \\ &< \sum_{l=1}^{i} \dim \mathcal{B}_l + \dim \hat{\mathcal{G}}, \end{aligned}$$

which is a contradiction. Equation (B.3) then follows immediately by noting that $\dim (A + BF_{\text{new}})^{i}\mathcal{B}_{i+1} = \dim \hat{\mathcal{G}} = \dim [\mathcal{F}^{i+1}/\mathcal{F}^{i}]$. Finally, it can be verified using (B.2), (B.8), (B.11) and (B.15) that $\bigoplus_{l=2}^{i+1} (A + BF_{\text{new}})^{l-2}\mathcal{B}_l \subset \mathcal{K}_2$. □

Remark B.2. Note that the proof of the above lemma is straightforward but notationally rather involved. An alternative proof could be given using the concept of train basis, see [14].

The above lemma is needed in the following:

Proof of Lemma 6.5. By Lemma B.1, there is a chain $\{\mathcal{B}_i\}_{i=1}^{n}$ in $\mathcal{B}$ and a map $F: \mathcal{X} \to \mathcal{U}$ such that $\mathcal{F}^n = \bigoplus_{i=1}^{n} A_F^{i-1}\mathcal{B}_i$ and such that (6.12) holds. Since $\mathcal{F}^n \subset \mathcal{K}_1$, also (6.11) holds. By (B.3), to prove (6.13) it is sufficient to show that, for $i \in \underline{n}$, $\dim A_F^{i-1}\mathcal{B}_i = \dim A_F^{i}\mathcal{B}_i$. Suppose the contrary, i.e., suppose that $\dim A_F^{i-1}\mathcal{B}_i > \dim A_F^{i}\mathcal{B}_i$ for some i. Then there is a vector $v \neq 0$ in $A_F^{i-1}\mathcal{B}_i$ such that $A_F v = 0$. Since a subspace $\mathcal{V} \subset \mathcal{X}$ is controlled invariant iff $A_F\mathcal{V} \subset \mathcal{V} + \mathcal{B}$ [22, Lemma 4.2] it follows that $\operatorname{span}\{v\}$ is controlled invariant. Since also $v \in \mathcal{K}_1$, v must be contained in $\mathcal{V}^*(\mathcal{K}_1)$, the largest controlled invariant subspace in $\mathcal{K}_1$. On the other hand, $v \in \mathcal{F}^n = \mathcal{R}_b^*(\mathcal{K}_2) \cap \mathcal{K}_1 \subset \mathcal{R}_b^*(\mathcal{K}_1) \cap \mathcal{K}_1$. By Lemma 2.2 it follows that $b \in \mathcal{R}_a^*(\mathcal{K}_1)$. Thus, $v \in \mathcal{R}_a^*(\mathcal{K}_1) \cap \mathcal{V}^*(\mathcal{K}_1)$, which by Lemma 2.2 contradicts the assumption that $\mathcal{R}^*(\mathcal{K}_1) = \{0\}$. Thus, we have proved formula (6.13).

Finally, to prove (6.10), note from (6.9) that $\mathcal{R}_b(\mathcal{K}_1, \mathcal{K}_2) = \mathcal{B} + A_F\mathcal{F}^n = \mathcal{B} + A_F\mathcal{B}_1 + \cdots + A_F^n\mathcal{B}_n$. It will be shown that this is, in fact, a direct sum. Suppose it is not. Then there are vectors $b_i \in \mathcal{B}_i$ and $b_0 \in \mathcal{B}$, not all zero, such that $\sum_{i=0}^{n} A_F^i b_i = 0$. Define $w := \sum_{i=1}^{n} A_F^{i-1} b_i$. Then we have $A_F w = -b_0 \in \mathcal{B}$. Since also $w \in \mathcal{K}_1$, w must be contained in $\mathcal{V}^*(\mathcal{K}_1)$. On the other hand, $w \in \mathcal{R}_a^*(\mathcal{K}_1)$, and it follows as above that $w = 0$. Hence, $b_0 = 0$. Repeating this argument with $w = \sum_{i=1}^{n} A_F^{i-1} b_i$ replaced by $w = \sum_{i=2}^{n} A_F^{i-2} b_i$ then yields $A_F w = -b_1$, and thus $b_1 = 0$, etc. In this way we find $b_i = 0$ $(i \in \underline{n})$, which is a contradiction. □

Appendix C. This appendix will be devoted to a proof of Lemma 7.4. The proof will be given through a series of smaller lemmas. For $\varepsilon > 0$, let $x_i(\varepsilon)$, $i \in \underline{k}$, a subspace $\mathcal{L}_\varepsilon$ and a map $F_\varepsilon: \mathcal{L}_\varepsilon \to \mathcal{U}$ be given by (7.3) to (7.5). Recall from Lemma 7.3 that a

matrix of the map $(A_F+BF_\varepsilon)|\mathscr{L}_\varepsilon$ with respect to the basis $X_\varepsilon := \{x_1(\varepsilon), \cdots, x_k(\varepsilon)\}$ is given by (7.6). Now, let $D_\varepsilon : \mathscr{L}_\varepsilon \to \mathscr{L}_\varepsilon$ be the linear map with matrix diag $(-1/\varepsilon, \cdots, -1/\varepsilon)$ with respect to X_ε. Define a nilpotent map $N_\varepsilon : \mathscr{L}_\varepsilon \to \mathscr{L}_\varepsilon$ by $N_\varepsilon := (A_F+BF_\varepsilon)|\mathscr{L}_\varepsilon - D_\varepsilon$. Obviously, the matrix of N_ε with respect to X_ε is given by

$$\text{(C.1)} \qquad \text{mat } N_\varepsilon = \begin{pmatrix} 0 & -\varepsilon^{-2} & \cdots & -\varepsilon^{-k} \\ 0 & 0 & \ddots & \vdots \\ \vdots & \vdots & \ddots & -\varepsilon^{-2} \\ 0 & 0 & \cdots & 0 \end{pmatrix}.$$

The following lemma is then immediate:

LEMMA C.1. *Let $i \in \underline{k}$. Then for $j = i, i+1, \cdots, k$ we have $N_\varepsilon^j x_i(\varepsilon) = 0$. On the other hand, for $j = 1, 2, \cdots, i-1$ the following holds*

$$\text{(C.2)} \qquad N_\varepsilon^j x_i(\varepsilon) = \sum_{l_1=1}^{i-1} \sum_{l_2=1}^{l_1-1} \cdots \sum_{l_j=1}^{l_{j-1}-1} (-1)^j \varepsilon^{l_j-i-j} x_{l_j}(\varepsilon).$$

(*For consistency, define $l_0 := i$.*)

Proof. Use (C.1) to obtain an expression for $N_\varepsilon x_i(\varepsilon)$. Apply N_ε to the result, etc. □

Another technical ingredient we will need in our proof is:

LEMMA C.2. *Let $i \in \underline{k}$. Then we have*

$$\text{(C.3)} \qquad x_i(\varepsilon) = A_F^{i-1} b - \varepsilon \sum_{l=1}^{i} A_F^{i-l+1} x_l(\varepsilon).$$

Proof. This follows immediately from (7.3), using induction. □

Finally, we will need the following result:

LEMMA C.3. *Under the assumptions of Lemma 7.4, the following holds for all $i \in \underline{k}$: $H_1 x_i(\varepsilon) = O(\varepsilon^{k-i})$ and $H_2 x_i(\varepsilon) = O(\varepsilon^{k-i-1})$.*

Proof. By iterating formula (C.3), we obtain

$$\begin{aligned} x_i(\varepsilon) = {} & A_F^{i-1} b - \varepsilon \sum_{l_1=1}^{i} A_F^i b + \varepsilon^2 \sum_{l_1=1}^{i} \sum_{l_2=1}^{l_1} A_F^{i+1} b + \cdots \\ & + (-\varepsilon)^{k-i-1} \left(\sum_{l_1=1}^{i} \sum_{l_2=1}^{l_1} \cdots \sum_{l_{k-i-1}=1}^{l_{k-i-2}} A_F^{k-2} b \right) \\ & + (-\varepsilon)^{k-i} \left(\sum_{l_1=1}^{i} \sum_{l_2=1}^{l_1} \cdots \sum_{l_{k-i}=1}^{l_{k-i-1}} A_F^{k-l_{k-i}} x_{l_{k-i}}(\varepsilon) \right) \end{aligned} \qquad \text{(C.4)}$$

(assume that $1 \leqq i \leqq k-2$). Under the assumptions of Lemma 7.4 we have $A_F^{i-1} b, \cdots, A_F^{k-3} b \in \mathscr{K}_2 \subset \mathscr{K}_1$ and $A_F^{k-2} b \in \mathscr{K}_1$. Thus, in (C.4) all terms but the last are in $\mathscr{K}_1$ and all terms but the last two are in $\mathscr{K}_2$. It follows then that $\lim_{\varepsilon \to 0} \varepsilon^{i-k} \cdot H_1 x_i(\varepsilon)$ exists and that $\lim_{\varepsilon \to 0} \varepsilon^{i+1-k} \cdot H_2 x_i(\varepsilon)$ exists.

For $i = k-1$, the existence of the former limit follows again from (C.4), while the existence of the latter is obvious. For $i = k$, the existence of both limits is obvious. □

Proof of Lemma 7.4. By the nilpotency of N_ε, note that for $i \in \underline{k}$

$$e^{(A_F+BF_\varepsilon)t} x_i(\varepsilon) = e^{N_\varepsilon t} e^{D_\varepsilon t} x_i(\varepsilon) = \left(\sum_{j=0}^{k-1} \frac{t^j N_\varepsilon^j}{j!} \right) e^{-(1/\varepsilon)t} x_i(\varepsilon).$$

By the triangle inequality it therefore suffices to prove the following: for $j = 0, 1, \cdots, k-1$, the sequence $\| t^j e^{-(1/\varepsilon)t} H_\alpha N_\varepsilon^j x_i(\varepsilon) \|_{L_1}$ tends to 0 as $\varepsilon \to 0$ for $\alpha = 1$ and is uniformly bounded with respect to ε if $\alpha = 2$. Since $\int_0^\infty t^j e^{-(1/\varepsilon)t} \, dt = j! \varepsilon^{j+1}$, it suffices to prove this asymptotic behaviour for $\| \varepsilon^{j+1} H_\alpha N_\varepsilon^j x_i(\varepsilon) \|$ (Euclidean norm!). Apply now Lemma

C.1 to obtain a representation of $N_\varepsilon^j x_i(\varepsilon)$. Again by the triangular inequality, it is then sufficient to prove that $\lim_{\varepsilon\to 0} \varepsilon^{j+1} H_\alpha \varepsilon^{l-i-j} x_l(\varepsilon)$ is 0 for $\alpha = 1$ and exists for $\alpha = 2$. (l is some index ranging between 1 and k.) Now, by Lemma C.3 $H_1 x_l(\varepsilon) = O(\varepsilon^{k-l})$ and $H_2 x_l(\varepsilon) = O(\varepsilon^{k-l-1})$. Thus, indeed $\lim_{\varepsilon\to 0} \varepsilon^{l-i-1} H_1 x_l(\varepsilon) = 0$ for *all* $l \in \underline{k}$ and $i \in \underline{k}$ and $\lim_{\varepsilon\to 0} \varepsilon^{l-i-1} H_2 x_l(\varepsilon)$ exists for *all* $l \in \underline{k}$ and $i \in \underline{k}$. This completes the proof of Lemma 7.4. □

Acknowledgments. The author would like to thank Professor Jan Willems for inspiring discussions and valuable comments and Dr. Kees Praagman for several helpful suggestions during the preparation of this paper.

REFERENCES

[1] C. COMMAULT AND J. M. DION, *Structure at infinity of linear multivariable systems, a geometric approach*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 693-696.

[2] C. A. DESOER AND M. VIDYASAGAR, *Feedback Systems: Input-Output Properties*, Academic Press, New York, 1975.

[3] B. A. FRANCIS AND K. GLOVER, *Bounded peaking in the optimal linear regulator with cheap control*, IEEE Trans. Automat. Control, AC-23 (1978), pp. 608-617.

[4] M. L. J. HAUTUS, *Stabilization, controllability and observability of linear autonomous systems*, Nederl. Akad. Wetensch. Proc. Ser. A, 73 (1970), pp. 448-455.

[5] ———, *(A,B)-invariant and stabilizability subspaces, a frequency domain description*, Automatica, 16 (1980), pp. 703-707.

[6] M. L. J. HAUTUS AND L. M. SILVERMAN, *System structure and singular control*, Lin. Algebra Appl., 50 (1983), pp. 369-402.

[7] M. HAZEWINKEL, *On families of linear systems: degeneration phenomena*, in Algebraic and Geometric Methods in Linear Systems Theory, Lectures in Applied Mathematics, vol. 18, American Mathematical Society, Providence, RI, 1980.

[8] E. HILLE, *Analytic Function Theory*, vol. II, Chelsea, New York, 1962.

[9] H. KIMURA, *A new approach to the perfect regulation and the bounded peaking in linear multivariable control systems*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 253-279.

[10] M. MALABRE, *A complement about almost controllability subspaces*, Systems Control Lett., 3 (1983), pp. 119-122.

[11] J. M. SCHUMACHER, *Dynamic Feedback in Finite and Infinite Dimensional Linear Systems*, Mathematical Centre Tracts 143, Amsterdam, 1981.

[12] ———, *Algebraic characterizations of almost invariance*, Intern. J. Control, 38 (1983), pp. 107-124.

[13] ———, *Almost stabilizability subspaces and high gain feedback*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 620-627.

[14] ———, *On the structure of strongly controllable systems*, Internat. J. Control, 38 (1983), pp. 525-545.

[15] H. L. TRENTELMAN, *On the assignability of infinite root loci in almost disturbance decoupling*, Internat. J. Control, 38 (1983), pp. 147-167.

[16] ———, *Reduction of observer order by differentiation, almost controllability subspace covers and minimal order* PID-*observers*, Systems Control Lett., 4 (1984), pp. 57-64.

[17] ———, *Almost invariant subspaces and high gain feedback*, Ph.D. dissertation, Rijksuniversiteit Groningen, May 1985.

[18] H. L. TRENTELMAN AND J. C. WILLEMS, *Guaranteed roll-off in a class of high gain feedback design problems*, Systems Control Lett., 3 (1983), pp. 361-369.

[19] J. C. WILLEMS, *Almost* A(mod B)*-invariant subspaces*, Astérisque, 75-76 (1980), pp. 230-249.

[20] ———, *Almost invariant subspaces: an approach to high gain feedback design-part* 1: *almost controlled invariant subspaces*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 235-253.

[21] ———, *Almost invariant subspaces: an approach to high gain feedback design-part* 2: *almost conditionally invariant subspaces*, IEEE Trans. Automat. Control, 27 (1982), pp. 1071-1085.

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

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 6, November 1986

006

DIAGONALLY MODIFIED CONDITIONAL GRADIENT METHODS FOR INPUT CONSTRAINED OPTIMAL CONTROL PROBLEMS*

J. C. DUNN†

Abstract. Many iterative methods for constrained minimization problems conform to the general scheme, $u_{i+1} \in A_i(\Omega, F, u_i)$, $u_1 \in \Omega$ where Ω is a set of feasible vectors in a Banach space X, F is a payoff function whose minimum is sought in Ω and A_i is a map with range in the set of subsets of Ω. If $\{(\Omega_i, F_i)\}$ is a sequence of related problems that approximate (Ω, F) in some sense, with $\Omega_i \subset \Omega_{i+1} \subset \Omega$, then the corresponding diagonal modification of the original algorithm generates iterates via the recursion, $u_{i+1} \in A_i(\Omega_i, F_i, u_i)$, $u_1 \in \Omega_1$. If (Ω_i, F_i) is properly selected, the diagonal modification can compute approximate solutions for (Ω, F) efficiently in circumstances where the original algorithm is difficult or impossible to implement for (Ω, F). In particular, this happens for certain gradient-related descent methods and increasingly refined finite-dimensional approximations to infinite-dimensional optimal control problems. Convergence rate estimates are obtained for a diagonally modified conditional gradient method, and this algorithm is applied to bounded input continuous-time optimal control problems.

Key words. conditional gradient method, diagonal modification, adaptive precision, optimal control, input constraints, convergence rates

AMS(MOS) subject classifications. 49D10, 49D37, 65B99

1. Introduction.

Conditional gradient algorithms belong to the general iterative scheme

$$u_{i+1} \in A_i(\Omega, F, u_i), \qquad u_1 \in \Omega \tag{1}$$

for $i = 1, 2, 3, \cdots$, where Ω is a set of "feasible" vectors in a real Banach space X, F is a real valued "payoff" function whose minimum is sought in Ω, and $A_i(\Omega, F, \cdot)$ is a map from X to the set of subsets of Ω, defined for each member of a certain class of problems (Ω, F). While it is often possible to clearly *imagine* the implementation of a given algorithm (1) and to determine its convergence properties *in theory*, an exact implementation may be computationally difficult (or even impossible) for certain (Ω, F)'s, and by design or otherwise one may end up generating iterates with

$$u_{i+1} \in A_i(\Omega_*, F_*, u_i), \qquad u_1 \in \Omega_*, \tag{2}$$

where (Ω_*, F_*) is in some sense close to the original problem. If (Ω_*, F_*) is held constant in (2), the iterates u_i will generally yield at best an approximate solution of (Ω, F) even in the limit as $i \to \infty$. On the other hand, if the quality of the approximation (Ω_*, F_*) is repeatedly upgraded during the iteration, the sequence $\{u_i\}$ may actually "solve" (Ω, F) as $i \to \infty$. In the simplest application of this basic principle, $\{u_i\}$ is recursively generated by

$$u_{i+1} \in A_i(\Omega_i, F_i, u_i), \qquad u_1 \in \Omega, \tag{3}$$

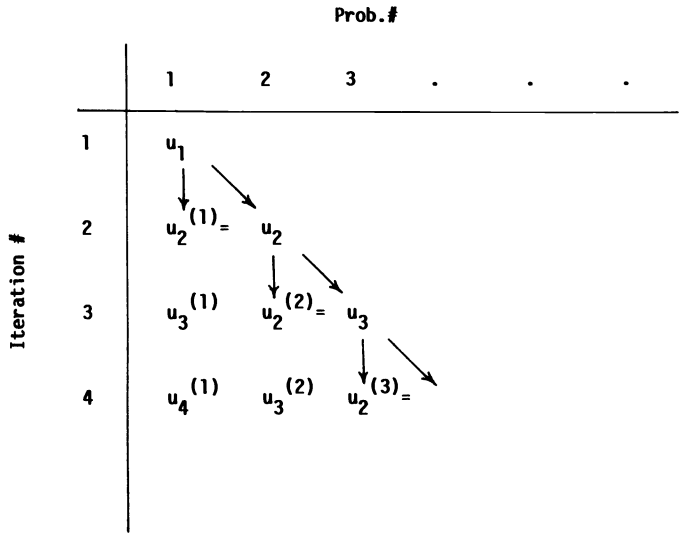
where $\{(\Omega_i, F_i)\}$ is a sequence of increasingly refined approximations to (Ω, F) satisfying

$$\Omega_i \subset \Omega_{i+1} \subset \Omega. \tag{4}$$

The iteration scheme (3) will be referred to here as a *diagonal modification* of (1) (see Fig. 1), following Tapia's terminology in [1].

* Received by the editors January 16, 1984, and in revised form May 2, 1985. This investigation was supported by National Science Foundation Research grant ECS-8005958.

† Mathematics Department, North Carolina State University, Raleigh, North Carolina 27695-8205.



$u_j^{(i)}$ = j^{th} iterate of (1.2) for $(\Omega_*, F_*) = (\Omega_i, F_i)$

u_i = i^{th} iterate of (1.3)

FIG. 1. *Diagonal modifications of* (1.1).

The so-called adaptive precision minimization algorithms formulated by Klessig and Polak [2] and Schittkowski [3] in the 1970's are essentially elaborations of (3) applied to Armijo rule descent methods for unconstrained minimization problems. These references deal extensively with "finite dimensional" approximations (Ω_i, F_i) to unconstrained Mayer final value problems (Ω, F), with

$$\Omega = \mathscr{L}^p([0,1], U), \tag{5a}$$

$$F(u(\cdot)) = P(x(1)), \tag{5b}$$

$$\frac{dx}{dt} = f(x, t, u(t)), \quad x \in \mathbb{R}^n, \quad t \in [0,1], \tag{5c}$$

$$x(0) = x_0, \tag{5d}$$

where $P: \mathbb{R}^n \to \mathbb{R}^1$ and $f: \mathbb{R}^n \times \mathbb{R}^1 \times \mathbb{R}^m \to \mathbb{R}^n$ are given functions, U is a given set in $\mathbb{R}^m$, $p \in [1, \infty)$ and x_0 is a given vector in $\mathbb{R}^n$. In reference [2], $U = \mathbb{R}^m$ and the corresponding approximations Ω_i consist of step functions in Ω that are *constant* on the intervals of a given net $\mathcal{N}_i$ for $[0,1]$, and the functions F_i are obtained by replacing the differential equation in (5) with an approximating *difference* equation on $\mathcal{N}_i$. Other possibilities for (Ω_i, F_i) are considered in [3]. Reference [2] describes two encouraging numerical experiments with an adaptive precision gradient method for (5), and shows that every subsequential limit of an iterate sequence produced by this method is a critical point of the limiting problem (Ω, F) under reasonable conditions on $\{(\Omega_i, F_i)\}$ and (Ω, F). Convergence theorems of a similar nature are established in [3]; however, convergence *rate* estimates are not developed in either study, and input constraints are not treated.

In the light of certain results developed in [4], [5] it seems likely that when $\{(\Omega_i, F_i)\}$ converges quickly enough to (Ω, F), the iterates generated by (1) and (3) will have the same *asymptotic* convergence properties, even though this may *not* be

true of (2) and (3) with (Ω_*, F_*) *fixed* at any particular (Ω_j, F_j) different from (Ω, F). This suggests that efficient computational procedures for a given (Ω, F) might be obtained by selecting an algorithm known to have good asymptotics for (Ω, F), and then implementing a diagonal modification of this algorithm with (Ω_i, F_i) chosen to simplify the computation of vectors in $A_i(\Omega_i, F_i, u)$, especially for small and moderate values of i. The guiding principle in the selection of (Ω_i, F_i) is to achieve reductions in computational cost mainly at the expense of degraded *asymptotics* for (2) with (Ω_*, F_*) fixed at any (Ω_j, F_j); so long as the latter algorithm substantially reduces F_j in the *first few* iterations, the diagonal modification (3) may well compute refined approximate solutions for (Ω, F) more efficiently than (2) for any fixed (Ω_*, F_*) near (Ω, F).

As outlined above, the diagonal modification principle amounts to a general strategy for deriving new and possibly more efficient iterative methods from existing algorithms. To realize the full potential of this idea in particular situations, it is necessary to know more about special structure in the problem at hand, and to exploit that structure in the selection of (1) and $\{(\Omega_i, F_i)\}$. Nevertheless, for large classes of minimization algorithms, problem structure influences the *asymptotic* behavior of (3) *only implicitly* through a few basic continuity, growth and convergence properties of (Ω, F) and $\{(\Omega_i, F_i)\}$, and for this reason it is possible to construct nontrivial general asymptotic convergence rate theories for (3) that can help in identifying likely candidate algorithms for diagonal modification in a variety of problem contexts. One such theory is formulated here for diagonal modifications of the conditional gradient method (DMCG). Together with certain results in [6]–[11], this theory indicates that DMCG methods should have acceptable asymptotic properties for bounded input versions of (5) with "bang-bang" solutions, notwithstanding the notoriously poor *generic* asymptotic behavior of the unmodified CG methods in finite-dimensional polytopes. Numerical results are also presented for a prototype bounded input control problem (5), and a DMCG method on nested nets $\mathcal{N}_i$ for $[0, 1]$. The net sequences employed in this example are generated by doubling the number of equally spaced net points at each iteration. Even for this crude scheme, i iterations of the DMCG appear to be $i/2$ times less costly than the same number of iterations of the unmodified CG method on the net $\mathcal{N}_i$. When $\mathcal{N}_i$ is highly refined, it should be noted that a single CG iterate is already much cheaper to implement than one iterate of a typical "higher order" method with better asymptotics.

The results presented here are part of a more extensive theory developed in [12] for diagonally modified recursive linear-quadratic programming methods.

2. A diagonally modified conditional gradient method. Let Ω be a nonempty convex set in a real Banach space X and suppose, without loss in generality, that X is the closed linear hull of Ω. Let F be a Frechet differentiable real valued function defined on some neighborhood of Ω in X, and assume that

$$\mathcal{T}(\Omega, F, u) \triangleq \arg\min_{v \in \Omega} F'(u)v \neq \varnothing, \qquad u \in \Omega \tag{1}$$

(this is certainly true if Ω is closed and bounded and X is reflexive). Fix δ in $(0, \frac{1}{2})$, and for $u \in \Omega$ and $\bar{u} \in \mathcal{T}(\Omega, F; u)$ let $W(\Omega, F; u, \bar{u})$ be a set of "step lengths" in $[0, 1]$ that satisfy Goldstein's rule [13], namely: for all $\omega \in W$,

$$\omega = 0 \quad \text{if } u \in \mathcal{T}(\Omega, F; u), \tag{2a}$$

$$\omega = 1 \quad \text{if } u \notin \mathcal{T}(\Omega, F; u) \quad \text{and} \quad \frac{F(u) - F(\bar{u})}{F'(u)(u - \bar{u})} \geqq \delta, \tag{2b}$$

$$\text{(2c)}\quad 1-\delta \geqq \frac{F(u)-F(u+\omega(\bar{u}-u))}{\omega F'(u)(u-\bar{u})} \geqq \delta \quad \text{if } u \notin \mathcal{T}(\Omega, F; u) \quad \text{and} \quad \frac{F(u)-F(\bar{u})}{F'(u)(u-\bar{u})} < \delta.$$

The associated recursion,

$$\text{(3a)}\qquad u_{i+1} = u_i + \omega_i(\bar{u}_i - u_i), \qquad u_1 \in \Omega,$$

$$\text{(3b)}\qquad \bar{u}_i \in \mathcal{T}(\Omega, F, u_i),$$

$$\text{(3c)}\qquad \omega_i \in W(\Omega, F; u_i, \bar{u}_i),$$

defines a *conditional gradient method* for the constrained minimization problem (Ω, F). For convex polyhedral sets in $\mathbb{R}^n$ and quadratic objective functions F, (3) reduces to the Frank-Wolf method with Goldstein step lengths.

Note 1. By definition ξ is an *extremal* of F in the set Ω iff

$$F'(\xi)(u-\xi) \geqq 0, \qquad u \in \Omega.$$

Equivalently,

$$u \text{ is an extremal of } F \text{ in } \Omega \Leftrightarrow \xi \in \mathcal{T}(\Omega, F; \xi).$$

Thus (3) "stops" at ξ iff ξ is an extremal. For convex Ω, every local minimizer of F in Ω is an extremal. For convex F, every extremal is a minimizer. □

For certain feasible sets Ω, the algorithm (3) costs little more to implement than simple unconstrained steepest descent. For example, if Ω is the unit cube,

$$\text{(4a)}\qquad \Omega = \{u \in \mathbb{R}^n : |u_j| \leqq 1; j = 1, 2, \cdots, n\},$$

then

$$\text{(4b)}\qquad \mathcal{T}(\Omega, F; u) = \left\{\bar{u} \in \Omega : \bar{u}_j \in -\operatorname{sgn}\frac{\partial F}{\partial u_j}(u); j = 1, 2, \cdots, n\right\}.$$

Similarly, if Ω is the unit simplex

$$\text{(5a)}\qquad \Omega = \left\{\bar{u} \in \mathbb{R}^n : \sum_{i=1}^{n} u_i = 1 \text{ and } u_i \geqq 0; i = 1, 2, \cdots, n\right\},$$

then

$$\text{(5b)}\qquad \mathcal{T}(\Omega, F; u) = \left\{\bar{u} \in \Omega : \sum_{j \in J} \bar{u}_j = 1 \text{ for } J = \arg\min_j \frac{\partial F}{\partial u_j}(u)\right\}.$$

More generally, if Ω is a convex polyhedral set in $\mathbb{R}^n$ then $\mathcal{T}(\Omega, F; u)$ consists of all solutions of a linear program with $F'(u)v$ as the objective function; if the linear constraints that define Ω are few in number, or are suitably structured, this program is easily solved and (3) is readily implemented. The sets $\mathcal{T}$ can also be written down at once for certain "round" convex sets Ω. Thus, if Ω is the unit ball,

$$\text{(6a)}\qquad \{u \in \mathbb{R}^n : \|u\| \leqq 1\},$$

then

$$\text{(6b)}\qquad \mathcal{T}(\Omega, F; u) = \begin{cases} \Omega & \text{if } \nabla F(u) = 0, \\ \left\{\dfrac{\nabla F(u)}{\|\nabla F(u)\|}\right\} & \text{if } \nabla F(u) \neq 0. \end{cases}$$

Still more generally, if Ω is a cartesian product

$$\text{(7a)}\qquad \Omega = \prod_{j=1}^{k} U^{(j)} = U^{(1)} \times U^{(2)} \times \cdots \times U^{(k)}$$

of convex sets $U^{(j)} \subset \mathbb{R}^{n_j}$, then

$$\mathcal{T}(\Omega, F; u) = \prod_{j=1}^{k} \arg \min_{v^{(j)} \in U^{(j)}} \nabla F^{(j)}(u) \cdot v^{(j)} \tag{7b}$$

where

$$\nabla F(u) = (\nabla F^{(1)}(u), \nabla F^{(2)}(u), \cdots, \nabla F^{(k)}(u)) \in \prod_{j=1}^{k} \mathbb{R}^{n_j}. \tag{7c}$$

Hence $\mathcal{T}(\Omega, F; u)$ will be easy to construct if the individual minimizer sets in (7b) are easy to compute.

Feasible sets with simple product structure arise naturally in discrete-time counterparts of the control problem (1.5). In fact, the set (1.5a) itself may be viewed as a kind of "infinite" product set with an associated decomposition formula analogous to (7b). Thus, for $p \in [1, \infty)$, let $q = p(p-1)^{-1} \in (1, \infty]$. If the real valued function F is Frechet differentiable at u in the set $\mathcal{L}^p([0, 1], \mathbb{R}^m)$, then by the Riesz representation theorem, there is a unique function g_u in $\mathcal{L}^q([0, 1], \mathbb{R}^m)$ such that,

$$F'(u)v = \int_0^1 g_u(t) \cdot v(t)\, dt, \qquad v \in \mathcal{L}^p([0, 1], \mathbb{R}^m) \tag{8}$$

where $\cdot$ signifies the standard inner product in $\mathbb{R}^m$. If U is a bounded closed convex set in $\mathbb{R}^m$ and if

$$\Omega = \mathcal{L}^p([0, 1], U) \tag{9a}$$

then it follows from (8) that

$$\mathcal{T}(\Omega, F; u) = \left\{ \bar{u} \in \mathcal{L}^p([0, 1], \mathbb{R}^m) \colon u(t) \overset{a.e.}{\in} \arg \min_{v \in U} g_u(t) \cdot v \right\}. \tag{9b}$$

This is the counterpart of formula (7b) for finite Cartesian products.

While the CG method (3) is easy to implement in specially structured finite-dimensional convex poyhedra, its generic convergence rate is known to be poor in *any* convex polyhedron [6]. On the other hand, when Ω is defined by nonlinear inequalities in $\mathbb{R}^n$, or when Ω is an *infinite* product of polyhedral sets U, the CG algorithm may converge well *in principle*, even though an exact computation of $\mathcal{T}(\Omega, F; u)$ is no longer possible [6], [7]. In such cases, one is led naturally to consider diagonally modified CG recursions

$$u_{i+1} = u_i + \omega_i(\bar{u}_i - u_i), \qquad u_1 \in \Omega, \tag{10a}$$

$$\bar{u}_i \in \mathcal{T}(\Omega_i, F_i; u_i), \tag{10b}$$

$$\omega_i \in W(\Omega_i, F_i; u_i, \bar{u}_i), \tag{10c}$$

$$\Omega_i \subset \Omega_{i+1} \subset \Omega, \tag{10d}$$

where F_i is Frechet differentiable on some neighborhood of Ω_i in the closed linear hull X_i of Ω_i, where $\{(\Omega_i, F_i)\}$ "converges" to (Ω, F) in some appropriate sense, and where $\mathcal{T}(\Omega_i, F_i; u)$ is relatively easy to compute, especially in the early iterations.

More specifically, $O(i^{-1})$ convergence is generic for (3) and convex F on convex polyhedral $\Omega \subset \mathbb{R}^n$ [6], [7]; however, *linear* convergence is generic for (3) and convex F on the set (9a), provided $p = 1$, F' is Lipschitz continuous, U is a convex polyhedron in $\mathbb{R}^m$, and the minimizer of F is a "bang-bang" (i.e., piecewise constant vertex-valued) function [4], [6]. Consider now that the set $\Omega = L^1([0, 1], U)$ is readily approximated

from below by sets Ω_i of *step functions* that have finite ranges in U and are constant on the subintervals of nets,

$$\mathcal{N}_i = \{t_0^{(i)}, t_1^{(i)}, \cdots, t_{n_i}^{(i)}\} \tag{11a}$$

with

$$0 = t_0^{(i)} < t_1^{(i)} < \cdots < t_{n_i}^{(i)} = 1, \tag{11b}$$

$$\mathcal{N}_i \subset \mathcal{N}_{i+1}, \qquad i = 1, 2, \cdots, \tag{11c}$$

$$\lim_{i \to \infty} \max_{1 \leqq j \leqq n_i} (t_j^{(i)} - t_{j-1}^{(i)}) = 0. \tag{11d}$$

In the case of (1.5), the objective function F is defined implicitly through a differential equation whose exact solutions are generally unknown, and hence F must also be approximated on the step function sets Ω_i; however, this is readily done by replacing (1.5c) with an approximating *difference* equation on the nets $\mathcal{N}_i$ (as in [2], [3]). Hence if U is a convex polyhedral set in $\mathbb{R}^m$, if (1.5) has a "bang-bang" solution, and if the nets $\mathcal{N}_i$ in (11) are refined quickly enough, it is plausible that the DMCG method will achieve the "theoretical" convergence rate for (3) and the limiting infinite-dimensional control problem (1.5). This conjecture is supported by the analysis in § 3, and by the numerical example in § 4. The results in § 3 can also be applied to continuous-time control problems in $\Omega = \mathcal{L}^p([0, 1], U)$, with U a "uniformly convex" set in $\mathbb{R}^m$.

3. Convergence rate estimates. When (Ω_*, F_*) is "close" to (Ω, F) it does *not* necessarily follow that the *asymptotic* convergence rate of (1.2) is like the corresponding rate for (1.1). For instance, it can happen that (Ω, F) has a "regular" minimizer for which (1.1) has good asymptotic properties, while (Ω_*, F_*) has a "singular" minimizer for which (1.2) behaves poorly. This point is illustrated in a control problem setting by the example in § 4 of [3], and § 4 of the present article. On the other hand, when (Ω_*, F_*) is near (Ω, F), these examples and the analysis in [5] show that (1.1) and (1.2) can behave similarly in the *exterior* of a small ball $B(\xi_*, \sigma_*)$ with radius σ_* and center $\xi_* \in \arg_{\Omega_*} \min F_*$, even though their asymptotic properties may be quite different *inside* the ball; in such cases the radius σ_* depends on the "distance" between (Ω_*, F_*) and (Ω, F) and typically *shrinks to zero* as $(\Omega_*, F_*) \to (\Omega, F)$ (roughly speaking, $B(\xi_*, \sigma_*)$ is like a "boundary-layer" in singular perturbation theory). These considerations suggest that the diagonal modifications (1.3) can have the *same* asymptotic behavior as (1.1) for (Ω, F) if the sequence $\{(\Omega_i, F_i)\}$ is chosen to make the balls $B(\xi_i, \sigma_i)$ shrink fast enough. This is the central idea in the convergence rate analysis presented below for conditional gradient method methods, and in the more extensive treatment of gradient-related methods in [12].

The convergence rate for the algorithm (2.3) is directly correlated with the growth properties of the local first order approximation $F'(\xi)\ (u - \xi)$ restricted to $u \in \Omega$ near a minimizer ξ for (Ω, F). To understand how this algorithm behaves for nearby problems (Ω_*, F_*), it is therefore important to know how the growth properties of $F'_*(\xi)(u - \xi_*)$ in Ω_* differ from those of $F'(\xi)(u - \xi)$ in Ω. A result of this kind is established in the following variant of Lemma 3.1 in [5].

LEMMA 1. *Let ξ and ξ_* be minimizers for (Ω, g) and (Ω_*, g_*) with $\Omega_* \subset \Omega$. Suppose that for some $a > 0$ and all $\sigma \geqq 0$*

$$\inf_{\substack{u \in \Omega \\ \|u - \xi\| \geqq \sigma}} g(u) - g(\xi) \geqq a\sigma^2. \tag{1}$$

Let a' be any fixed positive number less than a. Then for all $\varepsilon > 0$, the conditions

(2a) $$\sup_{u\in\Omega_*} |g(u) - g_*(u)| \leqq \frac{a\varepsilon^2}{3}$$

and

(2b) $$\inf_{u\in\Omega_*} g(u) - \inf_{u\in\Omega} g(u) \leqq \frac{a\varepsilon^2}{3}$$

imply that

(3) $$\|\xi - \xi_*\| \leqq \varepsilon$$

and

(4a) $$\inf_{\substack{u\in\Omega_* \\ \|u-\xi_*\|\geqq\sigma}} g_*(u) - g_*(\xi_*) \geqq a'\sigma^2$$

for

(4b) $$\sigma \geqq \sigma(\varepsilon) \triangleq 2a\varepsilon(a-a')^{-1}.$$

Proof. By hypothesis

$$\frac{a\varepsilon^2}{3} \geqq \inf_{u\in\Omega_*} g(u) - g(\xi) \geqq \inf_{u\in\Omega_*} g_*(u) - \frac{a\varepsilon^2}{3} - g(\xi)$$

$$= g_*(\xi_*) - \frac{a\varepsilon^2}{3} - g(\xi) \geqq g(\xi_*) - \frac{2a\varepsilon^2}{3} - g(\xi) \geqq a\|\xi_* - \xi\|^2 - \frac{2a\varepsilon^2}{3}.$$

Hence

$$\|\xi_* - \xi\| \leqq \varepsilon.$$

Furthermore, for all $u \in \Omega_*$,

$$g_*(u) - g_*(\xi_*) = g(u) - g(\xi) + g(\xi) - g_*(\xi_*) + g_*(u) - g(u)$$

$$\geqq g(u) - g(\xi) + g(\xi) - g_*(\xi_*) - \frac{a\varepsilon^2}{3}$$

and

$$g(\xi) - g_*(\xi_*) \geqq \inf_{u\in\Omega} g(u) - \inf_{u\in\Omega_*} g(u) - \frac{a\varepsilon^2}{3} \geqq \frac{-2a\varepsilon^2}{3}$$

and therefore,

$$g_*(\xi) - g_*(\xi_*) \geqq g(u) - g(\xi) - a\varepsilon^2.$$

Since $\Omega_* \subset \Omega$, and since $\|u - \xi_*\| \geqq \sigma$ implies that $\|u - \xi\| \geqq \sigma - \|\xi - \xi_*\| \geqq \sigma - \varepsilon$, it now follows that for all $\sigma \geqq \varepsilon$,

$$\inf_{\substack{u\in\Omega_* \\ \|u-\xi_*\|\geqq\sigma}} g_*(u) - g_*(\xi_*) \geqq \inf_{\substack{u\in\Omega \\ \|u-\xi\|\geqq\sigma-\varepsilon}} g(u) - g(\xi) - a\varepsilon^2$$

(5)

$$\geqq a[(\sigma-\varepsilon)^2 - \varepsilon^2] = a(\sigma^2 - 2\varepsilon\sigma).$$

Observe that

$$a(\sigma^2 - 2\varepsilon\sigma) \geqq a'\sigma^2$$

for all $\sigma \geqq \sigma(\varepsilon) = 2a\varepsilon(a-a')^{-1} > \varepsilon$ (Fig. 2). □

Note 1. *Inside* the ball with center ξ_* and radius $\sigma(\varepsilon)$, g_* may have an asymptotic growth rate completely different than g's growth rate near ξ.

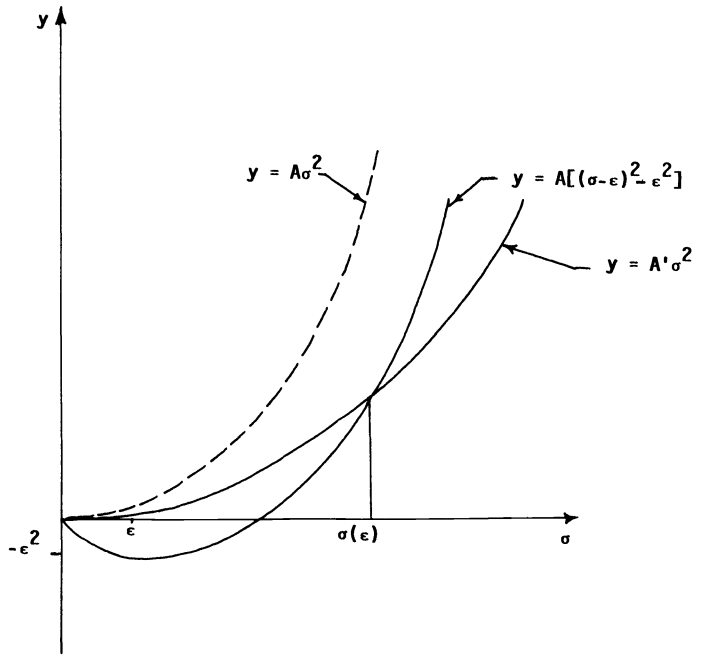


FIG. 2

Note 2. A result analogous to Lemma 1 is readily proved for *local* minimizers ξ, ξ_* satisfying

$$\inf_{\substack{u\in\Omega \\ \bar\sigma>\|u-\xi\|\geqq\sigma}} g(u)-g(\xi)\geqq a\sigma^2 \tag{6}$$

and

$$\|\xi-\xi_*\|<\bar\sigma \tag{7}$$

for some $a>0$, some $\bar\sigma>0$ and all $\sigma\in[0,\bar\sigma)$. In place of (4), one now obtains the growth condition

$$\inf_{\substack{u\in\Omega_* \\ \bar\sigma-\varepsilon>\|u-\xi_*\|\geqq\sigma}} g_*(u)-g_*(\xi_*)\geqq a'\sigma^2 \tag{8a}$$

for

$$\bar\sigma-\varepsilon\geqq\sigma\geqq\sigma(\varepsilon)=2a\varepsilon(a-a')^{-1}. \tag{8b}$$

In addition to Lemma 1, the convergence rate estimate in Theorem 1 below also depends on the following result.

LEMMA 2. *Let* $\{r_i\}$ *be a sequence of nonnegative real numbers such that for some* $\beta\in[0,1)$, $\lambda\in[0,1)$ *and* $c>0$,

$$r_{i+1}\leqq\beta r_i+c\lambda^i, \qquad i=1,2,3,\cdots. \tag{9}$$

If $\lambda\neq\beta$, *then*

$$r_i\leqq\beta^{i-1}r_1+\frac{c\lambda}{|\beta-\lambda|}\left|\beta^{i-1}-\lambda^{i-1}\right|, \qquad i=1,2,3,\cdots. \tag{10a}$$

On the other hand, if $\lambda = \beta$, *then*

$$r_i \leqq (r_1 + c(i-1))\beta^{i-1}, \qquad i = 1, 2, 3, \cdots. \tag{10b}$$

Proof. By induction,

$$r_i \leqq \beta^{i-1} r_1 + c \sum_{j=1}^{i-1} \beta^{i-j-1}\lambda^j \tag{11a}$$

and

$$(\beta - \lambda) \sum_{j=1}^{i-1} \beta^{i-j-1}\lambda^j = \lambda(\beta^{i-1} - \lambda^{i-1}) \tag{11b}$$

for $i = 2, 3, \cdots$. The inequalities (10) follow immediately from (11). □

With Lemmas 1 and 2, it is now possible to prove the following convergence rate theorem for the DMCG method (2.10). The growth condition (12) imposed in this theorem is known to hold for important classes of continuous-time optimal control problems (1.5) and for standard nonlinear programs in $\mathbb{R}^n$ with feasible sets Ω defined by finitely many uniformly convex inequalities (see Note 4 below).

THEOREM 1. *Let* Ω *be a nonempty convex set in a real Banach space* X, *let* F *be a convex Frechet differentiable real valued function defined on some neighborhood of* Ω *in* X, *and let* F *be Lipschitz continuous on bounded sets in* Ω. *Suppose that* ξ *minimizes* F *in* Ω *and that*

$$F'(\xi)(u - \xi) \geqq a\|u - \xi\|^2 \tag{12}$$

for some $a > 0$ *and all* $x \in \Omega$. *Let* $\{\Omega_i\}$ *be a nondecreasing sequence of nonempty convex subsets of* Ω, *and let* $\{F_i\}$ *be a sequence of convex Frechet differentiable real valued functions defined on neighborhoods of the corresponding sets* Ω_i *in their closed linear hulls* $X_i \subset X$. *Assume that* $\{F_i'\}$ *satisfies the uniform Lipschitz continuity condition*

$$\|F_i'(u) - F_i'(v)\| \triangleq \sup_{\substack{w \in X_i \\ \|w\| = 1}} |(F_i'(u) - F_i'(v))w| \leqq L\|u - v\| \tag{13}$$

for some $L > 0$, *all* $u, v \in \Omega_i$ *and* $i = 1, 2, 3, \cdots$. *In addition, let* $\{\xi_i\}$ *be a sequence of minimizers for the associated problems* (Ω_i, F_i), *and suppose that for some* $b > 0$, *some* $\lambda \in [0, 1)$ *and all* i,

$$\sup_{u \in \Omega_i} |F_i(u) - F(u)| \leqq b\lambda^i, \tag{14a}$$

$$\inf_{u \in \Omega_i} F(u) - \inf_{u \in \Omega} F(u) \leqq b\lambda^i, \tag{14b}$$

and

$$\sup_{u \in \Omega_i} |(F_i'(\xi_i) - F'(\xi))u| \leqq b\lambda^i. \tag{14c}$$

Finally, for $l > 0$, *let* S_l *denote the level set*

$$S_l = \{u \in \Omega: F(u) - F(\xi) \leqq l\}.$$

Then there is a number $\rho \in [0, 1)$, *independent of* l, *and positive numbers* C_1 *and* C_2 *depending on* l, *such that*

$$0 \leqq F(u_i) - F(\xi) \leqq C_1\rho^{i-1}, \qquad i = 1, 2, 3, \cdots \tag{15a}$$

and

$$0 \leqq F_i(u_i) - F_i(\xi_i) \leqq C_2\rho^{i-1}, \qquad i = 1, 2, 3, \cdots \tag{15b}$$

for every sequence $\{u_i\}$ that satisfies the diagonally modified conditional gradient recursion (2.10) *and originates at* $u_1 \in S_l$.

Proof. By Note 2.1, (14b) and the convexity of F,

$$\inf_{\Omega_i} F'(\xi)u - \inf_{\Omega} F'(\xi)u = \inf_{\Omega_i} F'(\xi)(u-\xi) \leqq \inf_{\Omega_i} F(u) - F(\xi) \leqq b\lambda^i. \tag{16}$$

Fix a' in $(0, a)$ and put

$$\alpha_1 = (12ab)^{1/2}(a-a')^{-1}, \tag{17a}$$

$$\alpha_2 = (3b/a)^{1/2}, \tag{17b}$$

$$\sigma_i = \alpha_1 \lambda^{i/2}. \tag{17c}$$

Let $g(u) = F'(\xi)u$ and $g_i(u) = F_i'(\xi_i)u$. Then by (12), (14), (16) and Lemma 1,

$$\|\xi_i - \xi\| \leqq \alpha_2 \lambda^{i/2} \tag{17d}$$

and

$$u \in \Omega_i \quad \text{and} \quad \|u - \xi_i\| \geqq \sigma_i \Rightarrow F_i'(\xi_i)(u-\xi_i) \geqq a'\|u-\xi_i\|^2. \tag{17e}$$

Let $\{u_i\}$ satisfy (2.10) with $\{\bar{u}_i\}$ and $\{\omega_i\}$, and put

$$r_i = F(u_i) - F(\xi) \geqq 0, \qquad i = 1, 2, 3, \cdots.$$

Suppose that

$$\|u_i - \xi_i\| < \sigma_i$$

for some particular value of i. Then by (17)

$$\|u_i - \xi\| \leqq \|u_i - \xi_i\| + \|\xi_i - \xi\| < (\alpha_1 + \alpha_2)\lambda^{i/2}. \tag{18}$$

Hence,

$$r_i \leqq K\|u_i - \xi\| \leqq \alpha_3 \lambda^{i/2} \tag{19a}$$

where K is a Lipschitz constant for F in the ball $\{u \in \Omega: \|u-\xi\| \leqq \alpha_1 + \alpha_2\}$, and

$$\alpha_3 = K(\alpha_1 + \alpha_2). \tag{19b}$$

It now follows from (2.10), (14) and (19) that

$$\begin{aligned} r_{i+1} &= F(u_{i+1}) - F(\xi) = F_i(u_{i+1}) + F(u_{i+1}) - F_i(u_{i+1}) - F(\xi) \\ &\leqq F_i(u_i) - F(\xi) + b\lambda^i \leqq r_i + 2b\lambda^i \leqq \alpha_4 \lambda^{i/2} \end{aligned} \tag{20a}$$

where

$$\alpha_4 = \alpha_3 + 2b\lambda^{1/2}. \tag{20b}$$

On the other hand, suppose that

$$\|u_i - \xi_i\| \geqq \sigma_i$$

for some particular i. Then (13), (14) and (17) can be used to estimate r_{i+1} as follows. First, observe that (2.10), (14) and the convexity of F_i imply that

$$\begin{aligned} r_i - r_{i+1} &= F(u_i) - F(u_{i+1}) \\ &= F_i(u_i) - F_i(u_{i+1}) + F(u_i) - F_i(u_i) + F_i(u_{i+1}) - F(u_{i+1}) \\ &\geqq \delta\omega_i F_1'(u_i)(u_i - \bar{u}_i) - 2b\lambda^i \\ &\geqq \delta\omega_i F_1'(u_i)(u_i - \xi_i) - 2b\lambda^i \\ &\geqq \delta\omega_i (F_i(u_i) - F_i(\xi_i)) - 2b\lambda^i \end{aligned} \tag{21}$$

and

$$|F_i(\xi_i) - F(\xi)| = \left|\inf_{\Omega_i} F_i = \inf_{\Omega} F\right| \leqq \left|\inf_{\Omega_i} F - \inf_{\Omega} F\right| + b\lambda^i \leqq 2b\lambda^i \tag{22}$$

and therefore

$$r_i - r_{i+1} \geqq \delta\omega_i r_i - 4b\lambda^i. \tag{23}$$

If $u_i \in \mathcal{T}(\Omega_i, F_i; u_i)$ then $r_i = 0$, by Note 2.1; in this case (23) gives

$$r_{i+1} \leqq 4b\lambda^i. \tag{24}$$

However, if $u_i \notin \mathcal{T}(\Omega_i, F_i; u_i)$ then $F_1'(u_i)(u_i - \bar{u}_i) > 0$, and (2.10), (13) and the Fundamental Theorem of Calculus imply that

$$\omega_i = 1 \tag{25a}$$

or

$$1 - \delta \geqq \frac{F_i(u_i) - F_i(u_{i+1})}{\omega_i F_i'(u_i)(u_i - \bar{u}_i)} \geqq 1 - \frac{\omega_i L\|u_i - \bar{u}_i\|^2}{2F_1'(u_i)(u_i - \bar{u}_i)}. \tag{25b}$$

Consequently, by (17) and the convexity of F_i,

$$\omega_i \geqq \min\left\{1, 2\delta\frac{F_1'(u_i)(u_i - \bar{u}_i)}{L\|u_i - \bar{u}_i\|^2}\right\} \geqq \min\left\{1, \frac{2\delta a'\|u_i - \xi_i\|^2}{L(\|u_i - \xi_i\| + \|\bar{u}_i - \xi_i\|)^2}\right\}. \tag{26}$$

Observe now that

$$\|\bar{u}_i - \xi_i\| < \sigma_i \Rightarrow \|\bar{u}_i - \xi_i\| < \|u_i - \xi_i\|$$

whereas

$$\begin{aligned}\|\bar{u}_i - \xi_i\| \geqq \sigma_i &\Rightarrow L\|u_i - \xi_i\| \cdot \|\bar{u}_i - \xi_i\| \\ &\geqq (F_i'(\xi_i) - F_i'(u_i))(\bar{u}_i - \xi_i) \geqq a'\|\bar{u}_i - \xi_i\|^2\end{aligned}$$

in view of (17). In all cases,

$$\|\bar{u}_i - \xi_i\| \leqq \gamma\|u_i - \xi_i\| \tag{27a}$$

with

$$\gamma = \max\left\{1, \frac{L}{a'}\right\}. \tag{27b}$$

With (23), (26) and (27) one obtains

$$\omega_i \geqq \omega \triangleq \min\left\{1, \frac{2\delta a'}{L(1+\gamma)^2}\right\} > 0 \tag{28a}$$

and

$$r_{i+1} \leqq \beta r_i + 4b\lambda^i \tag{28b}$$

with

$$\beta = \max\{0, 1 - \delta\omega\}. \tag{28c}$$

Together, the inequalities (20), (24) and (28) yield

$$r_{i+1} \leqq \beta r_i + c\lambda^{i/2} \tag{29a}$$

with

$$c = \max\{\alpha_4, 4b\lambda^{1/2}\}. \tag{29b}$$

If $\lambda \neq \beta$, then (10a) in Lemma 2 yields,

$$r_i \leqq C\rho^{i-1}, \qquad i = 1, 2, 3, \cdots \tag{30a}$$

with

$$\rho = \max\{\lambda, \beta\} \tag{30b}$$

and

$$C = r_1 + \frac{c\lambda}{|\beta - \lambda|}. \tag{30c}$$

On the other hand, if $\lambda = \beta$, let ρ be any fixed number in the interval

$$\beta < \rho < 1 \tag{31a}$$

and put

$$C = r_1 + c(e \ln(\rho/\beta))^{-1}. \tag{31b}$$

Since $x\mu^x \leqq -(e \ln \mu)^{-1}$ for $x \geqq 0$ and $\mu \in [0, 1)$, it now follows from (10b) that

$$r_i \leqq C\rho^{i-1}. \tag{31c}$$

Finally, (14), (22) and (30) or (31) imply that

$$F_i(u_i) - F_i(\xi_i) \leqq r_i + 3b\lambda^i \leqq (C + 3b\lambda)\rho^{i-1}, \qquad i = 1, 2, 3, \cdots. \tag{32}$$

Note 3. When $0 \leqq \lambda < \beta$, the factor ρ in (30)–(32) is equal to β, and β is essentially determined by parameters of the limiting problem (Ω, F). In fact, when $\lambda = 0$, (30) provides a convergence rate estimate for the unmodified CG method (2.3) applied to (Ω, F). This suggests that if λ is "sufficiently small" compared to β the DMCG method (2.10) ought to behave like (2.3) for (Ω, F), even though (2.3) may exhibit a much poorer asymptotic convergence rate when implemented for any one of the problems (Ω_i, F_i).

Note 4. If U is a polyhedral convex set in $\mathbb{R}^n$ and if $\Omega = \mathscr{L}^1([0, 1], U)$, then condition (12) typically holds in the $\mathscr{L}^1$ norm at "bang-bang" extremals, i.e., piecewise constant ξ with range in the vertex set for U [4], [6]. According to Lemma 1, (12) can therefore hold in the *exterior* of small neighborhoods of extremals ξ_* in polyhedral product approximations $\Omega_* = U^{n_*} \subset \mathbb{R}^{m \times n_*}$.

If Ω satisfies the uniform convexity condition,

$$x, y \in \Omega \quad \text{and} \quad \left\| z - \left(\frac{x+y}{2}\right)\right\| \leqq c\|x - y\|^2 \Rightarrow z \in \Omega \tag{33}$$

for some $c > 0$ and all x, y, z, and if $F'(\xi) \neq 0$ at an extremal in the boundary of Ω, then (12) holds with $a = 2c\|F'(\xi)\|$ [6]. The convexity condition (33) is satisfied by any ball in a Hilbert space, and more generally by sets defined by finitely many smooth convex nonlinear inequalities,

$$h_i(u) \leqq 0, \qquad i = 1, \cdots, l \tag{34a}$$

with

$$h_i''(u)vv \geqq \mu\|v\|^2 \tag{34b}$$

for some $\mu > 0$, all i, and all $u \in \Omega$ [6].

Finally, if $U \subset \mathbb{R}^m$ satisfies (33), and $\Omega = \mathscr{L}^1([0, 1], U)$, then (12) holds at extremals ξ with continuous nonvanishing Riesz representors g_ξ in (2.8) [10]; extremals of this type are continuous and have range in the boundary of U.

Note 5. The convexity conditions imposed on F and F_i in Theorem 1 may be replaced with the weaker "uniform pseudoconvexity" condition described in [8].

Note 6. A result analogous to Theorem 1 is established in [12] for projected gradient methods; in both cases, the proofs are readily modified for Armijo step lengths [14]. References [15] and [16] establish similar results for full-step Newton and quasi-Newton methods.

4. An optimal control problem. As explained earlier, the continuous-time optimal control problem (1.5) can be approximated by discrete-time problems (Ω_i, F_i), where the Ω_i's are sets of step functions that have range in the control input set U and are constant on intervals of the nets $\mathcal{N}_i$, and where the F_i's are computed by replacing the differential equation (1.5c) with a standard "numerical integrator" on $\mathcal{N}_i$ (for example, an Euler, modified Euler, or Runge-Kutta difference equation). Conditions (3.13) and (3.14) will hold for this type of approximation if the functions P and G are "sufficiently smooth" and the lengths of the largest intervals in $\mathcal{N}_i$ decrease geometrically with increasing i. Moreover, condition (3.12) typically holds in $\mathscr{L}^1$ if U is a convex polyhedron in $\mathbb{R}^m$ and (Ω, F) has a "bang-bang" solution ξ, or if U satisfies the uniform convexity condition (3.33) and $F'(\xi)$ has a continuous nonvanishing Riesz representor (Note 3.4). For control problems of this sort, Theorem 1 suggests that i iterations of the DMCG method (2.10) will yield values of F_i comparable to those produced by the same number of iterations of the CG method (2.3) for (Ω_i, F_i), provided the nets N_j are refined at a suitable rate for $1 \leqq j \leqq i$. If this is so, the DMCG method may have a decided advantage, since the cost of one CG iterate for (Ω_j, F_j) increases roughly in proportion to the number n_j of intervals in the net N_j. More precisely if $n_j \cong \nu^j$ for some $\nu > 1$, then

$$\frac{\text{CG cost } (i \text{ iterations for } (\Omega_i, F_i))}{\text{DMCG cost } (i \text{ iterations})} = in_i \left(\sum_{j=1}^{i} n_j \right)^{-1} \cong i\nu^i \cdot \left(\sum_{j=1}^{i} \nu^j \right)^{-1} = \frac{i(\nu-1)\nu^i}{\nu(\nu^i-1)} \geqq \frac{i(\nu-1)}{\nu}. \tag{1}$$

Example 1. Consider a particle of unit mass that moves along a straight line in response to a time-dependent force $u(t)$ whose magnitude cannot exceed 1. Suppose that at time $t = 0$ the particle's position coordinate x_1 and velocity x_2 are both zero. How should one choose $u(t)$ on the interval $0 \leqq t \leqq 1$ in order to make $x(1)$ large while keeping $|x_2(1)|$ small? There are several ways to formulate this problem in precise mathematical terms. For present purposes, the formulation (1.5) will serve with $p = 1$, $m = 1$, $n = 2$, $x_1(0) = x_2(0) = 0$, $U = [-1, 1]$, $f_1 = x_2$, $f_2 = u$, $P = -x_1 + \frac{1}{2}\Lambda x_2^2$, where Λ is a positive penalty constant (the larger the value assigned to Λ, the more stress is placed on bringing the particle to rest at $t = 1$). The corresponding optimal control problem (Ω, F) has a bang-bang solution,

$$\xi(t) = \begin{cases} 1, & 0 \leqq t < t_*, \\ -1, & t_* \leqq t \leqq 1, \end{cases} \tag{2a}$$

with

$$t_* = \frac{1+\Lambda}{1+2\Lambda}; \tag{2b}$$

and

$$F(\xi) = -\frac{1+\Lambda}{2(1+2\Lambda)}. \tag{3}$$

Moreover, this solution satisfies condition (3.12) in $\mathscr{L}^1$ [4].

Now let Ω_i consist of step function controls that have range in $[-1, 1]$ and are constant on intervals of the net $\mathcal{N}^{(i)} = \{t_0^{(i)}, t_1^{(i)}, \cdots t_{n_i}^{(i)}\}$ with

$$n_i = 2^i,$$
$$t_0^{(i)} = 0,$$
$$t_{l+1}^{(i)} = t_l^{(i)} + \Delta t_i, \qquad l = 0, 1, \cdots, n_i - 1,$$

and

$$\Delta t_i = \frac{1}{n_i}.$$

Construct F_i by replacing the differential equations $\dot{x}_1 = x_2$, $\dot{x}_2 = u(t)$ by a corresponding pair of modified Euler difference equations on the net $\mathcal{N}^{(i)}$. According to (1), i iterations of the CG method (2.3) applied to (Ω_i, F_i) will cost roughly $i/2$ times as much as i iterations of the DMCG method. Furthermore, numerical computations performed for this example indicate that i iterations of the two method produce comparable values of F_i. Some of these results are presented in Table 1 for $\Lambda = 6$ and $\delta = .1$.

TABLE 1

	A		B		C		D	
i	CG	DMCG	CG	DMCG	CG	DMCG	CG	DMCG
1	0.0	0.0	.01563	.01563	−.06250	−.06250	.2539	.2539
2	.01563	.01563	.03247	.03247	.02051	.02051	.2540	.2540
3	.02832	.02832	.06123	.06172	.2500	.2500	.2581	.2561
4	.06123	.06123	.08220	.07571	.1865	.2539	.2592	.2583
5	.1367	.1367	.1895	.2500	.1973	.2552	.2584	.2594
6	.1443	.1895	.2500	.2539	.2023	.2569	.2592	.2602
7	.1762	.1973	.1904	.2552	.2184	.2573	.2638	.2625
8	.1923	.2095	.2007	.2569	.2219	.2584	.2644	.2638
9	.2007	.2177	.2061	.2579	.2264	.2599	.2655	.2643
10	.2061	.2227	.2284	.2629	.2390	.2603	.2667	.2650

(NOTE: Multiply entries by −1 to obtain payoff values $F_i(u_{i+1})$).

Table 1 contains four pairs of columns labelled A, B, C, D, corresponding to four different starting functions in Ω_1, namely $u_1^{\mathrm{A}}(\cdot) \equiv 1$, $u_1^{\mathrm{B}}(\cdot) \equiv 0$, $u_1^{\mathrm{C}}(\cdot) \equiv -.5$ and the step function $u^{\mathrm{D}}(\cdot)$ with $u^{\mathrm{D}}(t) = 1$ for $t \in [0, \frac{1}{2}]$ and $u^{\mathrm{D}}(t) = -1$ for $t \in (\frac{1}{2}, 1]$. The second column in each pair lists the values of $-F_i(u_{i+1})$ obtained after i iterations of the DMCG method, for $i = 1, 2, \cdots, 10$. The first column in each pair lists the corresponding values of $-F_i(u_{i+1})$ obtained by doing i iterations of the *unmodified* CG method for the *fixed* problem (Ω_i, F_i). It can be seen that for each i, the modified conditional gradient method typically produces a sub-optimal solution for (Ω_i, F_i) that is comparable to, or significantly better than, the approximate solution produced by the same number of iterations with the unmodified method. What is more, since DMCG uses crude nets in the early iterations, it is less costly to implement; in particular, for $i = 10$ DMCG has a better than 5 to 1 cost advantage over CG, as explained above.

Computations were halted in the present example after 10 iterations, since the simple grid refinement rule used here produces a net $\mathcal{N}_{10}$ with 1024 intervals; nevertheless for three out of four starting points, the DMCG values $F_{10}(u_{11})$ in Table 1 are already close to the optimal value $F(\xi) = -.2692\cdots$ for the *limiting continuous-time problem* (Ω, F) with $\Lambda = 6$ (see (3)). It seems likely that still better approximations to $F(\xi)$ could be obtained more efficiently with variable Δt nets $\mathcal{N}_i$ that increase in size more slowly than 2^i but concentrate their smallest intervals where the $(i-1)$st iterate function is changing most rapidly (for example, near "switching points"). It might also pay to increase the precision of the difference equation approximation as the iteration commences.

REFERENCES

[1] R. A. TAPIA, *Diagonalized multiplier methods and quasi-Newton methods for constrained optimization*, J. Optim. Theory Appl., 22 (1977), pp. 135–194.

[2] R. KLESSIG AND E. POLAK, *An adaptive precision gradient method for optimal control*, this Journal, 11 (1973), pp. 80–93.

[3] K. SCHITTKOWSKI, *An adaptive precision method for nonlinear optimization problems*, this Journal, 17 (1979), pp. 82–98.

[4] J. C. DUNN, *Extremal types for certain $\mathcal{L}^p$ minimization problems and associated large sale nonlinear programs*, Appl. Math. Optim., 9 (1983), pp. 303–335.

[5] J. C. DUNN AND E. SACHS, *The effects of perturbations on the convergence rates of optimization algorithms*, Appl. Math. Optim., 10 (1983), pp. 143–157.

[6] J. C. DUNN, *Rates of convergence for conditional gradient algorithms near singular and nonsingular extremals*, this Journal, 17 (1979), pp. 187–211.

[7] ———, *Convergence rates for conditional gradient sequences generated by implicit step length rules*, this Journal, 18 (1980), pp. 473–487.

[8] ———, *Global and asymptotic convergence rate estimates for a class of projected gradient processes*, this Journal, 19 (1981), pp. 368–400.

[9] ———, *Newton's method and the Goldstein step length rule for constrained minimization problems*, this Journal, 18 (1980), pp. 659–674.

[10] G. C. HUGHES AND J. C. DUNN, *Newton-Goldstein convergence rates for convex constrained minimization problems with singular solutions*, Appl. Math. Optim., 12 (1984), pp. 203–230.

[11] C. G. HUGHES, *Convergence rate analysis for iterative minimization schemes with quadratic subproblems*, Ph.D. dissertation, Mathematics Dept., North Carolina State Univ., Raleigh, NC, 1981.

[12] J. C. DUNN, *Diagonal modifications of iterative minimization schemes*, unpublished report, presented at the 6th Annual Symposium on Mathematical Programming with Data Perturbations, Washington, DC, May 1984.

[13] A. A. GOLDSTEIN, *Minimizing functions on Hilbert space*, in Computer Methods in Optimization Problems, Academic Press, New York, 1964.

[14] L. ARMIJO, *Minimization of functionals having continuous partial derivatives*, Pacific J. Math., 16 (1966), pp. 1–3.

[15] E. SACHS, *Rates of convergence for adaptive Newton methods*, submitted.

[16] R. S. DEMBO, S. C. EISENSTAT AND T. STEIHAUG, *Inexact Newton methods*, SIAM J. Numer. Anal., (1982), pp. 400–408.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 6, November 1986

007

CONTROLLABILITY OF CONVEX PROCESSES*

JEAN-PIERRE AUBIN†, HALINA FRANKOWSKA† AND CZESŁAW OLECH‡

Abstract. The purpose of this paper is to provide several characterizations of controllability of differential inclusions whose right-hand sides are convex processes. Convex processes are the set-valued maps whose graphs are convex cones; they are the set-valued analogues of linear operators. Such differential inclusions include linear systems where the controls range over a convex cone (and not only a vector space). The characteristic properties are couched in terms of invariant cones by convex processes, or eigenvalues of convex processes, or a rank condition. We also show that controllability is equivalent to observability of the adjoint inclusion.

Key words. convex process, differential inclusion, tangent cone, duality, invariant cones, viability domain

AMS(MOS) subject classifications. 49A55, 93B05, 93C05, 93C10

Introduction. A *convex process* A from $\mathbb{R}^n$ to itself is a set-valued map satisfying

$$\forall x, y \in \operatorname{Dom} A, \quad \lambda, \mu \geqq 0, \qquad \lambda A(x) + \mu A(y) \subset A(\lambda x + \mu y), \tag{0.1}$$

or, equivalently, a set-valued map whose graph is a convex cone. Convex processes are the set-valued analogues of linear operators. We shall say that a convex process is *closed* if its graph is closed and that it is *strict* if its domain is the whole space.

We associate with a strict closed convex process A the Cauchy problem for the differential inclusion: find an absolutely continuous function $x(\cdot)$ satisfying

$$\text{for almost all } t \in [0, T], \quad x'(t) \in A(x(t)), \quad x(0) = 0. \tag{0.2}$$

We denote by R_T the *reachable set* at *time* T defined by

$$R_T := \{x(T) : x(\cdot) \text{ is a solution to (0.2)}\}. \tag{0.3}$$

We also say that

$$R := \bigcup_{T>0} R_T \text{ is the reachable set} \tag{0.4}$$

and that the differential inclusion (0.2) (or the convex process A) is *controllable* if the reachable set R is equal to the whole space $\mathbb{R}^n$.

Convex processes were introduced and thoroughly studied in Rockafellar [1967], [1970], [1974] and in Aubin and Ekeland [1984], for instance. Derivatives of set-valued maps (see Aubin and Ekeland [1984, Chap. 7]) provide examples of closed convex processes. These are used, for instance, in Frankowska [1984], [1985] for deriving local controllability of differential inclusions from the controllability of convex processes which "approximate" in some sense the original differential inclusion around the equilibrium:

THEOREM (Frankowska). *Let F be a set-valued map from $\mathbb{R}^n$ into the compact subsets of $\mathbb{R}^n$, Lipschitzian around zero and $0 \in F(0)$. Denote by $F'(0)$ the derivative of F at zero and by L the closed convex cone spanned by* co $F(0)$ *(convex hull of $F(0)$). Set*

$$A(x) = \operatorname{cl}(F'(0)x + L).$$

*Received by the editors November 16, 1984, and in revised form September 3, 1985.

† International Institute of Applied Systems Analysis, (IIASA) Laxenburg, Austria, and Centre de Recherche de Mathématiques de la Décision (CEREMADE), Université de Paris-Dauphine, 75775 Paris, France.

‡ International Institute of Applied Systems Analysis, (IIASA) Laxenburg, Austria, and Institute of Mathematics, Polish Academy of Sciences, Warsaw, Poland.

Then the differential inclusion

$$x' \in F(x), \qquad x(0) = 0$$

is locally controllable around zero at time T if the "linearized" inclusion

$$x' \in A(x)$$

is controllable at time T.

We know that for linear problems the reachable sets are invariant. Hence we wish to extend the usual concept of invariant subspace by a linear operator. This can be done in two different ways: let A be a convex process and P be a closed convex cone contained in Dom A. We recall that the *tangent cone* $T_P(x)$ at a point $x \in P$ is defined by

$$T_P(x) := \operatorname{cl}\left(\bigcup_{h>0} \frac{1}{h}(P - x) \right) = \operatorname{cl}(P + \mathbb{R}x). \tag{0.5}$$

We shall say that P is *invariant by* A if

$$\forall x \in P, \qquad A(x) \subset T_P(x) \tag{0.6}$$

and that P is a *viability domain* for A if

$$\forall x \in P, \qquad A(x) \cap T_P(x) \neq \varnothing. \tag{0.7}$$

When P is a vector space, then $T_p(x) = P$, so that a subspace is invariant by A if $\forall x \in P$, $A(x) \subset P$ and is a viability domain for A if $\forall x \in P$, $A(x) \cap P \neq \varnothing$.

The first example of an invariant cone is provided by the closure of the reachable set.

THEOREM 0.1. *Let A be a strict closed convex process. Then the closure of the reachable set is the smallest closed convex cone invariant by A and the subspace $R - R$ spanned by R is the smallest subspace invariant by A.*

Furthermore, if $R - R = \mathbb{R}^n$ and $R \neq \mathbb{R}^n$, there exists $\lambda \in \mathbb{R}$ such that $\operatorname{Im}(A - \lambda I) \neq \mathbb{R}^n$.

We could say that a real number λ such that $\operatorname{Im}(A - \lambda I) \neq \mathbb{R}^n$ is an *eigenvalue of* A.

We shall prove this theorem by "duality." Indeed, convex processes can be transposed, as can linear operators. Let A be a convex process; we define its *transpose* A^* by

$$p \in A^*(q) \Leftrightarrow \forall (x, y) \in \operatorname{Graph} A, \langle p, x \rangle \leqq \langle q, y \rangle. \tag{0.8}$$

We also replace the orthogonality between subspaces by polarity between cones. If G is a subset of $\mathbb{R}^n$, we denote by G^+ its (positive) *polar cone* defined by

$$G^+ := \{ p \in \mathbb{R}^n \mid \forall x \in G, \langle p, x \rangle \geqq 0 \}. \tag{0.9}$$

We recall that the separation theorem implies that

$$G^{++} \text{ is the closed convex cone spanned by } G. \tag{0.10}$$

Therefore, it is convenient to bear in mind that

$$(q, p) \in \operatorname{Graph}(A^*) \Leftrightarrow (-p, q) \in \operatorname{Graph}(A)^+ \tag{0.11}$$

so that when A is a closed convex process, then $A = A^{**}$.

Example. Let F be a linear operator from $\mathbb{R}^n$ to itself, let L be a closed convex cone of controls and let A be the strict closed convex process defined by

$$A(x) := Fx + L. \tag{0.12}$$

Then its transpose is equal to

$$A^*(q) = \begin{cases} F^*q & \text{if } q \in L^+, \\ \varnothing & \text{if } q \notin L^+. \end{cases} \tag{0.13}$$

When $L = \{0\}$, i.e., when $A = F$, we deduce that $A^* = F^*$, so that transposition of convex processes is a legitimate extension of transposition of linear operators.

When A is a strict closed convex process, we shall prove that A^* is upper semi-continuous with convex compact values, that $A^*(0) = \{0\}$, $\operatorname{Dom} A^* = A(0)^+$ is closed and that the restriction of A^* to the vector space $\operatorname{Dom} A^* \cap -\operatorname{Dom} A^*$ is a linear (single-valued) operator.

As expected, we associate with the differential inclusion (0.2) the *adjoint inclusion*

$$\text{for almost all } t \in [0, T], \qquad -q'(t) \in A^*(q(t)). \tag{0.14}$$

We introduce the cones Q_T and Q defined by

$$\begin{aligned} &\text{i)} \quad Q_T := \{\eta \mid \exists q(\cdot), \text{ a solution to (0.14) satisfying } q(T) = \eta\}, \\ &\text{ii)} \quad Q := \bigcap_{T>0} Q_T. \end{aligned} \tag{0.15}$$

To say that $Q = \{0\}$ amounts to saying that the only solution to (0.14) defined on $[0, \infty[$ is $q \equiv 0$, or, in the language of systems theory, that the *adjoint system is observable.*

The "duality" method lies in the following statement.

THEOREM 0.2. *Let A be a strict closed convex process. Then*

$$R_T^+ = Q_T \quad \textit{and} \quad R^+ = Q. \tag{0.16}$$

Furthermore, a closed convex cone $P \supset A(0)$ is invariant by A if and only if its polar cone $P^+ \subset \operatorname{Dom} A^*$ is a viability domain for A^*.

Indeed, it allows one to derive Theorem 0.1 from

THEOREM 0.3. *Let A be a strict closed convex process. The cone Q is the largest closed convex cone which is a viability domain for A^* and $Q \cap -Q$ is the largest subspace invariant by (the linear operator) A^*.*

Furthermore, if Q is not reduced to $\{0\}$ and contains no line, there exists a solution $q \neq 0$ and $\lambda \in \mathbb{R}$ to the inclusion $\lambda q \in A^*(q)$.

We could say that such a q is an *eigenvector of A^**.

It will be convenient to introduce the following definition. We say that A satisfies the *rank condition* if

$$\begin{aligned} &\text{the subspace spanned by the cone } A^m(0) \text{ is the whole space} \\ &\mathbb{R}^n \text{ for some integer } m \geqq 1. \end{aligned} \tag{0.17}$$

This is motivated by the terminology used for linear systems. Indeed, when $A(x) := Fx + L$ where F is a linear operator and L is a convex cone of controls, we observe that $A^m(0) = L + FL + \cdots + F^{m-1}L$.

We shall derive from these results the following characterization of controllability of convex processes.

THEOREM 0.4. *Let A be a strict closed convex process. The following conditions are equivalent.*

(a) *differential inclusion* (0.2) *is controllable* (*i.e.*, $R = \mathbb{R}^n$),

(b) *differential inclusion* (0.2) *is controllable at some time* $T > 0$ (*i.e.*, $R_T = \mathbb{R}^n$),

(c) *the adjoint inclusion* (0.14) *is observable* (*i.e.*, $Q = \{0\}$),

(d) *the adjoint inclusion* (0.14) *is observable at some time* $T > 0$ (*i.e.*, $Q_T = \{0\}$),

(e) $\mathbb{R}^n$ *is the smallest closed convex cone invariant by A,*

(f) $\{0\}$ *is the largest closed convex cone which is a viability domain for* A^*,

(g) A *has neither proper invariant subspace nor eigenvalues,*

(h) A^* *has neither proper invariant subspace nor eigenvectors,*

(i) *the rank condition holds true and* A *has no eigenvalues,*

(j) *the rank condition is satisfied and* A^* *has no eigenvectors,*

(k) *for some* $m \geqq 1$, $A^m(0) = (-A)^m(0) = \mathbb{R}^n$.

Example. Let F be a linear operator from $\mathbb{R}^n$ to itself and let L be a closed convex cone of controls. We consider the differential inclusion

$$x'(t) \in Fx(t) + L, \qquad x(0) = 0, \tag{0.18}$$

and its adjoint inclusion

$$-q'(t) = F^*q(t), \qquad \forall t \geqq 0, \quad q(t) \in L^+. \tag{0.19}$$

COROLLARY 0.5. *The following conditions are equivalent:*

(a) *the system* (0.18) *is controllable,*

(b) *the adjoint equation* (0.19) *is observable* (*the only solution of* $-q' = F^*q$ *remaining in* L^+ *on* $[0, \infty[$ *is* $q \equiv 0$),

(c) $\{0\}$ *is the largest closed convex cone contained in* L^+ *which is invariant by* F^*,

(d) F^* *has neither proper invariant subspace contained in* L^+ *nor eigenvector in* L^+,

(e) *the subspace spanned by* $L, FL, \cdots, F^{n-1}L$ *is equal to* $\mathbb{R}^n$ *and* F^* *has no eigenvector in* L^+ (*see Brammer* [1972]),

(f) *for some* $m \geqq 1$, $L + FL + \cdots + F^mL = L - FL + \cdots + (-1)^mF^mL = \mathbb{R}^n$ (*see Korobov* [1980]).

This example also illustrates another advantage of duality, because some properties bearing on the adjoint system have a simpler formulation. This explains why some criteria mentioned in Theorem 0.4 disappear in Corollary 0.5.

When L is a vector space, statements (c), (d) and (f) are the same and the mention of eigenvector in statement (e) is redundant. This is not the case when L is a proper cone. It is sufficient to consider the example

$$x' \in -x + \mathbb{R}_+, \qquad x(0) = 0.$$

The rank condition is satisfied ($A^2(0) = \mathbb{R}$) and the reachable set is $\mathbb{R}_+$.

We summarize in § 1 the results on convex processes and their transpose that we will need later. Section 2 is devoted to the proof of the duality Theorem 0.2, characterizing the positive polar cones of the reachable set. We then derive the characterization of the closure of the reachable set as the smallest invariant cone by A and its dual version in § 3 and the existence of eigenvalues of A and eigenvectors of A^* in § 4. These results are used to prove Theorem 0.4 in § 5.

1. Convex processes and their transposes.

DEFINITION 1.1. A set-valued map from $\mathbb{R}^n$ to $\mathbb{R}^n$ is said to be a *convex process* if its graph is a convex cone. It is *closed* if its graph is closed. It is called *strict* if

$$\text{Dom } A := \{x \in \mathbb{R}^n \mid A(x) \neq \varnothing\} \text{ is the whole space.}$$

DEFINITION 1.2. Let X be a Hilbert space and let $G \subset X$ be a subset. We denote by G^+, the (positive) *polar cone* of G, the closed convex cone defined by

$$G^+ := \{p \in X^* \mid \forall x \in G, \langle p, x \rangle \geqq 0\}. \tag{1.1}$$

The separation theorem implies that the "bipolar" G^{++} is the closed convex cone spanned by G. We shall use the following consequence of this fact.

LEMMA 1.3 (closed image lemma). *Let X, Y be two Hilbert spaces, let ϕ be a continuous linear operator from X to Y and let L be a closed convex cone of Y. Assume that*

$$\text{Im } \phi - L = Y \quad (\textit{surjectivity condition}). \tag{1.2}$$

Then

$$\phi^{-1}(L)^+ = \phi^*(L^+). \tag{1.3}$$

Proof. (a) We prove first that $\phi^*(L^+)$ is closed. Let $q_n \in L^+$ be a sequence such that $\phi^*(q_n)$ converges to some p in X^* and let us prove that p belongs to $\phi^*(L^+)$.

We begin by showing that q_n is weakly bounded. Indeed, for any $v \in Y$, there exist $x \in X$ and $y \in L$ such that $v = \phi(x) - y$. Hence

$$\langle q_n, v\rangle = \langle \phi^*(q_n), x\rangle - \langle q_n, y\rangle \leqq \langle \phi^*(q_n), x\rangle \leqq \|\phi^*(q_n)\| \cdot \|x\|.$$

Therefore, since X is reflexive, the sequence q_n is in a weakly compact subset and a subsequence $q_{n'}$ converges weakly to some $q \in Y^*$. Since L^+ is closed and convex, and thus is weakly closed, q belongs to L^+. Since $\phi^*(q_{n'})$ converges weakly to $\phi^*(q)$ and strongly to p, we deduce that $p = \phi^*(q) \in \phi^*(L^+)$.

(b) We observe that $\phi^*(L^+)^+ = \phi^{-1}(L)$ because $x \in \phi^*(L^+)^+$ if and only if $\langle \phi^* q, x\rangle = \langle q, \phi(x)\rangle \geqq 0$ for all $q \in L^+$, i.e., if and only if $\phi(x)$ belongs to $L^{++} = L$. Hence, since $\phi^*(L^+)$ is closed, we deduce that

$$\phi^*(L^+) = \phi^*(L^+)^{++} = \phi^{-1}(L)^+. \qquad \square$$

We now recall some properties of convex processes, some of them already are known (see Rockafellar [1967], [1970, § 39], [1974], Aubin and Ekeland [1984, Chap. 3]).

DEFINITION 1.4. Let A be a convex process from $\mathbb{R}^n$ to itself. The transpose A^* of A is the set-valued map from $\mathbb{R}^n$ to itself given by

$$p \in A^*(q) \Leftrightarrow \forall (x, y) \in \text{Graph } (A), \qquad \langle p, x\rangle \leqq \langle q, y\rangle. \tag{1.4}$$

In other words,

$$(q, p) \in \text{Graph } (A^*) \Leftrightarrow (-p, q) \in (\text{Graph } A)^+. \tag{1.5}$$

The transpose of A^* is obviously a closed convex process and $A = A^{**}$ if and only if the convex process A is closed. When A is a linear operator its transpose as a linear operator coincides with its transpose as a convex process.

LEMMA 1.5. *If A is a closed convex process, then*

$$A(0) = (\text{Dom } A^*)^+. \tag{1.6}$$

Proof. We observe that y belongs to $A(0)$ if and only if $0 = \langle p, 0\rangle \leqq \langle q, y\rangle$ for all $q \in \text{Dom } A^*$ and $p \in A^*(q)$, i.e., if and only if $\langle q, y\rangle \geqq 0$ for all $q \in \text{Dom } A^*$. $\square$

DEFINITION 1.6. Let B denote the unit ball. When A is a closed convex process, we define

$$\|A\| := \sup_{x \in B \cap \text{Dom } A} \inf_{y \in A(x)} \|y\| \in [0, +\infty]. \tag{1.7}$$

PROPOSITION 1.7. *Let A be a strict closed convex process. Then*

(a) $\forall x, y \in \mathbb{R}^n$, $A(x) \subset A(y) + \|A\| \, \|x - y\| B$ *(i.e., A is Lipschitzian with finite Lipschitz constant equal to $\|A\|$).*

(b) $\text{Dom } A^* = A(0)^+$ *and A^* is upper semicontinuous with compact convex images, mapping the unit ball into the ball or radius $\|A\|$.*

(c) *The restriction of* A^* *to the vector space* $\operatorname{Dom} A^* \cap -\operatorname{Dom} A^*$ *is single-valued and linear* (*and thus,* $A^*(0)=0$).

Proof. (a) The first statement is a reformulation of the Robinson-Ursescu theorem (see Robinson [1972], Ursescu [1975], and Aubin and Ekeland [1984, Cor. 3.3.3, p. 132]).

(b) We observe that

$$\forall q \in \operatorname{Dom} A^*, \qquad \sup_{p \in A^*(q)} \|p\| \leqq \|A\| \|q\|, \tag{1.8}$$

because for all $x \in \operatorname{Dom} A = \mathbb{R}^n$ and for all $p \in A^*(q)$ we have

$$\|p\| = \sup_{x \in \mathbb{R}^n} \frac{\langle p, x\rangle}{\|x\|} \leqq \sup_{x \in \mathbb{R}^n} \inf_{y \in A(x)} \frac{\langle q, y\rangle}{\|x\|}$$

$$\leqq \sup_{x \in \mathbb{R}^n} \inf_{y \in A(x)} \frac{\|q\| \|y\|}{\|x\|} = \|A\| \|q\|.$$

Then A^* maps bounded sets to bounded sets. Since its graph is a closed convex cone we deduce that A^* is upper semicontinuous with compact convex images. By Lemma 1.5, $\operatorname{Dom} A^* = A(0)^+$. Therefore it remains to prove that $\operatorname{Dom} A^*$ is closed. Indeed let $q_n \in \operatorname{Dom} A^*$ be a sequence converging to some q and let $p_n \in A^*(q_n)$. The sequence $\{p_n\}$ being bounded contains a subsequence $\{p_{n'}\}$ converging to some p. Thus

$$(q, p) = \lim_{n' \to \infty} (q_{n'}, p_{n'}), \qquad (q_{n'}, p_{n'}) \in \operatorname{graph} A^*.$$

The graph of A^* being closed, we proved that $q \in \operatorname{Dom} A^*$.

Statement (c) follows from Rockafellar [1970, Thm. 39.1, p. 414]. □

We observe that we always have

$$\sup_{p \in A^*(q_0)} \langle p, x_0\rangle \leqq \inf_{y \in A(x)} \langle q_0, y\rangle.$$

LEMMA 1.8. *Let* A *be a closed convex process.*

For any $x_0 \in \operatorname{Int} \operatorname{Dom} A$, *and* $q_0 \in \operatorname{Dom} A^*$,

$$\sup_{p \in A^*(q_0)} \langle p, x_0\rangle = \inf_{y \in A(x_0)} \langle q_0, y\rangle. \tag{1.9}$$

(See Rockafellar [1970, Thm. 39.3, p. 419]).

We now extend to the case of closed convex cones the concepts of invariant subspace. When K is a subspace and F is a linear operator, we recall that K is invariant by F when $Fx \in K$ for all $x \in K$. When A is a convex process, there are two ways of extending this notion: we shall say that K is invariant by A if, for any $x \in K$, $A(x) \subset K$ and that K is a viability domain for A if, for any $x \in K$, $A(x) \cap K \neq \emptyset$. We also need to extend these notions to the case when K is a closed convex cone. For that purpose, we recall the

DEFINITION 1.9. If K is a closed convex set and x belongs to K, we say that

$$T_K(x) := \operatorname{cl}\left(\bigcup_{h>0} \frac{1}{h}(K - x) \right)$$

is the *tangent cone* to K at x.

LEMMA 1.10. *When* K *is a vector subspace, then for all* $x \in K$, $T_K(x) = K$ *and when* K *is a closed convex cone then,*

$$\forall x \in K, \qquad T_K(x) = \operatorname{cl}(K + \mathbb{R}x). \tag{1.10}$$

(See Aubin and Ekeland [1984, Prop. 4.1.9, p. 171]).

Now, we can introduce

DEFINITION 1.11. Let K be a closed convex cone and let A be a convex process. We say that K is *invariant* by A if

$$\forall x \in K, \quad A(x) \subset T_K(x) \tag{1.11}$$

and that K is a *viability domain* for A if

$$\forall x \in K, \quad A(x) \cap T_K(x) \neq \emptyset. \tag{1.12}$$

These are dual notions, as the following proposition shows.

PROPOSITION 1.12. *Let A be a strict closed convex process and let K be a closed convex cone containing $A(0)$. Then K is invariant by A if and only if K^+ is a viability domain for A^*.*

Proof. By Proposition 1.7 (b) the condition $A(0) \subset K$ implies that $K^+ \subset A(0)^+ = \text{Dom } A^*$. To say that K is invariant by A amounts to saying that

$$\forall x \in K, \quad \forall q \in T_K(x)^+, \quad \inf_{y \in A(x)} \langle q, y \rangle \geqq 0. \tag{1.13}$$

Lemma 1.10 states $T_K(x) = \text{cl}\,(\mathbb{R}x + K)$, $T_{K^+}(q) = \text{cl}\,(\mathbb{R}q + K^+)$. Therefore

$$q \in T_K(x)^+ \Leftrightarrow \{\langle q, x \rangle = 0, x \in K, q \in K^+\} \Leftrightarrow x \in T_{K^+}(q)^+. \tag{1.14}$$

On the other hand, Lemma 1.8 implies that

$$\inf_{y \in A(x)} \langle q, y \rangle = \sup_{p \in A^*(q)} \langle p, x \rangle.$$

Therefore condition (1.13) is equivalent to the condition

$$\forall q \in K^+, \quad \forall x \in T_{K^+}(q)^+, \quad \sup_{p \in A^*(q)} \langle p, x \rangle \geqq 0. \tag{1.15}$$

By Proposition 1.7 (b) for all $q \in K^+$ the set $A^*(q)$ is compact. The separation theorem implies that $A^*(q)$ has a nonempty intersection with $T_K{}^+(q)$ if and only if for all $x \in \mathbb{R}^n$,

$$\sup_{p \in A^*(q)} \langle p, x \rangle \geqq \inf_{z \in T_{K^+}(q)} \langle z, x \rangle.$$

Since $T_{K^+}(q)$ is a cone the latter inequality is equivalent to (1.15). This ends the proof. □

We now introduce the concepts of eigenvalues and eigenvectors of closed convex processes.

DEFINITION 1.13. We shall say that $\lambda \in \mathbb{R}$ is an *eigenvalue* of a convex process A if $\text{Im}\,(A - \lambda I) \neq \mathbb{R}^n$ and that $x \in \text{Dom } A$ is an *eigenvector* of A if $x \neq 0$ and if there exists $\lambda \in \mathbb{R}$ such that $\lambda x \in A(x)$.

We observe that half-lines spanned by eigenvectors of A^* are viability domains for A^*.

LEMMA 1.14. *Let A be a strict convex process. Then A^* has an eigenvector if and only if $\text{Im}\,(A - \lambda I) \neq \mathbb{R}^n$ for some $\lambda \in \mathbb{R}$.*

Proof. (a) Let η be an eigenvector of A^*, a solution to $\lambda\eta \in A^*(\eta)$, $\eta \neq 0$. Thus, for all $y \in A(x)$, $\langle \eta, y - \lambda x \rangle \geqq 0$ and thus, $\text{Im}\,(A - \lambda I) \subset \{\eta\}^+ \neq \mathbb{R}^n$.

(b) Conversely, assume that for some $\lambda \in \mathbb{R}$, $\text{Im}\,(A - \lambda I) \neq \mathbb{R}^n$. Since it is a convex cone of a finite-dimensional space, there exists a nonzero $\eta \in \mathbb{R}^n$ such that $\langle \eta, z \rangle \geqq 0$ for all $z \in \text{Im}\,(A - \lambda I)$. This implies that for all $x \in \mathbb{R}^n$ and $y \in A(x)$,

$$\lambda \langle \eta, x \rangle \leqq \langle \eta, y \rangle.$$

By the very definition of A^*, we deduce that $\lambda\eta$ belongs to $A^*(\eta)$. □

Example 1.15. Let F be a linear operator from $\mathbb{R}^n$ to itself, let L be a closed convex cone of controls and let A be the strict closed convex process defined by $A(x) := Fx + L$.

A cone K is invariant by A if

$$\forall x \in K, \qquad Fx + L \subset T_K(x),$$

and λ is an eigenvalue of A if

$$\operatorname{Im}(F - \lambda I) + L \neq \mathbb{R}^n.$$

The transpose A^* of A is defined by

$$A^*q = \begin{cases} F^*q & \text{if } q \in L^+, \\ \varnothing & \text{if } q \notin L^+. \end{cases}$$

A cone $P \subset L^+ = \operatorname{Dom} A^*$ is a viability domain for A^* if and only if

$$\forall q \in P, \qquad F^*q \in T_P(q).$$

An element $q \neq 0$ is an eigenvector of A^* if and only if q is an eigenvector of F^* which belongs to the cone L^+.

Example 1.16. Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be a single-valued map and $L \subset R^n$ be a closed convex cone, to which we associate the set-valued map A defined by

$$\forall x \in \mathbb{R}^n, \qquad A(x) := f(x) + L. \tag{1.16}$$

We observe that A is a closed convex process if and only if

$$\forall q \in L^+, x \to \langle q, f(x) \rangle \text{ is convex, lower semicontinuous and positively homogeneous} \tag{1.17}$$

or, in other words, if and only if

$$\forall q \in L^+, x \to \langle q, f(x) \rangle \text{ is the support function of a compact convex subset which we denote by } F^*(q). \tag{1.18}$$

We then verify that the transpose A^* of A is equal to

$$A^*(q) := \begin{cases} F^*(q) & \text{if } q \in L^+, \\ \varnothing & \text{if not.} \end{cases} \tag{1.19}$$

We observe that

$$\forall q_1, q_2 \in L^+, \qquad A^*(q_1 + q_2) = F^*(q_1) + F^*(q_2) = A^*(q_1) + A^*(q_2). \tag{1.20}$$

Furthermore, $F^*(q) = A^*(q)$ is single-valued when q ranges over $L^+ \cap -L^+$.

To say that q is an eigenvector of A^* is to say that

$$q \neq 0, \quad q \in L^+, \quad \lambda q \in F^*(q), \tag{1.21}$$

and that $\lambda \in \mathbb{R}$ is an eigenvalue of A is to say that

$$\operatorname{Im}(f - \lambda I) + L \neq \mathbb{R}^n. \tag{1.22}$$

We check that

$$\sum_{k=1}^{m} f^k(L) \subset \sum_{k=1}^{m} A^k(0). \tag{1.23}$$

2. The duality theorem. We devote this section to the duality theorem, which characterizes the polar cones of the reachable sets.

We denote by $W^{1,p}(0, T)$, $p \in [1, \infty]$, the Sobolev space of functions $x \in L^p(0, T; \mathbb{R}^n)$ such that $x'(\cdot)$ belongs to $L^p(0, T; \mathbb{R}^n)$.

Let us consider the Cauchy problem for the differential inclusion

$$\text{(2.1)} \quad \begin{cases} \text{(i)} & x'(t) \in A(x(t)) \quad \text{for almost all } t \in [0, T], \\ \text{(ii)} & x(0) = 0. \end{cases}$$

We recall that the reachable set R_T is defined by

$$R_T := \{x(T) \mid x \in W^{1,1}(0, T) \text{ is a solution to (2.1)}\}. \tag{2.2}$$

We shall characterize its positive polar cone R_T^+. For that purpose, we associate with the differential inclusion (2.1) the adjoint inclusion

$$\text{(2.3)} \quad \begin{cases} \text{(i)} & -q'(t) \in A^*(q(t)) \quad \text{for almost all } t \in [0, T], \\ \text{(ii)} & q(T) = \eta, \end{cases}$$

and we denote by $Q_T \subset \operatorname{Dom} A^*$ the set of "final" values η such that the differential inclusion (2.3) has a solution.

$$Q_T := \{\eta \mid \exists q \in W^{1,1}(0, T) \text{ a solution to (2.3)}\}. \tag{2.4}$$

The statement of the duality theorem is the following.

THEOREM 2.1. *Let A be a strict closed convex process. Then*

$$R_T^+ = Q_T. \tag{2.5}$$

We need the following technical lemma.

LEMMA 2.2. *Let A be a strict closed convex process. Then the $W^{1,\infty}(0, T)$ solutions to* (2.1) *are dense in $W^{1,1}(0, T)$ solutions to* (2.1) *in the metric of uniform convergence on* $[0, T]$.

Proof. Indeed let $w \in W^{1,1}(0, T)$ be a solution of (2.1) and $\varepsilon > 0$ be a given number. Denote by $C \geqq 1$ a Lipschitz constant of A which exists thanks to Proposition 1.7 (a). Let $M \subset [0, T]$ be such that w' is bounded on $[0, T] \setminus M$ and

$$2C(T+1)e^{CT} \int_M (\|w(s)\| + \|w'(s)\|)\, ds < \varepsilon. \tag{2.6}$$

Set

$$y'(t) := \begin{cases} 0 & \text{if } t \in M, \\ w'(t) & \text{otherwise,} \end{cases}$$

and

$$y(t) := \int_0^t y'(s)\, ds.$$

Then

$$\|y(t) - w(t)\| \leqq \int_M \|w'(s)\|\, ds \leqq \varepsilon/2 \tag{2.7}$$

and

$$p(t) := \operatorname{dist}(y'(t), A(y(t))) \leqq \begin{cases} C\|y(t)\| & \text{if } t \in M, \\ C\|w(t) - y(t)\| & \text{otherwise.} \end{cases}$$

Thus

$$\int_0^T p(t)\,dt \leqq C\left(\int_M \|w(t)\|\,dt + \int_0^T \|w(t)-y(t)\|\,dt\right) \leqq C\int_M (\|w(s)\| + T\|w'(s)\|)\,ds. \tag{2.8}$$

By a Filippov theorem (see Aubin and Cellina [1984, p. 120]) there exists a solution $x(\cdot)$ to (2.1) satisfying, by (2.6) and (2.8),

$$\text{(i)}\quad \|x(t)-y(t)\| \leqq e^{CT}\int_0^T p(t)\,dt < \varepsilon/2,$$

$$\text{(ii)}\quad \|x'(t)-y'(t)\| \leqq C\,e^{CT}\int_0^T p(t)\,dt + p(t) \quad \text{a.e.} \tag{2.9}$$

Since $p(\cdot)$ is a bounded function and $y \in W^{1,\infty}(0, T)$, the solution $x(\cdot)$ belongs to $W^{1,\infty}(0, T)$. Moreover by (2.7), (2.9), for all $t \in [0, T]$,

$$\|x(t)-w(t)\| \leqq \|x(t)-y(t)\| + \|y(t)-w(t)\| < \varepsilon.$$

Since ε is an arbitrary positive number the proof ensues. □

Proof of Theorem 2.1. (a) We denote by S the closed convex cone of solutions to the differential inclusion (2.1) in the Hilbert space

$$X := \{x \in W^{1,2}(0, T) \,|\, x(0) = 0\}. \tag{2.10}$$

Consider the continuous linear operator

$$\gamma_T : x(\cdot) \in X \to x(T) \in \mathbb{R}^n.$$

The transpose γ_T^* maps $\mathbb{R}^n$ into the dual X^* of X and for all $\eta \in R_T^+$

$$\forall x \in S, \qquad \langle \gamma_T^* \eta, x\rangle = \langle \eta, \gamma_T x\rangle \geqq 0. \tag{2.11}$$

By Lemma 2.2, S is dense in the $W^{1,1}(0, T)$ solutions to (2.1) in the metric of uniform convergence on $[0, T]$. This and (2.11) yield

$$R_T^+ = \{\eta : \gamma_T^* \eta \in S^+\}. \tag{2.12}$$

Let us set

$$\text{(i)}\quad Y := L^2(0, T; \mathbb{R}^n) \times L^2(0, T; \mathbb{R}^n),$$

$$\text{(ii)}\quad L := \{(x, y) \in Y : y(t) \in A(x(t)) \text{ a.e.}\}, \tag{2.13}$$

$$\text{(iii)}\quad D, \text{ the differential operator defined on } X \text{ by } Dx = x'.$$

Then $S = (1 \times D)^{-1}(L)$. The closed image Lemma 1.3 applied to the continuous linear operator $\phi = (1 \times D)$ states that

$$S^+ = (1 \times D)(L^+) \tag{2.14}$$

provided that the "surjectivity assumption"

$$\operatorname{Im}(1 \times D) - L = Y \tag{2.15}$$

is satisfied.

(b) It can be written

$$\forall (u, v) \in Y \text{ there exists } x \in X \text{ such that } x'(t) \in A(x(t) - u(t)) + v(t) \text{ a.e.} \tag{2.16}$$

Since the domain of A is the whole space, then A is Lipschitzian. The set-valued map $F(t, x) := A(x-u(t))+v(t)$ is then measurable in t, Lipschitzian with respect to x, has closed images and satisfies the following estimate

$$d(0, F(t, 0)) \leqq \|A\| \|u(t)\| + \|v(t)\|.$$

The function $t \to \|A\| \|u(t)\| + \|v(t)\|$ being in $L^1(0, T)$ we can apply a Filippov theorem [1967] (see Clarke [1983]) which states the existence of a solution $x(\cdot)$ to the differential inclusion $x'(t) \in F(t, x(t))$, $x(0) = 0$, satisfying

$$\|x'(t)\| \leqq \|A\| e^{\|A\|T} \int_0^T d(0, F(t, 0))\, dt + d(0, F(t, 0)).$$

Thus $x \in X$ and the surjectivity assumption (2.15) holds true.

(c) Therefore, by (2.12) and (2.14), we obtain the formula

$$R_T^+ = \{\eta : \gamma_T^* \eta \in (1 \times D)^*(L^+)\}. \tag{2.17}$$

Let $\eta \in Q_T$ and q be a solution to the adjoint inclusion (2.3). By Proposition 1.7(b), $q(\cdot) \in W^{1,\infty}(0, T)$ and for all $x \in S$

$$\langle \eta, x(T)\rangle = \langle (q', q), (x, x')\rangle_Y.$$

This is nonnegative by the definition of A^*. Thus $Q_T \subset R_T^+$.. To prove the opposite, let η belong to R_T^+. By (2.17), there exists $(p, q) \in L^+$ such that

$$\langle \eta, \gamma_T x\rangle = \langle p, x\rangle_{L^2} + \langle q, Dx\rangle_{L^2} \quad \forall x \in X. \tag{2.18}$$

By taking x so that $x(T) = 0$ we deduce that $p = Dq$ in the sense of distributions. Since p and q belong to L^2, we infer that q belongs to the Sobolev space $W^{1,2}(0, T)$. Thus $Dq = q'$. Integrating by parts in equation (2.18) and taking into account that $x(0) = 0$, we obtain

$$\langle \eta, \gamma_T x\rangle = \langle p - q', x\rangle_{L^2} + \langle q(T), x(T)\rangle = \langle q(T), x(T)\rangle.$$

The surjectivity of γ_T implies that $\eta = q(T)$. Thus $q(\cdot)$ is a solution to (2.3) and then, η belongs to Q_T. This achieves the proof. □

3. Invariant cones and viability domains. We devote this section to a thorough study of the viability domains for A^*, the transpose of a strict closed convex process. We then derive, thanks to the duality theorem, corresponding properties of the invariant cones.

We consider the Cauchy problem for the differential inclusion

$$\text{for almost all } t \in [0, T], \quad x'(t) \in A(x(t)), \quad x(0) = 0, \tag{3.1}$$

the reachable sets R_T defined by (2.2), the adjoint differential inclusion

$$\text{for almost all } t \in [0, T], \qquad -q'(t) \in A^*(q(t)). \tag{3.2}$$

We associate with any $\eta \in \operatorname{Dom} A^*$ the "solution set" $S_T(\eta)$ of solutions to the differential inclusion (3.2) satisfying $q(T) = \eta$ and we denote by Q_T the domain of the "solution map" S_T

$$Q_T := \{\eta \in \operatorname{Dom} A \mid S_T(\eta) \neq \emptyset\}. \tag{3.3}$$

We shall use the following technical lemma.

LEMMA 3.1. *Let A be a strict closed convex process. The following properties hold true*:

(a) *the graph of the restriction of S_T to any compact subset of* Dom A^* *is compact in* $\mathbb{R}^n \times \mathscr{C}(0, T; \mathbb{R}^n)$.

(b) *Any viability domain P for A^* is contained in Q_T.*

Proof. (a) Let $\mathscr{C}$ be a compact subset of Dom A^* and let us consider a sequence (η_n, q_n) where $\eta_n \in C$ and $q_n \in S_T(\eta_n)$. Then a subsequence (again denoted η_n) of η_n converges to some $\eta \in C$ because C is compact.

For almost all $t \in [0, T]$

$$\left|\frac{d}{dt}\|p_n(t)\|^2\right| = 2|\langle p_n(t), p_n'(t)\rangle| \leqq 2\|p_n(t)\|\,\|p_n'(t)\| \leqq 2\|A\|\,\|p_n(t)\|^2$$

(by formula (1.8), because $-p_n'(t) \in A^*(p_n(t))$).

Gronwall's lemma implies that

$$\|p_n(t)\| \leqq \|\eta_n\| \exp(\|A\|(t-T)). \tag{3.4}$$

This and formula (1.8) imply that for almost all $t \in [0, T]$,

$$\|p_n'(t)\| \leqq \|A\|\,\|\eta_n\| \exp(\|A\|(t-T)). \tag{3.5}$$

Thus, by the Banach-Alaoglu theorem, p_n' lies in a weakly compact subset of $L^\infty(0, T; \mathbb{R}^n)$ and by the Ascoli-Arzela theorem, p_n lies in a compact subset of $\mathscr{C}(0, T; \mathbb{R}^n)$. Therefore there exists a subsequence (again denoted) $p_n(\cdot)$ and an absolutely continuous function $p: [0, T] \to \mathbb{R}^n$ such that

$$\begin{array}{ll} \text{(i)} & p_n \text{ converges uniformly to } p \text{ on } [0, T], \\ \text{(ii)} & p_n' \text{ converges weakly } p' \text{ in } L^1(0, T; \mathbb{R}^n). \end{array} \tag{3.6}$$

The weak convergence of the pair (p_n, p_n') in $L^1(0, T; \mathbb{R}^n) \times L^1(0, T; \mathbb{R}^n)$ implies the strong convergence of convex combinations of elements of this sequence (Mazur's lemma). Since $(p_n(t), p_n'(t))$ belongs to Graph A^* for almost all $t \in [0, T]$ and since it is closed and convex, we infer that for almost all $t \in [0, T]$, $(p(t), p'(t)) \in$ Graph (A^*). Hence $p(\cdot)$ belongs to $S_T(\eta)$.

(b) Let P be a viability domain for A and $\eta \in P$. We shall show that there exists a solution $p \in S_T(\eta)$.

The viability theorem (see Haddad [1981]) implies that for all $t_0 \leqq T$ a solution p of the differential inclusion

$$-p'(t) \in A^*(p(t)), \quad p(t) \in P, \quad p(T) = \eta, \tag{3.7}$$

defined on a time interval $[t_0, T]$, can be extended to a solution of (3.7) defined on a larger time interval $[t_1, T]$, $t_1 < t_0$. Setting $\eta_n = \eta$ in (3.4) and (3.5), we obtain that

$$\begin{array}{ll} \text{(i)} & \|p(t)\| \leqq \|\eta\| \quad \text{for all } t \in [t_1, T], \\ \text{(ii)} & \|p'(t)\| \leqq \|A\|\,\|\eta\| \quad \text{for a.e. } t \in [t_1, T]. \end{array} \tag{3.8}$$

As in the case of ordinary differential equations, one can show that $p(\cdot)$ can be extended to a solution (again denoted $p(\cdot)$) defined on the time interval $[0, T]$. Thus $p(\cdot)$ belongs to $S_T(\eta)$ and thus, η belongs to A_T. □

We observe now that the sequence of the closed domains Q_T decreases

$$\text{if } T_1 \geqq T_2, \quad \text{then } Q_{T_1} \subset Q_{T_2}. \tag{3.9}$$

We introduce the intersection Q of these cones

$$Q := \bigcap_{T>0} Q_T. \tag{3.10}$$

Since the compact subsets $S^{n-1} \cap Q_T$ form a decreasing sequence, we observe that $Q \neq \{0\}$ if and only if all the cones Q_T are different from 0. We shall see that Q is the *largest viability domain*, thanks to the following theorem.

THEOREM 3.2. *Let A be a strict closed convex process. Then the closed convex cone Q is the largest closed convex cone which is a viability domain for A^*.*

Proof. Lemma 3.1 (b) implies that Q is a closed convex cone which contains any viability domain P. It remains to prove that Q is a viability domain, i.e., that

$$\forall q \in Q, \qquad A^*(q) \cap T_Q(q) \neq \emptyset. \tag{3.11}$$

Assume that $Q \neq \{0\}$.

Thanks to the necessary condition of the viability theorem (see Haddad [1981]), it is sufficient to prove that for some $T > 0$,

$$\forall \eta \in Q, \quad \exists p(\cdot) \in S_T(\eta) \quad \text{which is viable on } Q. \tag{3.12}$$

Since η belongs to Q_{nT} for all $n \geqq 2$, there exists a solution $p_n(\cdot) \in S_{nT}(\eta)$. By the very definition of Q_t, we know that $p(t) \in Q_t$ for all $t \leqq nT$.

Therefore, the translated function $\hat{p}_n(\cdot)$ defined on $[0, T]$ by

$$\hat{p}_n(t) := p_n(t + (n-1)T) \tag{3.13}$$

belongs to $S_T(\eta)$ and satisfies for all $t \in [0, T]$, $k \leqq n-1$,

$$\hat{p}_n(t) = p_n(t + (n-1)T) \in Q_{t+(n-1)T} \subset Q_{(n-1)T} \subset Q_{kT}. \tag{3.14}$$

By Lemma 3.1 (a), $S_T(\eta)$ is compact in $\mathscr{C}(0, T; \mathbb{R}^n)$. Thus there exists a subsequence of $\hat{p}_n(\cdot)$ converging to some $\hat{p}(\cdot) \in S_T(\eta)$ uniformly on $[0, T]$. By (3.14) for all $t \in [0, T]$, $k \geqq 1$, $\hat{p}(t) \subset Q_{kT}$. Therefore

$$\hat{p}(t) \subset \bigcap_{k \geqq 1} Q_{kT} = Q. \qquad \square$$

We now translate this result in terms of reachable sets R_T.

Since $0 \in A(0)$ the reachable cones $R(T)$ do form an increasing sequence. We define the reachable set of the inclusion (3.1) to be

$$R := \bigcup_{T>0} R(T). \tag{3.15}$$

It is a convex cone, which is equal to the whole space if and only if for some $T > 0$, $R(T) = \mathbb{R}^n$.

THEOREM 3.3. *Let A be a strict closed convex process. Then the closed convex cone $\bar{R}$ is the smallest closed convex cone invariant by A.*

Proof. Indeed Theorem 2.1 and the definition of $\bar{R}$ and Q imply that $\bar{R}^+ = Q$. By Theorem 3.2 and Proposition 1.12, $\bar{R}$ is the smallest closed convex cone containing $A(0) = (\text{Dom } A^*)^+$ which is invariant by A. $\square$

We consider now the largest subspace

$$Q \cap -Q \subset \text{Dom } A^* \cap -\text{Dom } A^* \tag{3.16}$$

of Q.

PROPOSITION 3.4. *Let A be a strict closed convex process. The subspace $Q \cap -Q$ is the largest subspace invariant by A^* and its orthogonal space $R - R$ is invariant by A in the sense that*

$$\forall x \in R - R, \qquad A(x) \subset R - R. \tag{3.17}$$

Proof. By Proposition 1.7 (c) the restriction of A^* to $Q \cap -Q$ is a linear (single-valued) operator. We have to check that $A^*(Q \cap -Q) \subset Q \cap -Q$. Let q belong to $Q \cap -Q$. Then by Theorem 3.2, since $\mathbb{R}q \subset Q \cap -Q$

$$A^*q \in T_Q(q) = \overline{(\mathbb{R}q + Q)} \subset Q + Q \cap -Q \subset Q.$$

Since $-q \in Q \cap -Q$, we also have

$$A^*q = -A^*(-q) \subset -Q.$$

Thus

$$A^*q \in Q \cap -Q.$$

Since $Q = \bar{R}^+$, the orthogonal space to $Q \cap -Q$ is the (closed) vector space spanned by R. Since we are in finite-dimensional space, we infer that

$$(Q \cap -Q)^\perp = R - R. \tag{3.18}$$

Proposition 1.12 implies that the vector space $R - R$ is invariant by A, because we have proved that $Q \cap -Q$ is a viability domain for A^*. □

We consider now the cones $A(0)$, $A^2(0) := A(A(0)), \cdots, A^k(0) = A(A^{k-1}(0))$, etc. Since 0 belongs to $A(0)$, these convex cones form an increasing sequence. We introduce the cone

$$N := \operatorname{cl}\left(\bigcup_{k \geq 1} A^k(0)\right) \tag{3.19}$$

and the vector subspace

$$M \text{ spanned by } N. \tag{3.20}$$

THEOREM 3.5. *Let A be a strict closed convex process. Then*

(a) $A(N) \subset N$,

(b) $\bar{R} \subset N \subset M \subset R - R$,

(c) $Q \cap -Q \subset \bigcap_{k \geq 1} A^k(0)^\perp \subset \bigcap_{k \geq 1} A^k(0)^+ \subset Q$.

Proof. (a) It is clear that $A(\bigcup_{k \geq 1} A^k(0)) \subset N$.

Let $x \in N$, $y \in A(x)$ and $x_n \in \bigcup_{j \geq 1} A^j(0)$ be a sequence converging to x. Since A is Lipschitzian, there exists a sequence $y_n \in A(x_n) \subset N$ converging to y, which belongs to N because it is closed.

(b) Since N is a closed invariant cone containing $A(0)$, Theorem 3.3 implies that N contains the reachable set $\bar{R}$. On the other hand, 0 belongs to $R - R$ and this vector space is invariant by A, thanks to Proposition 3.4. Therefore the cones $A^k(0) = A(A^{k-1}(0))$ are contained in $R - R$ and so does M.

(c) We deduce the other inclusions by polarity, noticing that $N^+ = \bigcap_{k \geq 1} A^k(0)^+$ and

$$M^\perp = \bigcap_{k \geq 1} A^k(0)^\perp.$$

Remark. When the reachable set R is a vector space, the subsets R, N, M and $R - R$ coincide. This happens when, for instance, A is symmetric (in the sense that $A(-x) = -A(x)$), i.e., when the graph of A is a vector subspace.

4. Eigenvectors and eigenvalues of convex processes. When $Q \cap -Q = \{0\}$ (or $R - R = \mathbb{R}^n$), there is no proper subspace invariant by A^* (or there is no proper subspace invariant by A). Moreover, when $Q \neq \{0\}$ (or $R \neq \mathbb{R}^n$), we can still prove the existence of an eigenvalue of A (see Definition 1.13 and Lemma 1.14), or eigenvectors of A^*.

Actually, eigenvectors η of A^*, nonzero solutions of the inclusion $\lambda\eta \in A^*(\eta)$, do belong to the largest viability domain Q because for all $T>0$ the function $p(t) := \eta \exp(\lambda(T-t))$ belongs to $S_T(\eta)$.

THEOREM 4.1. *Let A be a strict closed convex process. If the largest viability domain Q for A^* is different from $\{0\}$ and contains no line, then A^* has at least one eigenvector.*

By Lemma 1.14 and duality Theorem 2.1, the following dual version of this theorem holds true.

THEOREM 4.2. *Let A be a strict closed convex process. Assume that the reachable set R is different from $\mathbb{R}^n$ and spans the whole space. Then A has at least one eigenvalue.*

First we recall the following property

LEMMA 4.3. *Let Q be a closed convex cone of $\mathbb{R}^n$. The following properties are equivalent:*

(i) $Q \cap -Q = \{0\}$,

(ii) *Q is spanned by a compact convex subset which does not contain zero,*

(iii) *The interior of Q^+ is nonempty.*

If one of these properties holds true, then for all $x_0 \in \operatorname{Int} Q^+$, the compact convex subset

$$M := \{q \in Q : \langle q, x_0\rangle = 1\} \tag{4.1}$$

spans Q.

Proof. We provide the proof for the convenience of the reader.

Condition (i) means that zero is the extremal point of Q, which is equivalent to the assertion $0 \notin \operatorname{co}(Q \cap S^{n-1})$. Since the compact convex set $\operatorname{co}(Q \cap S^{n-1})$ spans the cone Q we proved the equivalence of (i) and (ii). Condition (iii) means that $Q^{++} = Q$ contains no line, which is precisely the statement (i).

If $x_0 \in \operatorname{Int} Q^+$ and $q,\ q_i \in M,\ i = 1, 2, \cdots$ are such that

$$\langle q_i, x_0\rangle = 1, \qquad \lim_{i\to\infty} q_i/\|q_i\| = q \in Q \cap S^{n-1}.$$

Then

$$0 < \langle q, x_0\rangle = \lim_{i\to\infty} \langle q_i, x_0\rangle/\|q_i\| = \lim_{i\to\infty} \|q_i\|^{-1}.$$

It implies that the norms $\|q_i\|$ are bounded and, therefore, M is bounded. Obviously it is also convex and closed. □

Proof of Theorem 4.1. Let $x_0 \in \operatorname{Int} Q^+$ and let M be defined by (4.1). Then for all $p \in M$

$$T_M(p) := \{v \in T_Q(x_0) : \langle v, x_0\rangle = 0\}. \tag{4.2}$$

We introduce the following projectors

$$\forall p \in M, \qquad \pi(p)q = q - \langle q, x_0\rangle p. \tag{4.3}$$

For all $p \in M$ and $q \in Q$, $\langle \pi(p)p, x_0\rangle = 0 = \langle \pi(p)q, x_0\rangle$. Hence the projector $\pi(p)$ maps the set $\mathbb{R}p + Q$ into $T_M(p)$. Since $T_Q(p) = \overline{\mathbb{R}p + Q}$ and $\pi(p)$ is a continuous linear operator, we obtain

$$\forall p \in M, \qquad \pi(p) \text{ maps } T_Q(p) \text{ into } T_M(p). \tag{4.4}$$

Consider the set-valued map $p \in M \to \pi(p)A^*(p)$. It is upper semicontinuous with nonempty compact convex images. By assumptions of Theorem 4.1, for all $p \in M \subset Q$, $A^*(p) \cap T_Q(p) \neq \emptyset$. Thus by (4.4)

$$\forall p \in M, \qquad \pi(p)A^*(p) \cap T_M(p) \neq \emptyset. \tag{4.5}$$

The assumptions of Aubin and Ekeland [1984, Thm. 6.4.11, p. 341] are satisfied. Therefore, for some $\bar{p} \in M$, $0 \in \pi(\bar{p})A^*(\bar{p})$. Hence there exists $\bar{q} \in A^*(\bar{p})$ such that $\langle \bar{q}, x_0 \rangle \bar{p} = \bar{q} \in A^*(\bar{p})$. In other words $\bar{p}$ is an eigenvector of A^* associated to the eigenvalue $\langle \bar{q}, x_0 \rangle$. □

5. Characterization of controllable convex processes. We shall deduce from the preceding results several characterizations of the controllability of differential inclusions

$$\text{for almost all } t \in [0, T], \quad x'(t) \in A(x(t)), \quad x(0) = 0 \tag{5.1}$$

or, equivalently, of the observability of the adjoint inclusion

$$\text{for almost all } t \in [0, T], \quad -q'(t) \in A^*(q(t)). \tag{5.2}$$

DEFINITION 5.1. We shall say that (5.1) is *controllable at time T* (respectively, *controllable*) if $R_T = \mathbb{R}^n$ (respectively, $R = R^n$). We shall say that the adjoint inclusion (5.2) is *observable at time T* (respectively, *observable*) if $Q_T = \{0\}$ (respectively, $Q = \{0\}$).

We also observe the following property.

LEMMA 5.2. *Let A be a strict closed convex process. The three following properties are equivalent.*

$$\begin{aligned} &\text{(a)} \quad \exists m \geqq 1 \text{ such that } A^m(0) - A^m(0) = \mathbb{R}^n, \\ &\text{(b)} \quad \exists m \geqq 1 \text{ such that } A^m(0)^{\perp} = \{0\}, \\ &\text{(c)} \quad \exists m \geqq 1 \text{ such that } \operatorname{Int} A^m(0) \neq \varnothing. \end{aligned} \tag{5.3}$$

It is convenient to introduce the

Rank condition 5.3. We say that a convex process A satisfies the rank condition if one of the equivalent properties (5.3) holds true.

LEMMA 5.4. *Consider the strict closed convex process $A(x) = Fx + L$, where $F \in \mathbb{R}^{n \times n}$ is a matrix and L is a vector subspace of $\mathbb{R}^n$. Then A satisfies the rank condition if and only if $A^n(0) - A^n(0) = \mathbb{R}^n$.*

Proof. The rank condition is satisfied if and only if for some $m \geqq 1$ the cone $L + AL + \cdots + A^{m-1}L$ spans the whole space. The Cayley-Hamilton theorem ends the proof. □

We begin by stating characteristic properties of observability of the adjoint system (5.2) and then use the duality results to infer the equivalent characteristic properties of system (5.1).

THEOREM 5.5. *Let A be a strict closed convex process. The following properties are equivalent*:

(a*) *the adjoint inclusion* (5.2) *is observable,*

(b*) *the adjoint inclusion* (5.2) *is observable at time $T > 0$ for some T,*

(c*) $\{0\}$ *is the largest closed convex cone which is a viability domain for A^*,*

(d*) *A^* has neither proper invariant subspace nor eigenvectors,*

(e*) *the rank condition is satisfied and A^* has no eigenvectors.*

Proof. (α) Since the intersections $Q_T \cap S^{n-1}$ of the cones Q_T and the unit sphere S^{n-1} form a decreasing sequence of compact subsets, we deduce that $Q \cap S^{n-1}$ is empty if and only if $Q_T \cap S^{n-1}$ is empty for some T, i.e., that $Q = \{0\}$ if and only if $Q_T = \{0\}$ for some $T > 0$. Thus (a*)$\Leftrightarrow$(b*).

(β) Property (c*) is equivalent to $Q = \{0\}$ by Theorem 3.2, i.e. (a*)$\Leftrightarrow$(c*).

(γ) When $Q = \{0\}$, then $Q \cap -Q = \{0\}$ (there is no proper invariant subspace) and there is no eigenvector (because an eigenvector is contained in Q).

When $Q \neq \{0\}$, then either $Q \cap -Q \neq \{0\}$ and, by Proposition 3.4, there is a proper invariant subspace or $Q \cap -Q = \{0\}$ and, by Theorem 4.1, there exists at least an eigenvector of A^*. This proves the equivalence of (d*) with $Q = \{0\}$, i.e., (a*)$\Leftrightarrow$(d*).

(δ) Since the sequence of cones $A^k(0)$ is increasing, the sequence of vector spaces $A^k(0)^\perp$ is decreasing, so that

$$\bigcap_{k \geq 1} A^k(0)^\perp = \{0\} \Leftrightarrow \exists m \geq 1 \quad \text{such that } A^m(0)^\perp = \{0\}$$
$$\Leftrightarrow \text{the rank condition is satisfied.}$$

Assume that $Q = \{0\}$. Then, by Theorem 3.5(c), and the above remark, the rank condition is satisfied and there is no eigenvector. Assume now that the rank condition is satisfied. Then $Q \cap -Q = \{0\}$ by Theorem 3.5(c). Then Theorem 4.1 implies that if A^* does not have an eigenvector, the cone Q is equal to $\{0\}$. Equivalence between (e*) and $Q = \{0\}$ ensues. □

THEOREM 5.6. *Let A be a strict closed convex process. The equivalent properties* (a*), (b*), (c*), (d*) *and* (e*) *of Theorem* 5.5 *are equivalent to the following properties*:

(a) *the differential inclusion* (5.1) *is controllable,*

(b) *the differential inclusion* (5.1) *is controllable at some time $T > 0$,*

(c) $\mathbb{R}^n$ *is the smallest closed convex cone invariant by A,*

(d) *A has neither proper invariant subspace nor eigenvalues,*

(e) *the rank condition is satisfied and A has no eigenvalues,*

(f) *for some $m \geq 1$, $A^m(0) = (-A)^m(0) = \mathbb{R}^n$.*

Proof. Statements (a) through (e) follow from the duality results (Proposition 1.12, Lemma 1.14 and Theorem 2.1) and Theorem 5.5. We shall show that (a) is also equivalent to (f).

Step 1. Consider the closed convex process $A_1(x) = A(-x)$. Then $A_1^* = -A^*$. We claim that (5.1) is controllable if and only if the inclusion

$$(5.1)' \qquad x' \in A_1(x), \qquad x(0) = 0$$

is controllable.

Indeed invariant subspaces and eigenvectors of A_1^* and A^* coincide and our claim follows from Theorem 5.5(d*).

Step 2. If (5.1) is controllable, then by Step 1 and Theorem 3.5(b)

$$\bigcup_{k \geq 1} A^k(0) = \bigcup_{k \geq 1} A_1^k(0) = \mathbb{R}^n.$$

Since $\{A^k(0)\}$ and $\{A_1^k(0)\}$ are increasing sequences of convex cones it implies that for some $m \geq 1$, $A^m(0) = A_1^m(0) = \mathbb{R}^n$. Moreover $A_1^m(0) = -(-A)^m(0)$. This implies (f).

Step 3. Assume that (f) holds true. If (5.1) is not controllable then there exist $\lambda \in \mathbb{R}$, $q \in A(0)^+$, $q \neq 0$ such that $\lambda q \in A^*(q)$. Then $(-\lambda)q \in A_1^*(q)$. Therefore,

$$\lambda^m q \in (A^*)^m(q) \quad \text{if } \lambda \geq 0,$$

$$(-\lambda)^m q \in (A_1^*)^m(q) \quad \text{if } \lambda \leq 0.$$

If $\lambda \geq 0$, then for all $y \in A^m(0)$, $0 = \langle \lambda^m q, 0 \rangle \leq \langle q, y \rangle$. If $\lambda \leq 0$, then for all $y \in A_1^m(0)$, $0 = \langle (-\lambda)^m q, 0 \rangle \leq \langle q, y \rangle$. In both cases we obtain a contradiction with (f). The proof is complete. □

So, the conjunction of Theorems 5.5 and 5.6 imply Theorem 0.4 stated in the Introduction.

Example 1. In the case when the set-valued map A is defined by $A(x) := Fx + L$, we derive known results due to Kalman when L is a vector space of controls and to

Brammer, and Saperstone and Yorke when L is an arbitrary set of controls containing 0.

Consider the linear control system in $\mathbb{R}^n$

$$x' = Fx + Gu, \quad u \in U, \quad x(0) = 0, \tag{5.4}$$

where $F \subset \mathbb{R}^{n\times n}$, $G \subset \mathbb{R}^{n\times m}$ are constant matrices and $U \subset \mathbb{R}^m$ is the given control set.

By Lemma 5.4 the rank condition 5.3 for the closed convex process $Ax = Fx + \operatorname{cl} G\{\lambda u: \lambda \geqq 0, u \in \operatorname{co} U\}$ is equivalent to

$$\sum_{i=0}^{n-1} F^i G(\operatorname{span} U) = \mathbb{R}^n. \tag{5.5}$$

This and Theorem 5.6(f) imply

THEOREM (Kalman). *If $U = \mathbb{R}^m$ then the control system* (5.4) *is controllable if and only if* $\operatorname{rank}[G, FG, \cdots, F^{n-1}G] = n$.

We shall study next the question of local controllability.

The control system (5.4) is said to be *locally controllable* around zero if zero is an interior point of the reachable set of (5.4).

To provide necessary and sufficient conditions for local controllability of (5.4) let us consider convex hull co U of U, and

$$N := \operatorname{cl}\{\lambda u: \lambda \geqq 0, u \in \operatorname{co} U\}$$

and the associated control system

$$x' \in Fx + \overline{GN}, \qquad x(0) = 0. \tag{5.6}$$

LEMMA 5.7. *If $0 \in \overline{\operatorname{co}}\, GU$ then the control system* (5.4) *is locally controllable around zero if and only if the system* (5.6) *is controllable.*

Proof. The reachable set of system (5.6) is a convex cone equal to

$$\left\{\int_0^t e^{F(t-s)}v(s)\,ds: t \geqq 0, v(s) \in \operatorname{cl} GN\right\} \tag{5.7}$$

and containing the reachable set of (5.4). Hence the local controllability of (5.4) implies the controllability of (5.6).

Because $0 \in \overline{\operatorname{co}}\, GU = \overline{G \operatorname{co} U}$, by a density argument, it is possible to verify that the cone given by (5.7) is equal to $\mathbb{R}^n$ if and only if

$$0 \in \operatorname{Int}\left\{\int_0^t e^{F(t-s)}Gu(s)\,ds: t \geqq 0, u(s) \in \operatorname{co} U\right\}. \tag{5.8}$$

Because the sets

$$\left\{\int_0^t e^{F(t-s)}Gu(s)\,ds: u(s) \in U\right\}$$

are convex and dense in

$$\left\{\int_0^t e^{F(t-s)}Gu(s)\,ds: t \geqq 0, u(s) \in \operatorname{co} U\right\}$$

(Lee and Markus [1967]) the inclusion (5.8) is equivalent to

$$0 \in \operatorname{Int}\left\{\int_0^t e^{F(t-s)}Gu(s)\,ds: t \geqq 0, u(s) \in U\right\}. \tag{5.9}$$

□

THEOREM 5.8. *Assume that* $0 \in \overline{\text{co}}\, GU$. *Then the system* (5.4) *is locally controllable around zero if and only if the rank condition* (5.5) *is satisfied and there is no eigenvector of* F^* *in* $(GU)^+$.

Proof. Observe that $GU^+ = (GN)^+$. By Lemma 5.7 it is enough to prove that the system (5.6) is controllable if and only if the rank condition 5.3 is satisfied and F^* has no eigenvector in $(GN)^+$. But this follows from Theorem 5.5(e*) and (5.5). □

In particular when $m = 1$ we obtain the result from Saperstone and Yorke [1971]. The above theorem is a generalization of Brammer's theorem [1972] (see also Jacobson [1977]). Theorem 5.6(f) and Example 1.15 imply

THEOREM 5.9. *Let* F *be an* $n \times n$ *matrix and let* L *be a closed convex subcone of* $\mathbb{R}^n$. *The control system*

$$x' = Fx + L, \qquad x(0) = 0$$

is controllable if and only if for some $m \geqq 1$

$$L + FL + \cdots + F^m L = L - FL + \cdots + (-1)^m F^m L = \mathbb{R}^n.$$

The last theorem together with Lemma 5.7 imply a result of Korobov [1980].

Example 2. Let $f = (f_1, \cdots, f_n)\colon \mathbb{R}^n \to \mathbb{R}^n$ be a single-valued map. Assume that for all $i = 1, \cdots, n$

f_i is lower semicontinuous, positively homogeneous convex function.

Observe that the assumption (1.17) holds true with $L = \mathbb{R}^n_+$. Consider the control system

$$x' = f(x) + u, \qquad u \in \mathbb{R}^n_+, \quad x(0) = 0. \tag{5.10}$$

THEOREM 5.10. *System* (5.10) *is controllable if and only if* f *has no proper invariant subspace containing* $\mathbb{R}^n_+$ *and if*

$$\sup_{\substack{q \in \mathbb{R}^n_+ \\ \lambda \in \mathbb{R}}} \inf_{x \in S^{n-1}} (\langle q, f(x)\rangle - \lambda \langle q, x\rangle) < 0$$

where S^{n-1} *denotes the unit sphere of* $\mathbb{R}^n$.

Proof. Set $A(x) := f(x) + \mathbb{R}^n_+$. By Theorem 0.4(g) the local controllability of (5.10) is equivalent to the two following properties:

(5.11) There is no proper subspace S satisfying $f(S) + \mathbb{R}^n_+ \subset S$.

(5.12) There is no $\lambda \in \mathbb{R}$ so that $\text{Im}\,(A - \lambda I) \neq \mathbb{R}^n$.

Since $f(0) = 0$ and $0 \in \mathbb{R}^n_+$ (5.11) is equivalent to

(5.11)′ there is no subspace S containing $\mathbb{R}^n_+$ such that $f(S) \subset S$.

By the separation theorem, (5.12) is equivalent to: for all $q \in \mathbb{R}^n_+$ and all $\lambda \in \mathbb{R}$ there exists $x \in \mathbb{R}^n$ such that $\langle f(x), q\rangle - \lambda \langle x, q\rangle < 0$. This ends the proof. □

Remark. Criterion (0.4)(k) can be used as well to derive a necessary and sufficient condition for the controllability of (5.10).

Let us consider the control system in $\mathbb{R}^2$

$$\begin{aligned} x' &= |x| + y + u, \qquad u \in \mathbb{R}_+, \\ y' &= |y| - x + v, \qquad v \in \mathbb{R}_+, \\ x(0) &= y(0) = 0. \end{aligned} \tag{5.12}′$$

We wish to know whether it is controllable. For this we consider the closed convex process $A: \mathbb{R}^2 \rightrightarrows \mathbb{R}^2$ defined by

$$A(x, y) = (|x| + y, |y| - x) + \mathbb{R}_+ \times \mathbb{R}_+ .$$

Then

$$A(0) = \mathbb{R}_+ \times \mathbb{R}_+, \qquad -A(0) = \mathbb{R}_- \times \mathbb{R}_-,$$
$$A^2(0) = \mathbb{R}_+ \times \mathbb{R}, \qquad (-A)^2(0) = \mathbb{R} \times \mathbb{R}_-,$$
$$A^3(0) = \mathbb{R} \times \mathbb{R}, \qquad (-A)^3(0) = \mathbb{R} \times \mathbb{R}.$$

By Theorem 0.4(k) the system (5.12) is controllable.

Acknowledgment. The authors acknowledge the opportunity offered by the International Institute of Applied Systems Analysis, Laxenburg, Austria, to work together.

REFERENCES

J. P. AUBIN AND A. CELLINA [1984], *Differential Inclusions*, Springer-Verlag, Berlin.

J. P. AUBIN AND I. EKELAND [1984], *Applied Nonlinear Analysis*, Wiley-Interscience, New York.

R. J. AUMANN [1985], *Integrals of set-valued functions*, J. Math. Anal. Appl., 12, pp. 1-12.

R. F. BRAMMER [1972], *Controllability in linear autonomous systems with positive controllers*, this Journal, 10, pp. 339-353.

F. H. CLARKE [1983], *Optimization and Nonsmooth Analysis*, Wiley-Interscience, New York.

A. F. FILIPPOV [1967], *Classical solutions of differential equations with multivalued right-hand side.* English translation: this Journal, 5, pp. 609-621.

H. FRANKOWSKA [1984], *Contrôlabilité locale et propriétés des semigroupes de correspondances*, CRAS, 299, pp. 165-168.

——— [1985], *Local controllability and infinitesimal generators of semi-groups of set-valued maps*, this Journal, 24 (1986), to appear.

G. HADDAD [1981], *Monotone trajectories of differential inclusions and functional differential inclusions with memory*, Israel J. Math., 39, pp. 83-100.

D. H. JACOBSON [1977], *Extensions of Linear-Quadratic Control, Optimization and Matrix Theory*, Academic Press, New York.

V. I. KOROBOV [1980], *A geometric criterion of local controllability of dynamical systems in the presence of constraints on the control*, Differential Equations, 15, pp. 1136-1142.

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

C. OLECH [1976], *Existence theory in optimal control*, in Control Theory and Topics in Functional Analysis, Vol. 1, pp. 291-328, Intern. At. Energy Agency, Vienna, 1976.

S. ROBINSON [1972], *Normed convex processes*, Trans. Amer. Math. Soc., 177, pp. 127-140.

R. T. ROCKAFELLAR [1967], *Monotone processes of convex and concave type*, Mem. Amer. Math. Soc., 77.

——— [1970], *Convex Analysis*, Princeton Univ. Press, Princeton, NJ.

——— [1974], *Convex algebra and duality in dynamic models of production*, in Mathematical Models in Economics, Łoś, ed., North-Holland, Amsterdam.

S. M. SAPERSTONE AND J. A. YORKE [1971], *Controllability of linear oscillatory systems using positive controls*, this Journal, 9, pp. 253-262.

C. URSESCU [1975], *Multifunctions with closed convex graph*, Czech. Math. J., 25, pp. 438-441.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 6, November 1986

008

SPECTRAL ASSIGNABILITY OF SYSTEMS WITH SCALAR CONTROL AND APPLICATION TO A DEGENERATE HYPERBOLIC SYSTEM*

LOP FAT HO†

Abstract. Some distributed parameter systems with scalar boundary control can be represented as systems in Hilbert spaces for which the input functional may not be continuous, but are admissible in some sense. We prove a spectral assignability result for such systems. The conditions we need are that the system be approximately controllable and that feedback relations of a certain type be continuous. We show that these conditions are satisfied by systems that are exactly controllable. We then apply the general results to a degenerate hyperbolic system. Having shown that it is exactly controllable, we obtain a spectral assignability result. Finally, we consider systems that may have multiple eigenvalues.

Key words. spectral assignability, controllability, linear feedback, Carleson measure

AMS(MOS) subject classification. 93B55

1. Introduction. It is a well-known fact that if a finite dimensional autonomous control system is completely controllable, then we can arbitrarily prescribe the eigenvalues of the closed loop system by means of a suitable linear feedback. The main purpose of this paper is to generalize this result, under some restriction, to infinite dimensional systems with scalar controls. Sun has obtained a similar result of this kind in [12]. The main difference between the work of this paper and that of Sun is that our work also includes systems with boundary controls. The problem of spectral assignability of stabilizability for various distributed parameter systems has been considered in [3], [11], [12], [13].

We will consider the system

$$\frac{d}{dt}x(t) = Ax(t) + u(t)b, \qquad x(t) \in X, \quad t \geqq 0 \tag{1.1}$$

where X is a Hilbert space, A is the generator of a strongly continuous semigroup, $u(\cdot)$ is a scalar value function and b is what we call an admissible input element, which will be defined in § 2. (See Definition 2.1 and also [9].) An admissible input element may not be in X. (In [12], b is required to be in X.) This less restrictive requirement on b allows us to include also systems with boundary controls in our consideration.

The problem of spectral assignability for the system (1.1) can be stated as follows.

For a given complex sequence $\{\mu_n\}_{n=-\infty}^{\infty}$ find a feedback relation $u(t) = \langle x(t), h\rangle$ such that the set of eigenvalues of the "closed loop system"

$$\frac{d}{dt}x(t) = Ax(t) + \langle x(t), h\rangle b$$

is exactly $\{\mu_n\}_{n=-\infty}^{\infty}$.

Most known results concerning this problem for the case of scalar controls say that the problem can be solved (i.e. we can find the desired h) if

$$\sum_{n=-\infty}^{\infty} \frac{|\lambda_n - \mu_n|^2}{|b_n|^2} < \infty$$

where $\{b_n\}$ is a sequence of complex numbers related to b (see [11], [12]).

* Received by the editors December 18, 1984, and in revised form June 10, 1985.

† Department of Mathematics, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong.

Also, most of the systems that have been considered possess, or are assumed to possess, the properties

1. that it is approximately controllable in some time $T>0$, and

2. that the numbers $\beta_n=\sum_{m\neq n} 1/[\lambda_n-\lambda_m|^2$, $-\infty<n<\infty$, are finite and the set $\{\beta_n\}$ is uniformly bounded.

The second property implies that the points λ_n cannot be too closed to each other asymptotically.

To prove the main spectral assignability result in this paper, we do not need property 2 as an assumption. Thus $\{\beta_n\}$ may be unbounded, but we will find out that the magnitudes of β_n will affect the class of sequences $\{\mu_n\}$ that can be assigned as eigenvalues of the closed loop system. More precisely, condition 2 will be replaced by

$$\sum_{h=-\infty}^{\infty} \frac{(1+\beta_n)|\mu_n-\lambda_n|^2}{|b_n|^2}<\infty.$$

Another assumption we need is that feedback relations of a certain type are continuous (see (4.3)). We will see that this assumption is valid if the system is exactly controllable in some time T. So we have a result of the type "exact controllability implies spectral assignability."

We now give an outline of the rest of this paper.

In § 2, we define the notion of an admissible input element and give meaning to the solution of the closed loop system. The definition we give here for admissible input element is slightly more restrictive than that given in [9].

In § 3, we study the spectrum of the closed loop system and show that it consists of zeros of a function analytic outside the spectrum of A, with possible points from the spectrum of A also. Our work here is somewhat complicated by the fact that we do not have a good expression for the generator of the closed loop system. We expect the generator to be $A_h x=Ax+\langle x, h\rangle b$, but if $b\notin X$, then this does not make sense; for example, when $\langle x, h\rangle\neq 0$. However, we can avoid this difficulty by defining A_h to be the adjoint of $Lx=A^*x+\langle x, b\rangle h$.

In § 4, we give an exact description of the spectrum of the closed loop system under the assumption that the spectrum of A consists of only simple eigenvalues. We establish the spectral assignability result, using a method which is a generalization of that used in [12], assuming that a certain class of feedback relations is continuous. We show that the continuity of such relations is implied by exact controllability, using a result on Carleson measure.

In § 5, we prove a controllability result for a degenerate hyperbolic system and apply the general theory we have developed to yield a result on spectral assignability.

Finally, in § 6, we show how we can modify the results we have developed to include the cases where there is a finite number of eigenvalues of A with multiplicities greater than one.

We will use $\sigma(\cdot)$ to denote the spectrum of the operator $\cdot$. Also we will use $\langle\cdot,\cdot\rangle$ to denote both inner product and duality pairing. $\langle x, y\rangle$ is linear in x and conjugate linear in y. We always identify X' so that $D(A^*)\subseteq X\subseteq D(A^*)'$ and the duality pairing on $D(A^*)'\times D(A^*)$ is an extension of the inner product on X. Also if $b\in D(A^*)'$ and $x\in D(A^*)$ then $\langle b, x\rangle=\overline{\langle x, b\rangle}$.

2. A scalar control system. We study the control system

$$\begin{aligned}&\frac{d}{dt}x(t)=Ax(t)+u(t)b, \qquad x(t)\in X,\\ &x(0)=x_0\end{aligned} \tag{2.1}$$

where X is a Hilbert space, A is the infinitesimal generator of a strongly continuous semigroup $S(t)$ of bounded operators on X and b is an admissible input element in the sense defined below.

DEFINITION 2.1. We suppose that $D(A^*)$ is considered as a Hilbert space equipped with the graph norm $\|x\|_{D(A^*)} \triangleq (\|x\|^2 + \|A^*x\|^2)^{1/2}$. A continuous linear functional b on $D(A^*)$ is said to be an *admissible input element* for the system (2.1) if for every $T > 0$, there exists a positive constant K_T, uniformly bounded for T sufficiently small, such that

$$\int_0^T |\langle b, S^*(t)y\rangle|^2 \, dt \leqq K_T^2 \|y\|^2 \quad \text{for all } y \in D(A^*). \tag{2.2}$$

Sufficient conditions for b to be an admissible input element can be found in [8].

If b is an admissible input element, then (cf. reference just cited) there exists, for each $t > 0$, a bounded linear operator $B(t)$ from $L^2[0, t]$ into X such that

$$\langle B(t)u(\cdot), y\rangle = \int_0^t \langle b, S^*(t-s)y\rangle u(s) \, ds \quad \text{for all } y \in D(A^*). \tag{2.3}$$

The solution of (2.1) is, by definition, equal to

$$x(t) = S(t)x_0 + B(t)u(\cdot). \tag{2.4}$$

We also remark that from (2.2) and (2.3), we have

$$\|B(t)\| \leqq K_T. \tag{2.5}$$

LEMMA 2.1.

$$B(t+s)u(\cdot) = S(t)B(s)u(\cdot) + B(t)u(s+\cdot). \tag{2.6}$$

Proof. Let $y \in D(A^*)$, then

$$\begin{aligned}
\langle B(t+s)u(\cdot), y\rangle &= \int_0^{t+s} \langle b, S^*(t+s-\tau)y\rangle u(\tau) \, d\tau \\
&= \int_0^s \langle b, S^*(s-\tau)S^*(t)y\rangle u(\tau) \, d\tau \\
&\quad + \int_s^{t+s} \langle b, S^*(t+s-\tau)y\rangle u(\tau) \, d\tau \\
&= \langle B(s)u(\cdot), S^*(t)y\rangle + \int_0^t \langle b, S^*(t-\tau)y\rangle u(s+\tau) \, d\tau \\
&= \langle S(t)B(s)u(\cdot), y\rangle + \langle B(t)u(s+\cdot), y\rangle.
\end{aligned}$$

Since $D(A^*)$ is dense in X, it follows that (2.6) holds. This finishes the proof of Lemma 2.1.

Remark. It follows from Lemma 2.1 that if $x(\cdot)$ is the solution of (2.1) with initial condition $x(0) = x_0$, then $x_t(\cdot) = x(t+\cdot)$ is the solution of (2.1) with initial condition $x_t(0) = x(t)$ and control $u(t+\cdot)$.

LEMMA 2.2. *Let h be a continous linear functional on X; then there exists a continuous function $x(\cdot)$ from $[0, \infty)$ into X such that*

$$x(t) = S(t)x_0 + B(t)\langle x(\cdot), h\rangle \quad \textit{for all } t > 0. \tag{2.7}$$

Proof. Fix $t_1>0$. Let M be the space of continuous functions from $[0, t_1]$ into X equipped with the supremum norm

$$\|x(\cdot)\|_\infty = \sup_{t\in[0,t_1]} \|x(t)\|, \quad x(\cdot)\in M.$$

We define a mapping F from M into itself by the equation $(Fx(\cdot))(t) = S(t)x_0 + B(t)\langle x(\cdot), h\rangle$, $0\leqq t\leqq t_1$. We have

$$\begin{aligned}\|Fx_1(\cdot) - Fx_2(\cdot)\|_\infty &= \sup_{t\in[0,t_1]} \|B(t)\langle x_1(\cdot) - x_2(\cdot), h\rangle\| \\ &\leqq \sup_{t\in[0,t_1]} \|B(t)\|\,\|h\| t_1^{1/2} \|x_1 - x_2\|_\infty.\end{aligned}$$

Here $\sup_{t\in[0,t_1]} \|B(t)\|$ is finite because of (2.5) and the boundedness of K_T. Hence if t_1 is chosen such that

$$\sup_{t\in[0,t_1]} \|B(t)\|\,\|h\| t_1^{1/2} < 1, \tag{2.8}$$

then F is a contraction mapping. Hence F has a unique fixed point $x_1(\cdot)$ in M. Such an x_1 then satisfies (2.7) in $[0, t_1]$. The condition (2.8) is independent of x_0, so we can also find $x_2(\cdot)\in M$ such that $x_2(t) = S(t)x_1(t_1) + B(t)\langle x_2(\cdot), h\rangle$, $0\leqq t\leqq t_1$. By virtue of the remark after Lemma 2.1, the function

$$x(s) = \begin{cases} x_1(s), & 0\leqq s\leqq t_1, \\ x_2(s-t_1), & t_1 < s\leqq 2t_1 \end{cases}$$

satisfies (2.7) for $0\leqq t\leqq 2t_1$. In a similar way, we can extend the function x to $[0, 3t_1]$, $[0, 4t_1]$, etc., on which (2.7) is satisfied. This finishes the proof of the lemma.

The unique solution of (2.7) will be denoted by $S_h(t)x_0$. For each t, $S_h(t)$ is then defined, in this way, as a linear operator from X into itself. And from the remark after Lemma 2.1, we have the semigroup property $S_h(\tau+t) = S_h(\tau)S_h(t)$. Furthermore, Lemma 2.2 implies that for any $x_0\in X$, the function mapping t to $S_h(t)x_0$ is continuous on $[0, \infty)$. Hence we have proved

LEMMA 2.3. *Let the solution in* (2.7) *be denoted by* $S_h(t)x_0$, *then* $S_h(t)$ *is a strongly continuous semigroup on* X.

We will define the solution of the system (2.1) with feedback relation $U(t) = \langle x(t), h\rangle$ to be $S_h(t)x_0$. The infinitesimal generator of S_h will be denoted by A_h.

3. Spectrum of the closed loop system. In this section, we will study the spectrum of the generator A_h of the semigroup S_h (see Lemma 2.3) associated with the closed loop system. Thinking of A_h as a perturbation of A, we will see that such a perturbation produces some new eigenvalues which are zeros of a certain function outside $\sigma(A)$. Part of $\sigma(A_h)$ may still be in $\sigma(A)$. To have a more precise characterization of $\sigma(A)$, we must make some more assumptions on A. This we will do in the next section.

We recall that if b is an admissible input element, then $b\in D(A^*)'$. Now if λ is not in the spectrum of A, then $R(\bar\lambda, A^*)$ is a bounded linear operator from X onto $D(A^*)$. Hence the expression $\langle b, R(\bar\lambda, A^*)x\rangle$ makes sense.

LEMMA 3.1. *The function* $\Delta_h(\lambda) = \langle b, R(\bar\lambda, A^*)h\rangle - 1$ *is defined and analytic on the resolvent set of* A.

Proof. Let λ be in the resolvent set of A. We have $R(\bar\mu, A^*) = R(\bar\lambda, A^*)\sum_{n=0}^\infty (\bar\lambda - \bar\mu)^n R(\bar\lambda, A^*)^n$, the sum being convergent in $L(X, D(A^*))$ equipped with uniform norm, whenever $|\bar\lambda - \bar\mu|\,\|R(\bar\lambda, A^*)\|_{L(X,D(A^*))} < 1$. It follows that for such μ, $\Delta_h(\mu) = \sum_{n=0}^\infty (\lambda - \mu)^n \langle b, R(\bar\lambda, A^*)^{n+1}h\rangle - 1$. This shows that Δ_h is analytic at λ and hence finishes the proof of the lemma.

To find the spectrum of A_h, we will (in Theorem 3.3) show that A_h is the adjoint of the operator L defined by

$$Ly = A^*y + \langle y, b\rangle h, \qquad D(L) = D(A^*). \tag{3.1}$$

We remark that L is closed because the mapping $Ty = \langle y, b\rangle h$ is A^* bounded with A^*-bound equal to zero. (T has finite rank). (See [10, p. 190].)

We first find a set that contains $\sigma(L)$. Let

$$\Lambda = \text{set of zeros of } \Delta_h. \tag{3.2}$$

LEMMA 3.2. *Suppose $\lambda \notin \sigma(A) \cup \Lambda$. Then given any x in X, the equation $(\bar{\lambda} I - L)y = x$ has the unique solution*

$$y = -(\langle R(\bar{\lambda}, A^*)x, b\rangle) R(\bar{\lambda}, A^*)h/(\bar{\Delta}_h(\lambda)) + R(\bar{\lambda}, A)x$$

in $D(A^*)$. If $\lambda \in \Lambda$ then $(\bar{\lambda} I - L)R(\bar{\lambda}, A^*)h = 0$.

Proof. That the given y indeed satisfies $(\bar{\lambda} I - L)y = 0$ can be proved by direct computation. To prove uniqueness, we note that if $(\bar{\lambda} I - L)y = 0$, then $y = R(\bar{\lambda}, A^*)\langle y, b\rangle h$. Hence $\langle y, b\rangle = \langle y, b\rangle \langle R(\bar{\lambda}, A^*)h, b\rangle$. Because $\Delta_h(\lambda) = \langle R(\bar{\lambda}, A^*)h, b\rangle - 1$ is assumed to be nonzero, it follows that $\langle y, b\rangle = 0$. Hence $y = 0$. Finally, the equality $(\bar{\lambda} I - L)R(\bar{\lambda}, A^*)h = 0$ can be proved by direct computation. This finishes the proof of the lemma.

THEOREM 3.3. *Let L be as in* (3.1) *and A_h be the generator of the closed loop system* (*cf. end of* § 2). *Then $A_h = L^*$. Consequently, $\sigma(A_h) \subseteq \sigma(A) \cup \Lambda$.*

Proof. Let $x \in D(A_h)$. Then for $y \in D(A^*) = D(L)$, we have

$$\begin{aligned}
\langle A_h x, y\rangle &= \lim_{t\to 0} \frac{1}{t}\langle S_h(t)x - x, y\rangle \\
&= \lim_{t\to 0}\left(\frac{1}{t}\langle S(t)x - x, y\rangle + \frac{1}{t}\int_0^t \langle b, S^*(t-s)y\rangle\langle S_h(s)x, h\rangle\, ds\right) \\
&= \langle x, A^*y\rangle + \langle b, y\rangle\langle x, h\rangle \\
&= \langle x, Ly\rangle.
\end{aligned}$$

Hence we have made use of the fact that $\langle b, S^*(t-s)y\rangle\langle h, S_h(s)x\rangle$ is continuous in s.

Hence $x \in D(L^*)$ and $L^*x = A_h x$. So we have proved that L^* is an extension of A_h. It remains to show that $D(L^*) \subseteq D(A_h)$.

We know that L is closed. So Lemma 3.2 implies that $\sigma(L) \subseteq \{\lambda : \bar{\lambda} \in \sigma(A) \cup \Lambda\}$. It follows that $\sigma(L^*) \subseteq \sigma(A) \cup \Lambda$. Since both A and A_h are generators of strongly continuous semigroups, there exists a real number ω such that every number with real part greater than ω lies in the resolvent set of both A and A_h. Since Λ consists only of discrete points, we can certainly find a $\lambda \notin \Lambda$ with real part greater than ω. Such a λ does not belong to $\sigma(L^*) \cup \sigma(A_h)$. Now let $x \in D(A^*)$ and let $x_1 = R(\lambda, A_h)(\lambda I - L^*)x$. Then $x_1 \in D(A_h)$. Furthermore, $(\lambda I - L^*)x_1 = (\lambda I - A_h)x_1 = (\lambda I - L^*)x$. Since $\lambda \notin \sigma(L^*)$, it follows that $x = x_1$, showing that $x \in D(A_h)$. Hence indeed $L = A_h$. Finally, we have $\sigma(A_h) = \sigma(L^*) \subseteq \sigma(A) \cup \Lambda$. This completes the proof of the Theorem.

4. Spectral assignability. In this section, we make the following assumption on A.

Assumption A. The spectrum of A consists of simple eigenvalues $\{\lambda_n\}_{n=-\infty}^{\infty}$.

We will denote by $\{\varphi_n\}_{n=-\infty}^{\infty}$ the corresponding eigenvectors of A. We also let $\{\psi_n\}_{n=-\infty}^{\infty}$ be eigenvectors of A^* such that $\langle \psi_m, \varphi_n\rangle = \delta_{m,n}$.

PROPOSITION 4.1. *Let A satisfy assumption A. Then*

$$\sigma(A_h) = \Lambda \cup \{\lambda_n : \langle h, \varphi_n\rangle = 0 \text{ or } \langle \psi_n, b\rangle = 0\}.$$

Proof. It follows from Theorem 3.4 and the last statement of Lemma 3.2 that $\lambda \in \Lambda$ are eigenvalues of A_h.

Suppose $\langle h, \varphi_n \rangle = 0$. Let $y \in D(A^*)$. By Theorem 3.4 and (3.1), we have

$$\begin{aligned}\langle \lambda_n \varphi_n - A_h \varphi_n, y \rangle &= \langle \lambda_n \varphi_n, y \rangle - \langle \varphi_n, A^* y + \langle b, y \rangle h \rangle \\ &= \langle \lambda_n \varphi_n, y \rangle - \langle A \varphi_n, y \rangle \\ &= 0.\end{aligned}$$

Since $D(A^*)$ is dense in X, it follows that $\lambda_n \varphi_n - A_h \varphi_n = 0$. Hence $\lambda_n \in \sigma(A_h)$.

Now suppose $\langle \psi_n, b \rangle = 0$. Then $\bar{\lambda}_n \psi_n - L\psi_n = \bar{\lambda}_n \psi_n - A^* \psi_n = 0$. Hence $\bar{\lambda}_n$ is an eigenvalue of L and so λ_n is an eigenvalue of $L^* = A_h$.

Finally, suppose both $\langle h, \psi_n \rangle$ and $\langle \psi_n, b \rangle$ are nonzero. Since $\bar{\lambda}_h$ is an isolated eigenvalue of A^*, X can be decomposed into the direct sum of two subspaces $\langle \psi_n \rangle$ and M where $\langle \psi_n \rangle$ is the one-dimensional subspace generated by ψ_n and M is a closed subspace invariant under A^* such that the restriction of $\bar{\lambda}_n I - A^*$ to M has a bounded inverse [10, p.178]. We denote this inverse by $R_1(\bar{\lambda}_n, A^*)$.

We take any $y \in X$ and let $\alpha = \langle y, \varphi_n \rangle / \langle h, \varphi_n \rangle$ and $y_1 = y - \alpha h$. We claim that $y_1 \in M$. Indeed, because of the decomposition $X = \langle \varphi_n \rangle \oplus M$ and the boundedness of $R_1(\bar{\lambda}_n, A^*)$, we can write $y_1 = a\psi_n + (\bar{\lambda}_n I - A^*) y_2$, $y_2 \in M_2$. Since $\langle \varphi_n, y_1 \rangle = \langle \varphi_n, y \rangle - \bar{\alpha} \langle \varphi_n, h \rangle = 0$ and $\langle \varphi_n, (\bar{\lambda}_n I - A^*) y_2 \rangle = \langle (\lambda_n I - A) \varphi_n, y_2 \rangle = 0$, it follows that $\langle \varphi_n, a\psi_n \rangle = 0$. Hence $a = 0$ and so $y_1 \in M$.

Let $x_1 = R_1(\bar{\lambda}_n, A^*) y_1$ and let $t = -(\alpha + \langle x_1, b \rangle) / \langle \psi_n, b \rangle$. It is easy to verify that $x = t\psi_n + x_1$ satisfies $(\bar{\lambda}_n I - L)x = y$. Hence $\bar{\lambda}_n I - L$ is onto.

Now suppose $x = a\psi_n + x_1$, $x_1 \in M$ satisfies $(\bar{\lambda}_n I - L)x = 0$. Then this implies that

$$(\bar{\lambda}_n I - A^*) x_1 - \langle a\psi_n + x_1, b \rangle h = 0. \tag{4.1}$$

Since $\bar{\lambda}_n I - A^*$ is one-to-one on M, it follows that $\langle a\psi_n + x_1, b \rangle \neq 0$ unless $x_1 = 0$. On the other hand, taking the inner product of (4.1) with φ_n, we obtain $-\langle a\psi_n + x_1, b \rangle \langle \varphi_n, h \rangle = 0$. This is impossible. So we must have $x_1 = 0$. (4.1) then implies that $a = 0$. Hence $x = 0$. This proves that $\bar{\lambda}_n I - L$ is one-to-one. We have already shown that $\bar{\lambda}_n I - L$ is onto. Hence $\bar{\lambda}_n \notin \sigma(L)$ and thus $\lambda_n \notin \sigma(L^*) = \sigma(A_h)$. This finishes the proof of the proposition.

We now make another assumption.

Assumption B. $b_n \triangleq \langle b, \psi_n \rangle \neq 0$ for all n.

PROPOSITION 4.2. *If the system* (2.1) *is approximately controllable in some time* T (*i.e., range of* $B(T)$ *is dense in* X), *then assumption B is satisfied.*

Proof.

$$\begin{aligned}\langle B(T)u(\cdot), \psi_n \rangle &\int_0^T \langle b, S^*(T-s)\psi_n \rangle u(s)\, ds \\ &= \int_0^T \langle b, e^{\bar{\lambda}_n (T-s)} \psi_n \rangle u(s)\, ds \\ &= b_n \int_0^T e^{\lambda_n (T-s)} u(s)\, ds.\end{aligned} \tag{4.2}$$

It follows that if $b_n = 0$, then $\langle x, \psi_n \rangle = 0$ for all x in the range of $B(T)$. So the range of $B(T)$ cannot be dense.

We will now study the effect of a feedback of the form

$$u(t) = \langle x(t), h \rangle = \sum_{n=-\infty}^{\infty} \langle x(t), \psi_n \rangle h_n, \quad \text{where} \sum_{n=-\infty}^{\infty} |h_n|^2 < \infty. \tag{4.3}$$

We will find sufficient conditions on the operator A so that the above h indeed represents a continuous linear functional on X. One such sufficient condition is that $\{\psi_n\}_{n=-\infty}^{\infty}$ forms a Riesz basis of X (see, for example, [7, p. 309] for the definition of Riesz basis). However, this condition is sometimes rather difficult to verify. We are going to give another sufficient condition which is based on the exact controllability of the system (2.1) and a certain property about the distribution of the eigenvalues of A related to the Carleson measure.

Let μ be a nonnegative measure defined on the Borel subsets of $\{z: \operatorname{Re} z > \alpha\}$. Then μ is a *Carleson measure* if for every τ and every $\nu > 0$, $\mu(\{z: \tau - \nu \leqq \operatorname{Im} z \leqq \tau + \nu, \alpha < \operatorname{Re} z < \alpha + \nu\}) \leqq C\nu$ for some positive constant C depending only on μ (not on ν).

If $\{a_n\}$ is a sequence of complex numbers with $\operatorname{Re} a_n > \alpha$, we can define a measure on Borel subsets of $\{z: \operatorname{Re} z > \alpha\}$ by

$$\mu(E) = \text{number of elements of } E \cap \{a_n: -\infty < n < \infty\}. \tag{4.4}$$

We will simply refer to μ as *the measure induced by* $\{a_n\}$.

Remark. If the system (2.1) is exactly controllable in some time $T > 0$ (i.e. $B(T)$ is onto), then in particular each $b_n \neq 0$. For convenience, we normalize $\{\phi_n\}$ and $\{\psi_n\}$ so that $\langle \psi_n, b\rangle = 1$. (Note that we always have $\langle \phi_n, \psi_n\rangle = 1$ so this will automatically give a corresponding normalization for ϕ_n.)

PROPOSITION 4.3. *Suppose the system* (2.1) *is exactly controllable in some time* $T > 0$ (*i.e., $B(T)$ is onto*) *and* $\{-\lambda\}_{n=-\infty}^{\infty}$ *induces a Carleson measure on* $\{z: \operatorname{Re} z > \alpha\}$ *where* α *is a real number less than* $-\operatorname{Re} \lambda_n$ *for all* n. *Then there exists a constant* $K > 0$ *so that*

$$\sum_{n=-\infty}^{\infty} |\langle x, \psi_n\rangle|^2 \leqq K\|x\|^2.$$

(Note that the above relation in fact depends on b because we require $\langle b, \psi_n\rangle = 1$. This is to be expected because exact controllability depends on b.)

To prove Proposition 4.3, we need to use the Carleson Theorem as applied to the space

$$H_\alpha^2 \equiv H^2\{z: \operatorname{Re} z > \alpha\}, \quad \alpha \text{ real.}$$

One way to define H_α^2 is to define it as the space of complex functions which are Laplace transforms of functions f on $[0, \infty)$ such that $e^{-\alpha t} f(t)$ is square integrable on $[0, \infty)$ and we have

$$\|\mathscr{L}f\|_{H_\alpha^2} = \int_0^\infty |e^{-\alpha t} f(t)|^2 \, dt.$$

H_α^2 is a Hilbert space. We have the following Carleson's theorem ([2], [5] or [9]).

THEOREM (Carleson). *If* μ *is a Carleson measure on* $\{z: \operatorname{Re} z > \alpha\}$ *then for* $\phi \in H_\alpha^2$, *we have*

$$\int_{\{z: \operatorname{Re} z > \alpha\}} |\phi(z)|^2 \, d\mu(z) \leqq K^2 \|\phi\|_{H_\alpha^2}^2 \tag{4.5}$$

where K *is a constant independent of* ϕ.

Proof of Proposition 4.3. Since $B(T)$ is onto, it has a bounded right inverse $B(T^{-1})$, say. (4.2) implies that for $x \in X$ (with $b_n = 1$ now),

$$\langle x, \psi_n\rangle = \int_0^T e^{\lambda_n (T-s)} (B(T)^{-1} x)(s) \, ds.$$

Letting

$$\phi_x(z) = \int_0^T \bar{e}^{z(T-s)}(B(T)^{-1}(x))(s)\, ds,$$

then $\phi_x \in H_\alpha^2$ and

$$\|\phi_x\|_{H_\alpha^2} = \left(\int_0^T |e^{-\alpha(T-s)}(B(T)^{-1}x)(s)|^2\, ds\right)^{1/2}$$
$$\leqq \max\{1, e^{-\alpha T}\}\|(B(T)\|^{-1}\|x\|.$$

Combining this with inequality (4.5) as applied to the measure induced by $\{-\lambda_n\}$, we have $(\sum_{n-\infty}^{\infty}|\phi_x(-\lambda_n)|^2)^{1/2} \leqq K_1\|x\|$ where $K_1 = K \max\{1, e^{-\alpha T}\}\|B(T)^{-1}\|$. Since $\phi_x(-\lambda_n) = \langle x, \psi_n\rangle$, the proof of the proposition is completed.

Remark. The conclusion of Proposition 4.3 is equivalent to saying that for every $\{h_n\} \in l^2$ the mapping $\langle x, h\rangle = \sum_{n=-\infty}^{\infty} h_n\langle x, \psi_n\rangle$ is a continuous linear functional in X.

THEOREM 4.4. *Suppose the system* (2.1) *is approximately controllable in some time T and the operator A has only simple eigenvalues λ_n, $-\infty < n < \infty$. (So Assumptions* A *and* B *are satisfied.) Suppose also that the set $\{b_n\}$ is uniformly bounded, $\{{}^{1/\lambda}n\} \in l^2$ and every h of the form in* (4.3) *represents a continuous linear functional on X. Denoting $\beta_n = \sum_{m\neq n} 1/|\lambda_m - \lambda_n|^2$, if $\{\mu_n\}_{n=-\infty}^{\infty}$ is a sequence of complex numbers such that*

$$\sum_{n=-\infty}^{\infty} \frac{(1+\beta_n)|\mu_n - \lambda_n|^2}{|b_n|^2} < \infty, \tag{4.6}$$

then there exists $h \in X$ of the form in (4.3) *such that $\sigma(A_h) = \{\mu_n\}_{n=-\infty}^{\infty}$.*

Proof. Let $h \in X'$ be defined by

$$\langle x, h\rangle = \sum_n \langle x, \psi_n\rangle h_n, \qquad \{h_n\} \in l^2.$$

We first find a more explicit form for $\Delta_h(\lambda)$ (cf. Lemma 3.1). Let $R_\lambda b$ be the unique element in X such that

$$\langle b, R(\bar{\lambda}, A^*)x\rangle = \langle R_\lambda b, x\rangle$$

for all x in X. Clearly, $\langle R_\lambda b, \psi_n\rangle = \langle b, \psi_n\rangle/(\lambda - \lambda_n)$.

Hence

$$\begin{aligned}\Delta_h(\lambda) &= \langle R_\lambda b, h\rangle - 1\\ &= \sum_n \langle R_\lambda b, \psi_n\rangle h_n - 1\\ &= \sum_n \frac{\langle b, \psi_n\rangle h_n}{\lambda - \lambda_n} - 1\\ &= \sum_n \frac{b_n h_n}{\lambda - \lambda_n} - 1.\end{aligned}$$

So in order that $\sigma(A_n)$ be equal to a given sequence $\{\mu_n\}$, a necessary and sufficient condition is that

$$\sum_{n\in\mathcal{N}} \frac{b_n h_n}{\mu_n - \lambda_n} = 1 \quad \text{for all } m \in \mathcal{N} \tag{4.7}$$

where $\mathcal{N} = \{n: \mu_n \neq \lambda_n\}$.

For simplicity of notation, we assume $\mathcal{N}=\mathcal{Z}$. The proof for the case $\mathcal{N}\subset\mathcal{Z}$ is similar. Multiplying both side of (4.7) by $(\mu_m-\lambda_m)/b_m$, it becomes

$$\sum_n \frac{b_n(\mu_m-\lambda_m)}{b_m(\mu_m-\lambda_n)}h_n=\frac{\mu_m-\lambda_m}{b_m} \quad \text{for all } m. \tag{4.8}$$

We can rewrite this in the form of a matrix equation

$$M\mathbf{h}=\mathbf{c}. \tag{4.9}$$

where M is the infinite matrix with i, j entries $(b_j(\mu_i-\lambda_i))/(b_i(\mu_i-\lambda_j))$, $\mathbf{h}$ and $\mathbf{c}$ are infinite column vectors with ith entry equal to h_i and $(\mu_i-\lambda_i)/b_i$ respectively. Notice that the diagonal elements of M are 1 and the sum of the squares of the off diagonal elements is finite.

Indeed (4.6) implies that

$$\sum_{\substack{n,m\\ n\neq m}}\left|\frac{\mu_n-\lambda_n}{\mu_m-\lambda_n}\right|^2<\infty.$$

Hence there is an integer $N>0$ such that

$$\left|\frac{\mu_n-\lambda_n}{\lambda_m-\lambda_n}\right|<\frac{1}{2} \quad \text{wherever } |m|\geqq N \text{ or } |n|\geqq N.$$

So

$$\begin{aligned}\frac{1}{|\mu_m-\lambda_m|}&\leqq\frac{1}{(|\lambda_m-\lambda_n|-|\mu_n-\lambda_n)|}\\ &\leqq\frac{1}{(|\lambda_m-\lambda_n|-\frac{1}{2}|\lambda_m-\lambda_n|)}=\frac{2}{|\lambda_m-\lambda_n|}\end{aligned}$$

if $|m|\geqq N$ or $|n|\geqq N$.

It follows that the set

$$\left\{\frac{|\lambda_m-\lambda_n|}{|\mu_m-\lambda_n|}: m\neq n\right\}$$

is uniformly bounded. Let K be an upper bound for this set. Then

$$\begin{aligned}\sum_{\substack{m,n\\ m\neq n}}\left|\frac{b_n(\mu_m-\lambda_n)}{b_m(\mu_m-\lambda_n)}\right|^2&\leqq\sup_n|b_n|^2K^2\sum_{\substack{m,n\\ m\neq n}}\left|\frac{\mu_m-\lambda_m}{b_m(\lambda_m-\lambda_n)}\right|^2\\ &=\sup_n|b_n|^2K^2\sum_m\left|\frac{\mu_m-\lambda_m}{b_m}\right|^2\beta_m\\ &<\infty.\end{aligned}$$

Hence the sum of the squares of the off diagonal elements of M is indeed finite.

So by Fredholm Theory [15, Chap. 9, § 17], (4.9) has a solution if the determinant of M is convergent, i.e. if $\det M_n$ converges to a nonzero limit where M_n is the truncated $2n+1\times 2n+1$ matrix

$$(a_{ij})_{\substack{-n\leqq i\leqq n\\ -n\leqq j\leqq n}}.$$

By [12, Lemma 3.1] the limit of $\det M_n$ is given by the infinite product

$$\prod_{n=-\infty}^{\infty}\prod_{m=n+1}^{\infty}\frac{(\mu_m-\mu_n)(\lambda_n-\lambda_m)}{(\lambda_m-\mu_n)(\lambda_n-\mu_m)}=\prod_{n=-\infty}^{\infty}\prod_{m=n+1}^{\infty}\left(1+\frac{(\mu_m-\lambda_m)(\lambda_n-\mu_n)}{(\lambda_m-\mu_n)(\lambda_n-\mu_m)}\right). \tag{4.10}$$

Now by Schwartz' inequality,

$$\sum_{n=-\infty}^{\infty}\sum_{m=n+1}^{\infty}\left|\frac{(\mu_m-\lambda_m)(\lambda_n-\mu_n)}{(\lambda_m-\mu_n)(\lambda_n-\mu_m)}\right| \leqq \sum_{n=-\infty}^{\infty}\sum_{\substack{m=-\infty\\ m\neq n}}^{\infty}\left|\frac{\mu_n-\lambda_n}{\lambda_m-\mu_n}\right|^2. \tag{4.11}$$

Since $\sum_{n=-\infty}^{\infty}\beta_n|\mu_n-\lambda_n|^2$ is finite, we can choose N sufficiently large such that $|\mu_n-\lambda_n|^2\beta_n<\frac{1}{4}$ if $|n|>N$. This implies that if $|n|>N$, then $|\mu_n-\lambda_n|^2<\frac{1}{4}|\lambda_m-\lambda_n|^2$ for all $m\neq n$. Hence (4.11) implies that

$$\sum_{n=-\infty}^{\infty}\sum_{m=n+1}^{\infty}\left|\frac{(\mu_m-\lambda_m)(\lambda_n-\mu_n)}{(\lambda_m-\mu_n)(\lambda_n-\mu_m)}\right|$$

$$\leqq \sum_{n=-N}^{N}\sum_{\substack{m=-\infty\\ m\neq n}}^{\infty}\left|\frac{\mu_n-\lambda_n}{\mu_n+\lambda_m}\right|^2+\sum_{|n|>N}\sum_{\substack{m=-\infty\\ m\neq n}}^{\infty}\frac{|\mu_n-\lambda_n|^2}{\|\lambda_m-\lambda_n|-|\mu_n-\lambda_n\|^2}.$$

The first term of the last expression is finite and the second term is less than

$$4\sum_{|n|>N}\sum_{\substack{m=-\infty\\ m\neq n}}^{\infty}\frac{|\mu_n-\lambda_n|^2}{|\lambda_m-\lambda_n|^2}=4\sum_{|n|>N}|\mu_n-\lambda_n|^2\beta_n.$$

So by familiar results, [14, p. 15], for example, the infinite product in (4.10) converges. So (4.9) has a solution $\mathbf{h}=\{h_n\}$ in l^2. This completes the proof of the theorem.

5. A degenerate hyperbolic system. We consider the two-dimensional control system

$$\frac{\partial}{\partial t}\begin{pmatrix}v\\ w\end{pmatrix}(x,t)=\begin{pmatrix}-\gamma_1 & 0\\ 0 & \gamma_2\end{pmatrix}\frac{\partial}{\partial x}\begin{pmatrix}v\\ w\end{pmatrix}(x,t)+B\begin{pmatrix}v\\ w\end{pmatrix}(x,t),$$

$$v(0,t)=0, \qquad w(1,t)=u(t), \tag{5.1}$$

$$v(x,0)=v_0(x), \qquad w(x,0)=w_0(x).$$

Here both γ_1 and γ_2 are positive numbers,

$$B=\begin{pmatrix}b_{11} & b_{12}\\ b_{21} & b_{22}\end{pmatrix}$$

is a 2×2 matrix and $(x,t)\in\Omega\triangleq\{(\xi,s)\colon 0\leqq\xi\leqq 1, 0\leqq s\}$. This system is a degenerate case of the more general system with the boundary condition in (5.1) being replaced by

$$\alpha_1 v(0,t)+\alpha_2 w(0,t)=0, \qquad \beta_1 v(1,t)+\beta_2 w(1,t)=u(t). \tag{5.2}$$

The fact that $\alpha_2=\beta_1=0$ in our case says roughly that information is lost as it reaches the boundary and thus this system is not time reversible.

To put this system in the abstract form (2.1), we let

$$X=\left\{\begin{pmatrix}v(\cdot)\\ W(\cdot)\end{pmatrix}\colon v(\cdot)\in H^1[0,1],\ w(\cdot)\in L^2[0,1],\ v(0)=0\right\}$$

equipped with the $H^1[0,1]\times L^2[0,1]$ norm.

$$A\begin{pmatrix}v\\ w\end{pmatrix}=\begin{pmatrix}-\gamma_1 & 0\\ 0 & \gamma_2\end{pmatrix}\frac{d}{dX}\begin{pmatrix}v\\ w\end{pmatrix}+B\begin{pmatrix}v\\ w\end{pmatrix} \tag{5.3}$$

with

$$D(A)=\left\{\begin{pmatrix} v(\cdot) \\ w(\cdot) \end{pmatrix}: v(\cdot)\in H^2[0,1],\ w(\cdot)\in H^1[0,1],\right.$$
$$\left. v(0)=w(1)=0,\ -\gamma_1 v'(0)+b_{12}w(0)=0\right\}.$$

We shall identify X' as a space of distributions so that

$$X\subseteq (L^2[0,1])^2\subseteq X'$$

and the duality pairing of $X\times X'$ when restricted to $X\times(L^2[0,1])^2$ coincides with the inner product on $(L^2[0,1])^2$.

We denote by A' the operator from X' into X' so that

$$D(A')=\{y\in X': \text{there exists a constant } K \text{ such that } \langle Ax, y\rangle_{X\times X'}\leqq K\|x\|_X \text{ for all } x\in D(A)\}$$

and $\langle x, A'y\rangle_{X\times X'}=\langle Ax, y\rangle_{X\times X'}$ for all $x\in D(A)$, $y\in D(A')$.

Let C be the isometry from X onto X' defined in such a way so that $\langle x, Cy\rangle_{X\times X'}=\langle x, y\rangle_{X\times X}$ for all $x, y\in X$. The right-hand side is the inner product of x and y in X. The relation between A' and A^* is as follows.

(i) $x\in D(A^*)$ if and only if $Cx\in D(A')$.

(ii) $CA^*=A'C$.

(iii) If we equip $D(A')$ with the norm

$$\|y\|_{D(A')}=\|y\|_{X'}+\|A'y\|_{X'}$$

then C is an isometry from $D(A^*)$ onto $D(A')$.

We define

$$\hat{b}\begin{pmatrix} y \\ z \end{pmatrix}=\gamma_2 z(1) \quad \text{for} \begin{pmatrix} y \\ z \end{pmatrix}\in D(A'). \tag{5.4}$$

We now prove some results showing that the various assumptions we made in the preceding sections are satisfied by their system.

LEMMA 5.1. *The linear functions $\hat{b}$ defined in* (5.2) *is a continuous linear functional on $D(A')$.*

Proof. We define another operator A_1 on x:

$$A_1\begin{pmatrix} v \\ w \end{pmatrix}=\begin{pmatrix} -\gamma_1 & 0 \\ 0 & \gamma_2 \end{pmatrix}\frac{d}{dx}\begin{pmatrix} v \\ w \end{pmatrix}+B\begin{pmatrix} v \\ w \end{pmatrix} \tag{5.5}$$

with

$$D(A_1)=\left\{\begin{pmatrix} v(\cdot) \\ w(\cdot) \end{pmatrix}: v\in H^2[0,1],\ w\in H^1[0,1],\ v(0)=0,\ -\gamma_1 v'(0)+b_{12}w(0)=0\right\}.$$

We claim that

$$\left\langle A_1\begin{pmatrix} v \\ w \end{pmatrix}, \begin{pmatrix} y \\ Z \end{pmatrix}\right\rangle-\left\langle \begin{pmatrix} v \\ w \end{pmatrix}, A'\begin{pmatrix} y \\ z \end{pmatrix}\right\rangle=w(1)\hat{b}\begin{pmatrix} y \\ z \end{pmatrix} \tag{5.6}$$

for all

$$\begin{pmatrix} v \\ w \end{pmatrix}\in D(A_1) \quad \text{and} \quad \begin{pmatrix} y \\ z \end{pmatrix}\in D(A').$$

If (5.6) indeed holds, then obviously $\hat{b}$ is a continuous linear functional on $D(A')$ because we can certainly find some $\binom{v}{w}$ in $D(A_1)$ such that $w(1)\neq 0$. So it remains to prove (5.6).

We identify the dual space of X as consisting of ordered pairs of distributions so that

$$X \subseteq (L^2[0,1])^2 \subseteq X'$$

and the duality pairing when restricted to $X \times (L^2[0,1])^2$ coincides with the inner product on $(L^2[0,1])^2$.

Let $F \in X'$, then there are functions f_1 and f_2 in $L^2[0,1]$ such that

$$\left\langle \begin{pmatrix} v \\ w \end{pmatrix}, F \right\rangle = \int_0^1 v'(x) f_1(x) + w(x) f_2(x)\, dx \quad \text{for all } \begin{pmatrix} v \\ w \end{pmatrix} \in X. \tag{5.7}$$

Suppose $F \in D(A')$ and $A'F = G$; then since

$$\left\langle A\begin{pmatrix} v \\ w \end{pmatrix}, F \right\rangle = \int_0^1 (-\gamma_1 v'' + b_{11} v' + b_{12} w') f_1 + (\gamma_2 w' + b_{21} v + b_{22} w) f_2\, dx$$

and

$$\left\langle \begin{pmatrix} v \\ w \end{pmatrix}, G \right\rangle = \int_0^1 v' g_1 + w g_2\, dx$$

for some $g_1, g_2 \in L^2[0,1]$, it follows that

$$\begin{aligned} &f_1, f_2, \gamma_2 f_1' - g_1 \in H'[0,1], \\ &f_1(1) = f_2(0) = (\gamma_1 f_1' - g_1)(1) = 0, \\ &(\gamma_1 f_1' - g_1)' + b_{11} f_1' + b_{21} f_2' = 0 \\ &b_{12} f_1' + \gamma_2 f_2' = b_{22} f_2 - g_2. \end{aligned} \tag{5.8}$$

Hence integrating the right-hand side of (5.7) by parts, we get

$$\begin{aligned} \left\langle \begin{pmatrix} v \\ w \end{pmatrix}, F \right\rangle &= \int_0^1 -v(x) f_1'(x) + w(x) f_2(x)\, dx \\ &= \int_0^1 v(x)\overline{y(x)} + w(x)\overline{z(x)}\, dx \end{aligned}$$

where $y(x) = \overline{-f_1'(x)}$ and $z(x) = \overline{f_2(x)}$. So we see that

$$D(A') = \left\{ \begin{pmatrix} y \\ z \end{pmatrix} : y \in L^2[0,1],\ z \in H^1[0,1] \text{ and } z(0) = 0 \right\}.$$

From (5.6), we have

$$\begin{aligned} &-\gamma_1 \bar{y} - g_1 \in H^1[0,1], \\ &(-\gamma_1 \bar{y} - g_1)(0) = 0, \\ &(-\gamma_1 \bar{y} - g_1)' + b_{11} \bar{y} + b_{21} \bar{z} = 0, \\ &g_2 = -\gamma_2 \bar{z}' + b_{12} \bar{y} + b_{22} \bar{z}. \end{aligned}$$

Hence

$$\left\langle \begin{pmatrix} v \\ w \end{pmatrix}, A'\begin{pmatrix} y \\ z \end{pmatrix} \right\rangle = \int_0^1 v' g_1 + w(-\gamma_2 \bar{z}' + b_{12} \bar{y} + b_{22} \bar{z})\, dx. \tag{5.9}$$

Now if $\binom{v}{w} \in D(A_1)$, then $\binom{v}{w} - \binom{0}{xw(1)} \in D(A)$. Hence for $\binom{v}{w} \in D(A_1)$ and $\binom{y}{z} \in D(A')$, we have

$$\left\langle A_1\begin{pmatrix} v \\ w \end{pmatrix}, \begin{pmatrix} y \\ z \end{pmatrix}\right\rangle - \left\langle \begin{pmatrix} v \\ w \end{pmatrix}, A\begin{pmatrix} y \\ z \end{pmatrix}\right\rangle = \left\langle A_1\begin{pmatrix} 0 \\ xw(1) \end{pmatrix}, \begin{pmatrix} y \\ z \end{pmatrix}\right\rangle - \left\langle \begin{pmatrix} 0 \\ xw(1) \end{pmatrix}, A'\begin{pmatrix} y \\ z \end{pmatrix}\right\rangle$$

(by (5.9))
$$= \int_0^1 b_{12}xw(1)\bar{y} + \gamma_2 w(1) + b_{22}xw(1)\bar{z}\, dx$$
$$- \int_0^1 xw(1)(-\gamma_2\bar{z}' + b_{12}\bar{y} + b_{22}\bar{z})\, dx$$
$$= \int_0^1 \gamma_2 w(1)(\bar{z} + x\bar{z}')\, dx$$
$$= \int_0^1 \gamma_2 w(1)\frac{d}{dx}(x\bar{z})\, dx$$
$$= \gamma_2 w(1)\overline{z(1)}.$$

So (5.6) indeed holds. This completes the proof of the lemma.

We now consider the following system.

$$\frac{\partial}{\partial t}\begin{pmatrix} v \\ w \end{pmatrix}(x, t) = \begin{pmatrix} -\gamma_1 & 0 \\ 0 & \gamma_2 \end{pmatrix}\frac{\partial}{\partial x}\begin{pmatrix} v \\ w \end{pmatrix}(x, t) + B\begin{pmatrix} v \\ w \end{pmatrix}(x, t) + \begin{pmatrix} 0 \\ g(x, t) \end{pmatrix},$$

(5.10)
$$v(0, t) = 0, \qquad w(1, t) = u(t),$$
$$v(x, 0) = v_0(x), \qquad w(x, 0) = w_0(x).$$

For $T > 0$, we denote

$$\Omega_T \triangleq \{(\xi, s): 0 \leqq \xi \leqq 1, 0 \leqq s \leqq T\}.$$

PROPOSITION 5.2. *Suppose that* $v_0(x)$, $w_0(x) \in C^1[0, 1]$, $g \in C^1(\Omega_T)$ *and* $u(t)$ *is continuous and* $v_0(0) = 0$. *Then* (5.4) *admits a unique solution. Furthermore, there exists a constant* K *independent of* v_0, w_0, u *and* $0 \leqq t \leqq T$ *such that*

$$\left\|\begin{pmatrix} v \\ w \end{pmatrix}(t)\right\|_X \leqq K\left(\left\|\begin{pmatrix} v_0 \\ w_0 \end{pmatrix}\right\|_X + \|u\|_{L^2[0,T]} + \|g\|_{L^2(\Omega_T)}\right). \tag{5.11}$$

Proof. Existence and uniqueness of the solution with the given data is well known (see, e.g., [4, Chap. 5]).

For a point (x, t) in Ω_T, we will denote by $C_1(x, t)$ and $C_2(x, t)$ respectively the backward characteristics passing through (x, t) with slopes $1/\gamma_1$ and $-1/\gamma_2$ respectively and by $P_1(x, t)$ and $P_2(x, t)$ respectively the points at which these characteristics hit the boundary of Ω_T. We parametrize the characteristics by the t variable. Take a fixed point (x, t) in Ω_T. By integrating the first component equation of (5.4) along $C_1(x, t)$, we have

$$v(x, t) = \int_{C_1(x,t)} (b_{11}v + b_{12}w)(x(s), s)\, ds + v(P_1) \tag{5.12}$$

from which we can easily get the estimate

$$\int_0^1 |v(x, t)|^2\, dx \leqq K_1\left(\int_0^t \int_0^1 |v(x, s)|^2 + |w(x, s)|^2\, dx\, ds + \int_0^1 |v_0(x)|^2\, dx\right) \tag{5.13}$$

where K_1 is a constant which can be chosen to be independent of $t \in [0, T]$. Similarly, we have

$$w(x, t) = \int_{C_2(x,t)} (b_{21}v + b_{22}w + g)(x(s), s)\, ds + w(P_2) \tag{5.14}$$

from which we can obtain the inequality

$$\int_0^1 |w(x, t)|^2\, dx \leqq K_2\Big(\int_0^t \int_0^1 |v(x, s)|^2 + |w(x, s)|^2 + |g(x, s)|^2\, dx\, ds + \int_0^1 |v_0(x)|^2\, dx + \int_0^t |u(s)|^2\, ds\Big), \tag{5.15}$$

K_2 a constant independent of t.

Combining (5.13) and (5.15) and applying the Gronwall inequality, we have

$$\int_0^1 |v(x, t)|^2 + |w(x, t)|^2\, dx \leqq K_3\Big(\int_0^T \int_0^1 |g(x, s)|^2\, dx\, ds + \int_0^T |u(s)|^2\, ds + \int_0^1 |v_0(x)|^2 + |w_0(x)|^2\, dx\Big). \tag{5.16}$$

Next, we estimate $\partial v/\partial x$. First,

$$\begin{aligned} \frac{\partial}{\partial x}\int_{C_1(x,t)} v(x(s), s)\, ds &= \frac{\partial}{\partial x}\int_{t-x/\gamma_1}^t v(x + \gamma_1(s-t), s)\, ds \\ &= \frac{1}{\gamma_1} v(0, t - x/\gamma_1) + \int_{t-x/\gamma_1}^t \frac{\partial}{\partial x} v(x + \gamma_1(s-t), s)\, ds \\ &= \frac{1}{\gamma_1} v(P_1(x, t)) + \int_{C_1(x,t)} \frac{\partial}{\partial x} v(x(s), s)\, ds. \end{aligned} \tag{5.17}$$

Also,

$$\frac{\partial}{\partial x}\int_{C_1(x,t)} w(x(s), s)\, ds = \frac{1}{\gamma_1} w(0, t - x/\gamma_1) + \int_{t-x/\gamma_1}^t \frac{\partial}{\partial x} w(x + \gamma_1(s-t), s)\, ds.$$

We can rewrite the second term in the above expression. Since

$$\begin{aligned} &\int_{t-x/\gamma_1}^t \frac{\partial}{\partial x} w(x + \gamma_1(s-t), s) + \frac{\partial}{\partial s} w(x + \gamma_1(s-t), s)\, ds \\ &\quad = \int_{t-x/\gamma_1}^t \frac{d}{ds} w(x + \gamma_1(s-t), s)\, ds \\ &\quad = w(x, t) - w(P_1(x, t)) \end{aligned} \tag{5.18a}$$

and because of (5.10), we have

$$\gamma_2 \frac{\partial}{\partial x} w(x, s) - \frac{\partial w}{\partial s}(x, s) + b_{21}v(x, s) + b_{22}w(x, s) + g(x, s) = 0, \tag{5.18b}$$

it follows from (5.18a) and (5.18b) that

$$\int_{C_1(x,t)} (\gamma_1 + \gamma_2)\frac{\partial}{\partial x} w(x(s), s) + b_{21}v(x(s), s) + b_{22}w(x(s), s) + g(x(s), s)\, ds = w(x, t) - w(P_1(x, t)).$$

Hence,

$$\frac{\partial}{\partial x}\int_{C_1(x,t)} w(x(s), s)\,ds = \left(\frac{1}{\gamma_1}-\frac{1}{\gamma_1+\gamma_2}\right)w(P_1(x,t))+\frac{1}{\gamma_1+\gamma_2}w(x,t)$$

$$+\frac{1}{\gamma_1+\gamma_2}\int_{C_1(x,t)} -b_{21}v(x(s), s)-b_{22}w(x(s), s)-g(x(s), s)\,ds. \tag{5.19}$$

Combining (5.12), (5.17) and (5.19) we can easily obtain

$$\int_0^1 \frac{\partial}{\partial x}|v(x,t)|^2\,dx \leqq K_4\Bigg(\int_0^t\int_0^1 \left|\frac{\partial}{\partial x}v(x,s)\right|^2+|w(x,s)|^2\,dx\,ds$$

$$+\int_0^T\int_0^1 |g(x,s)|^2\,dx\,ds \tag{5.20}$$

$$+\int_0^1 |v_0(x)|^2+|w_0(x)|^2+\left|\frac{\partial}{\partial x}v_0(x)\right|^2 dx+\int_0^T |u(s)|^2\,ds\Bigg).$$

The desired result then follows by applying Gronwall's inequality to (5.20) and combining the result with (5.16). This finishes the proof of the proposition.

Letting u and g be zero in (5.10), we have the estimate

$$\left\|\begin{pmatrix} v \\ w\end{pmatrix}(t)\right\|_X \leqq K\left\|\begin{pmatrix} v_0 \\ w_0\end{pmatrix}\right\|_X,$$

where K is a constant independent of $t\in[0, T]$. Since T is arbitrary, we have in general

$$\left\|\begin{pmatrix} v \\ w\end{pmatrix}(t)\right\|_X \leqq K_t\left\|\begin{pmatrix} v_0 \\ w_0\end{pmatrix}\right\|$$

where K_t is some constant depending on t. This implies that when we apply no control on the system (5.1), the mapping carrying $\binom{v_0}{w_0}$ to $\binom{v(t)}{w(t)}$ can be extended to a continuous linear operator mapping X into itself. We denote this operator by $S(t)$. It is easy to see that the family $\{S(t): t\geqq 0\}$ forms a strongly continuous semigroup of continuous linear operators on X. It is easy to verify that the generator of this semigroup is exactly A. So we have proved.

PROPOSITION 5.3. *The operator as defined in* (5.3) *generates a strongly continuous semigroup in X.*

PROPOSITION 5.4. *Let $\hat{b}$ be the linear function on $D(A')$ as defined in* (5.4). *Then the linear functional $b=\hat{b}C$ defined on $D(A^*)$ is an admissible input element. Here C denotes the isometry from X into X' such that $\langle x, Cy\rangle_{X\times X'}=\langle x, y\rangle_{X\times X}$ for all $x, y\in X$.*

Proof. Let v and w be the solution of

$$\frac{\partial}{\partial t}\begin{pmatrix} v \\ w\end{pmatrix}=\begin{pmatrix} -\gamma_1 & 0 \\ 0 & \gamma_2\end{pmatrix}\frac{\partial}{\partial x}\begin{pmatrix} v \\ w\end{pmatrix}+B\begin{pmatrix} v \\ w\end{pmatrix},$$

$$v(0,t)=0,\quad w(1,t)=u(t),\quad v(x,0)=w(x,0)=0.$$

We assume that $u\in c^1(0,\tau)$, $\tau>0$. For each fixed t, $0\leqq t\leqq\tau$, we have

$$\begin{pmatrix} v(x,t) \\ w(x,t)\end{pmatrix}\in D(A_1)$$

and

$$\frac{d}{dt}\begin{pmatrix} v(x,t) \\ w(x,t)\end{pmatrix}=A_1\begin{pmatrix} v(x,t) \\ w(x,t)\end{pmatrix}.$$

Let

$$\begin{pmatrix} y_0 \\ z_0 \end{pmatrix} \in D(A').$$

Then (5.6) implies that

$$\begin{aligned} &\left\langle \frac{d}{dt}\begin{pmatrix} v(x,t) \\ w(x,t) \end{pmatrix}, S'(\tau-t)\begin{pmatrix} y_0 \\ z_0 \end{pmatrix} \right\rangle - \left\langle \begin{pmatrix} v(x,t) \\ w(x,t) \end{pmatrix}, A'S'(\tau-t)\begin{pmatrix} y_0 \\ z_0 \end{pmatrix} \right\rangle \\ &\quad = u(t)\overline{\hat{b}\left(S'(\tau-t)\begin{pmatrix} y_0 \\ z_0 \end{pmatrix} \right)}. \end{aligned} \tag{5.20)'}$$

Note that the left-hand side of the above equality equals

$$\frac{d}{dt}\left\langle \begin{pmatrix} v(x,t) \\ w(x,t) \end{pmatrix}, S'(\tau-t)\begin{pmatrix} y_0 \\ z_0 \end{pmatrix} \right\rangle.$$

Hence integrating from 0 to τ, one has

$$\left\langle \begin{pmatrix} v(x,t) \\ w(x,t) \end{pmatrix}, \begin{pmatrix} y_0 \\ z_0 \end{pmatrix} \right\rangle = \int_0^\tau u(t)\overline{\hat{b}\left(S'(\tau-t)\begin{pmatrix} y_0 \\ z_0 \end{pmatrix} \right)}\, dt.$$

But by Proposition 5.2, we have (on letting $v_0 = w_0 = 0$, $g = 0$)

$$\left\| \begin{pmatrix} v(\cdot,\tau) \\ w(\cdot,\tau) \end{pmatrix} \right\|_X \leqq K\left(\int_0^\tau |u(s)|^2\, ds \right)^{1/2}$$

the constant K being independent of τ for τ sufficiently small. Hence (5.20)' implies that

$$\int_0^\tau u(t)\overline{\hat{b}\left(S(\tau-t)\begin{pmatrix} y_0 \\ z_0 \end{pmatrix} \right)}\, dt \leqq K\left(\int_0^\tau |u(s)|^2\, ds \right)^{1/2} \left\| \begin{pmatrix} y_0 \\ z_0 \end{pmatrix} \right\|_{X'}.$$

Since u is an arbitrary function in $C^1(0, \tau)$, it follows that

$$\int_0^\tau \left| \hat{b}\left(S'(\tau-t)\begin{pmatrix} y_0 \\ z_0 \end{pmatrix} \right) \right| dt \leqq K^2 \left\| \begin{pmatrix} y_0 \\ z_0 \end{pmatrix} \right\|_{X'}^2$$

for all $\binom{y_0}{z_0} \in D(A')$. Let $C\mathbf{y} = \binom{y_0}{z_0}$. Then

$$\hat{b}\left(S'(\tau-t)\begin{pmatrix} y_0 \\ z_0 \end{pmatrix} \right) = \hat{b}(S'(\tau-t)C\mathbf{y}) = \hat{b}(CS^*(\tau-t)\mathbf{y}) = b(S^*(\tau-t)\mathbf{y}).$$

So we have $\int_0^\tau |b(S^*(\tau-t)\mathbf{y})|^2 \leqq K^2\|\mathbf{y}\|_{X'}^2$. This proves that b is an admissible input element.

Next, we prove a controllability result. We will denote

$$T = \frac{1}{\gamma_1} + \frac{1}{\gamma_2}.$$

THEOREM 5.5. *For the system* (5.1), *if* $b_{12} \neq 0$ *and* $|b_{21}|$ *is sufficiently small, then given any* $\binom{v_T}{w_T}$ *in* X, *we can find a unique* $u \in L^2[0, T]$ *such that the solution of* (5.1) *with initial condition* $v_0 = w_0 = 0$ *satisfies*

$$\begin{pmatrix} v \\ w \end{pmatrix}(x, T) = \begin{pmatrix} v_T \\ w_T \end{pmatrix}(x), \qquad 0 \leqq x \leqq 1. \tag{5.21}$$

Proof. Case 1: $b_{21}=0$. In this case, by integrating along characteristic $C_2(x, t)$ we have the relation

$$w(x, T)=\exp[b_{22}(1-x)/\gamma_2]u(T-(1-x)/\gamma_2)$$

from which we can solve for u as a function in $L^2[1/\gamma_1, T]$.

Next,

$$\begin{aligned} v(x, T) &= \int_{-x/\gamma_2}^{0} b_{12}\exp[-b_{11}s+b_{22}(1-x-\gamma_1)s/\gamma_2] \\ &\quad \times u(T+s-(1-x-\gamma_1 s)/\gamma_2)\,ds \\ &= \int_{(1-x)/\gamma_1}^{t-(1-x)/\gamma_2} b_{12}\gamma_2/(\gamma_1+\gamma_2)\cdot\exp[\alpha_1(1-x)+\alpha_2(t-\xi)]u(\xi)\,d\xi \end{aligned}$$

where $\alpha_1=(b_{22}-b_{11})/(\gamma_1+\gamma_2)$ and $\alpha_2=(b_{11}\gamma_1+b_{22}\gamma_2)/(\gamma_1+\gamma_2)$.

Differentiating with respect to x, we obtain

$$\begin{aligned} \frac{d}{dx}v_T(x) &= -\alpha_1 v_T(x)+b_{12}\gamma_2\exp[b_{22}(1-x)/\gamma_2]u(T-(1-x)/\gamma_2)/(\gamma_1+\gamma_2) \\ &\quad - b_{12}\gamma_2\exp[(b_{22}/\gamma_2)-(b_{11}\gamma_1)]u\left(\frac{1}{\gamma_2}-\frac{x}{\gamma_1}\right)\Big/(\gamma_1+\gamma_2). \end{aligned}$$

Since $u(T-(1-x)/\gamma_2)$ is known for $0\leqq x\leqq 1$, we could solve for $u(T-x/\gamma_1-1/\gamma_2)$ as a function in L^2. Because $\tau=T-x/\gamma_1-1/\gamma_2$ ranges from 0 to $1/\gamma_1$ as x ranges from 0 to 1, this gives u as a function in $L^2[0, 1/\gamma_1]$. This finishes the proof for Case 1.

Case 2: $b_{21}\neq 0$ *but is sufficiently small.* We prove by a fixed point argument. We introduce the two systems

$$\frac{\partial}{\partial t}\begin{pmatrix}\hat{v}\\ \hat{w}\end{pmatrix}(x, t)=\begin{pmatrix}-\gamma_1 & 0\\ 0 & \gamma_2\end{pmatrix}\frac{\partial}{\partial x}\begin{pmatrix}\hat{v}\\ \hat{w}\end{pmatrix}(x, t)+B\begin{pmatrix}\hat{v}\\ \hat{w}\end{pmatrix}(x, t),$$

$(\tilde{\ })$ $\quad \hat{v}(0, t)=0, \qquad \hat{w}(1, t)=\hat{u}(t),$

$$\hat{v}(x, 0)=\hat{w}(x, 0)=0;$$

$$\frac{\partial}{\partial t}\begin{pmatrix}\tilde{v}\\ \tilde{w}\end{pmatrix}(x, t)=\begin{pmatrix}-\gamma & 0\\ 0 & \gamma_2\end{pmatrix}\frac{\partial}{\partial x}\begin{pmatrix}\tilde{v}\\ \tilde{w}\end{pmatrix}(x, t)+B\begin{pmatrix}\tilde{v}\\ \tilde{w}\end{pmatrix}(x, t)+\begin{pmatrix}0 & 0\\ b_{21} & 0\end{pmatrix}\begin{pmatrix}\hat{v}\\ \hat{w}\end{pmatrix}(x, t),$$

$(\hat{\ })$ $\quad \tilde{v}(0, t)=\tilde{w}(1, t)=0,$

$$\tilde{v}(x, 0)=\tilde{w}(x, 0)=0.$$

We define an operator M from $L^2[0, T]$ into itself as follow.

For $u\in L^2[0, T]$, let $\begin{pmatrix}\hat{v}\\ \hat{w}\end{pmatrix}$ be the solution of $(\hat{\ })$ with $\hat{u}=u$ and $\begin{pmatrix}\tilde{v}\\ \tilde{w}\end{pmatrix}$ be the corresponding solution of $(\tilde{\ })$. $M(u)$ is the control so that if $\hat{u}=M(u)$ in $(\hat{\ })$, then its solution $\begin{pmatrix}\hat{v}_1\\ \hat{w}_1\end{pmatrix}$ satisfies

$$\begin{pmatrix}\hat{v}_1\\ \hat{w}_1\end{pmatrix}(x, T)=\begin{pmatrix}v_T\\ w_T\end{pmatrix}(x)-\begin{pmatrix}\tilde{v}\\ \tilde{w}\end{pmatrix}(x, T).$$

Applying Proposition 5.2, we see that $\begin{pmatrix}\tilde{v}\\ \tilde{w}\end{pmatrix}\in X$ and hence by Case 1, $M(u)$ exists and is unique. Now, we want to show that M is a contraction mapping. Indeed, if u_1 and u_2 are in $L^2[0, T]$, let

$$\begin{pmatrix}\hat{v}_1\\ \hat{w}_1\end{pmatrix}, \qquad \begin{pmatrix}\tilde{v}_1\\ \tilde{w}_1\end{pmatrix}, \qquad \begin{pmatrix}\hat{v}_2\\ \hat{w}_2\end{pmatrix}, \qquad \begin{pmatrix}\tilde{v}_2\\ \tilde{w}_2\end{pmatrix}$$

be respectively the corresponding solutions of ($\hat{}$) and ($\tilde{}$). Applying Proposition 5.1 to ($\tilde{}$) and then to ($\hat{}$), we have

$$\left\|\begin{pmatrix}\tilde{v}_1\\ \tilde{w}_1\end{pmatrix}(\cdot, T)-\begin{pmatrix}\tilde{v}_2\\ \tilde{w}_2\end{pmatrix}(\cdot, T)\right\| \leqq Kb_{21}\int_0^T\int_0^1 |\hat{v}_1(x, w)-\hat{v}_2(x, s)|^2\, dx\, ds$$

$$\leqq K^2 b_{21}\int_0^T |u_1(s)-u_2(s)|^2\, ds.$$

We remark that although K may depend on b_{21}, it remains bounded if b_{21} is bounded. So we may choose b_{21} small enough so that $b_{21}K^2<1$. M is then a contraction mapping and has a unique fixed point u, say. It is easy to see that u makes (5.21) true if and only if it is a fixed point of M. This finishes the proof of the theorem.

Next, we want to determine the location of the eigenvalues of A.

PROPOSITION 5.6. *The eigenvalues of A as defined in* (5.3) *are given by the asymptotic formula*

$$\lambda_n=\frac{1}{T}(-2\log|2n\pi|+i2n\pi+\varepsilon i\pi)+\omega+o(1) \tag{5.22}$$

where

$$\varepsilon=\begin{cases}-1 & \text{when } n>0,\\ 1 & \text{when } n<0.\end{cases}$$

Here, we assume that both b_{12} and b_{21} are nonzero.

Proof. If λ is an eigenvalue of A, then λ satisfies

$$\begin{pmatrix}-\gamma_1 & 0\\ 0 & \gamma_2\end{pmatrix}\frac{d}{dx}\begin{pmatrix}v\\ w\end{pmatrix}+B\begin{pmatrix}v\\ w\end{pmatrix}=\lambda\begin{pmatrix}v\\ w\end{pmatrix},\qquad v(0)=w(1)=0. \tag{5.23}$$

Once λ is fixed, (5.23) can be solved explicitly in terms of initial values $v(0)$ and $w(0)$. Indeed, we can rewrite (5.23) in the form

$$\frac{d}{dx}\begin{pmatrix}v\\ w\end{pmatrix}=M\begin{pmatrix}v\\ w\end{pmatrix} \tag{5.24}$$

where

$$M=\begin{bmatrix}(b_{11}-\lambda)/\gamma_1 & b_{12}/\gamma_1\\ -b_{21}/\gamma_2 & -(b_{22}-\lambda)/\gamma_2\end{bmatrix}.$$

Hence the solution of (5.23) is

$$\begin{pmatrix}v(x)\\ w(x)\end{pmatrix}=e^{xM}\begin{pmatrix}v(0)\\ w(0)\end{pmatrix}. \tag{5.25}$$

Because $v(0)=w(1)=0$, we have

$$\begin{pmatrix}v(1)\\ 0\end{pmatrix}=e^{M}\begin{pmatrix}0\\ w(0)\end{pmatrix}. \tag{5.26}$$

But $e^M=rI+sM$ where

$$r=(\lambda_2e^{\lambda_1}-\lambda_1e^{\lambda_2})/(\lambda_2-\lambda_1) \tag{5.27}$$

and

$$s=(e^{\lambda_2}-e^{\lambda_1})/(\lambda_2-\lambda_1) \tag{5.28}$$

[6, p. 101] and λ_1, λ_2 are the characteristic roots (which we may assume to be distinct when $|\lambda|$ is sufficiently large) of M. It follows from (5.26) that $E_{22}=0$ where E_{22} is the entry on the second row and second column of e^M.

Hence

$$E_{22}=r-s(b_{22}-\lambda)/\gamma_2=0. \tag{5.29}$$

Substituting (5.27) and (5.28) into (5.29), we have

$$(\lambda_2-\lambda_1)^{-1}\left(\left(\lambda_2+\frac{b_{22}-\lambda}{\gamma_1}\right)e^{\lambda_1}-\left(\lambda_1+\frac{b_{22}-\lambda}{\gamma_2}\right)e^{\lambda_2}\right)=0.$$

Now $\lambda_1=\frac{1}{2}((b_{11}-\lambda)/\gamma_1-(b_{22}-\lambda)/\gamma_2+\sqrt{D})$ and $\lambda_2=\frac{1}{2}((b_{11}-\lambda)/\gamma_1-(b_{22}-\lambda)/\gamma_2-\sqrt{D})$ where $D=((b_{11}-\lambda)/\gamma_1+(b_{22}-\lambda)/\gamma_2)^2-\alpha$ with $\alpha=b_{12}b_{21}/\gamma_1\gamma_2$. Letting $\xi=-(b_{11}-\lambda)/\gamma_1-(b_{22}-\lambda)/\gamma_2=-\alpha+\lambda T$, then (5.29) holds if and only if

$$((-\xi-\sqrt{D})\,e^{1/2\sqrt{D}}-(-\xi+\sqrt{D})\,e^{1/2\sqrt{D}})/\sqrt{D}=0 \tag{5.30}$$

where $D=\xi^2-b_{12}b_{21}/\gamma_1\gamma_2$.

An application of the Rouche's Theorem shows that the roots of (5.30) are asymptotically the same as that of the equation $\xi^2e^{\xi}-\alpha/4=0$. By [1, p. 410], we see that the roots of this equation are of the form.

$$\xi_n=-2\log|2n\pi|+i2n\pi+\varepsilon i\pi+\omega_1+o(1).$$

It follows that $\lambda_n=(1/T)(\xi_n+\alpha)$ is asymptotically of the form (5.22). This completes the proof of the theorem.

From the above proposition, it is easy to see that $\{-\lambda_n\}$ induces a Carleson measure on some half plane C_α. Furthermore, the numbers $\beta_n=\sum_{m\neq n}1/|\lambda_n-\lambda_m|^2$ are uniformly bounded if all the eigenvalues of A are simple. Proposition 4.3 implies that every feedback relation of the form (4.3) is continuous. (Again, we normalize ψ_n so that $b_n=\langle\psi_n, b\rangle=1$.) So we may apply Theorem 4.4 to yield immediately the following result.

THEOREM 5.7 (spectral assignability). *Suppose that in the system* (5.1) *the operator* A *as defined in* (5.3) *has only simple eigenvalues. Suppose also that both* b_{12} *and* b_{21} *are nonzero and* $|b_{21}|$ *is sufficiently small so that the conclusion of Theorem* 5.5 *holds. Then if* $\{\mu_j\}_{j=-\infty}^{\infty}$ *is any sequence of complex numbers such that*

$$\sum_{j=-\infty}^{\infty}|\mu_j-\lambda_j|^2<\infty$$

then there exists a feedback of the form $u(t)=\langle\binom{v}{w}(\cdot, T), h\rangle$ *where* h *is a continuous linear functional on* X *such that the set of eigenvalues of the closed loop system is exactly* $\{\mu_j\}_{j=-\infty}^{\infty}$.

6. Case of multiple eigenvalues. The system in § 5 may have multiple eigenvalues. So it seems that the assumption A in § 4 is rather arbitrary for this system. In this section, we will see how the theory we have developed can be modified to cover the case with multiple eigenvalues. We will only state the results and omit the proofs.

One can show that for the system in § 5, all except a finite number of eigenvalues are simple. We will take this last fact as a basic assumption. In other words, we now suppose that λ_n has multiplicity M_n, where $M_n>1$, for only finitely many n.

Let

$X_n=\{x=(\lambda_nI-A)^{M_n}x=0\}$,

$P_n=$ projection of X onto X_n such that $P_nA=AP_n$,

$P_n^*=$ adjoint of P_n (also a projection and $P_n^*A^*=A^*P_n^*$),

$X_n^*=$ image of P_n^*.

We denote by $\underset{\sim}{b}_n$ the unique element in X such that $\langle P_n^* y, b\rangle = \langle y, \underset{\sim}{b}_n\rangle$ for all $y \in X$. Then we see that $\underset{\sim}{b}_n \in \operatorname{Ker}(P_n^*)^\perp$ = closure of range of $P_n = X_n$.

We have the following generalization of Proposition 4.2.

PROPOSITION 6.1 (rank condition). *If the system* (2.1) *is approximately controllable in some time T, then for all n the set* $\{A^{M_n-1}\underset{\sim}{b}_n, A^{M_n-2}\underset{\sim}{b}_n, \ldots, \underset{\sim}{b}_n\}$ *spans* X_n.

Also, we have the following generalization of Theorem 4.4.

THEOREM 6.2. *Suppose the system* (2.1) *is approximately controllable in some time T and the eigenvalue* λ_n *has multiplicity* M_n, $-\infty < n < \infty$ *with* $M_n > 1$ *for only finitely many n. Suppose also that* $\{b_{j,n}\}_{j=1}^{M_n-1}{}_{n=-\infty}^{\infty}$ *is uniformly bounded,* $\{1/\lambda_n\} \in l^2$ *and every h of the form in* (6.4) *represents a continuous linear functional on X. Denoting* $\beta_n = \sum_{m\neq n} 1/|\lambda_m - \lambda_n|^2$, *if* $\{\mu_{j,n}\}_{j=0}^{M_n-1}{}_{n=-\infty}^{\infty}$ *is a sequence of complex numbers such that*

$$\sum_{n=-\infty}^{\infty} \sum_{j=0}^{M_n-1} \frac{(1+\beta_n)|\mu_{j,n} - \lambda_n|^2}{|b_{j,n}|^2} < \infty,$$

then there exists $h \in X$ *such that* $\sigma(A_h) = \{\mu_{j,n}\}_{j=0}^{M_n-1}{}_{n=-\infty}^{\infty}$.

REFERENCES

[1] R. BELLMAN AND K. L. COOKE, *Differential-Difference Equations*, Academic Press, New York, 1963.

[2] L. CARLESON, *Interpolations by bounded analytic functions and the corona problem*, Ann. Math., 76 (1962), pp. 547-559.

[3] D. 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.

[4] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics*, Vol. II, Wiley-Interscience, New York, 1962.

[5] P. L. DUREN, *Theory of H^P Spaces*, Academic Press, New York, 1970.

[6] F. R. GANTMACHER, *The Theory of Matrices*, Vol. I, Chelsea, New York, 1959.

[7] I. C. GOHBERG AND M. G. KREĬN, *Introduction to the Theory of Linear Nonselfadjoint Operators*, American Mathematical Society, Providence, RI, 1969.

[8] L. F. HO, *Controllability and spectral assignability of a class of hyperbolic control systems with retarded control canonical forms*, Thesis, Univ. Wisconsin, Madison, August 1981.

[9] L. F. HO AND D. L. RUSSELL, *Admissible input elements for systems in Hilbert space and a Carleson measure criterion*, this Journal, 21 (1983), pp. 614-640.

[10] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin-Heidelberg, 1976.

[11] 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.

[12] S. H. SUN, *On spectrum distribution of completely controllable linear systemss*, Acta Mathematica Sinica, 21 (1978), pp. 193-205 (in Chinese). English translation, this Journal, 19 (1981), pp. 730-743.

[13] R. G. TEGLAS, *A control canonical form for a class of linear hyperbolic systems*, Thesis, Univ. Wisconsin, Madison, June 1981.

[14] E. C. TITCHMARSH, *The Theory of Functions*, Oxford Univ. Press, Oxford, 1939.

[15] A. C. ZAANAN, *Linear Analysis*, North-Holland, Amsterdam, 1953.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 6, November 1986

009

THE REGULAR LOCAL NONINTERACTING CONTROL PROBLEM FOR NONLINEAR CONTROL SYSTEMS*

H. NIJMEIJER† AND J. M. SCHUMACHER‡

Abstract. We study the Noninteracting Control Problem for affine nonlinear control systems under the assumption that the number of scalar inputs equals the number of vector outputs. Our purpose is to find a static state feedback law for the system which achieves noninteraction. Using the recently developed differential geometric approach to nonlinear systems theory and working under a set of regularity assumptions, we give necessary and sufficient conditions for the local solvability of the problem. This work extends earlier results in the "geometric approach" for linear systems.

Key words. nonlinear control systems, noninteracting control, controlled invariance, controllability distributions

AMS(MOS) subject classifications. 93C10, 49E05

1. Introduction. This paper is intended as a contribution to the theory of noninteracting control of nonlinear systems. In general terms, the problem can be described as follows. Suppose that a dynamical system has been given, in which two sets of variables have been designated as instruments and as targets, respectively. The targets and instruments may be either scalar variables or vectors. One says that we have a situation of *noninteraction* (or *input-output decoupling*) if each instrument affects only one target and none of the others. If the given system does not have this property, one may ask whether it is possible to add control loops to the system in such a way, that noninteraction is obtained. This is the problem of noninteracting control.

To arrive at a precise problem statement, one has to specify: (i) the class of systems under study, (ii) the precise nature of the noninteraction one wants to obtain, and (iii) the control format. In this paper, we shall consider "affine" [13], [18], [22] systems, which constitute a class of nonlinear systems that has received considerable attention. We will assume that the input variables ("instruments") are scalars, but we allow the output variables ("targets") to be vectors. The condition of noninteraction will be defined using the concepts of "controllability distributions" [11], [14] and "output controllability" [16], [17]. The control schemes we shall consider consist of locally defined state feedback and (state-dependent) precompensation. This combination is often referred to in the literature simply as "static state feedback" [8], [18] or even just "feedback". Definitions of the concepts mentioned here will be given below.

The noninteracting control problem has been studied extensively and from various points of view. Most of the literature is concerned with linear systems. In this field, one has the option of dealing with the problem via the transfer function, and this approach has been taken in some of the earliest work in noninteraction [5], [10] as well as in recent contributions [6]. Within the state-space framework, a breakthrough was made around 1970 ([24]; see also [1]). The innovation centered around the introduction of the notion of "controllability subspace" as a means of expressing the intuitive notion of "subsystem", which is of obvious importance in the theory of noninteraction. Controllability subspaces came to play a key role in a successful line of research that has been termed the "geometric approach" to linear systems theory.

* Received by the editors May 17, 1983, and in revised form July 10, 1985.

† Department of Applied Mathematics, Twente University of Technology, P.O. Box 217, 7500 AE Enschede, the Netherlands.

‡ Centre for Mathematics and Computer Science, Kruislaan 413, 1098 SJ Amsterdam, the Netherlands.

For more detailed accounts of the long history of noninteracting control of linear systems, we refer to [12] and [6]. In the nonlinear domain, progress has been slower. "Although efforts have been made to develop a decoupling theory for nonlinear systems, (...) considerable difficulties (...) have so far inhibited substantial progress", wrote the authors of [12] in 1971. A criterion for achievability of noninteraction by "static state feedback" was given by Sinha [20] for the case of scalar inputs and scalar outputs. Further work under the same restriction was reported in [3].

In the award winning paper [8] Isidori et al. were the first authors to give a geometric formulation for the general noninteracting control problem. They presented a solution to this problem [8, Thm. 5.1]; however, their necessary and sufficient conditions for solvability are, in the general case (allowing for vector outputs), not constructive (see [8, Thm. 5.1]). The purpose of this paper is to give constructive (verifiable) necessary and sufficient conditions also in the case of vector outputs, pertaining to the solvability of what we will call the *regular local noninteracting control problem.* In contrast to the results of [20], [3], [2] and also [8] we will use a nonlinear version of the concept of controllability subspaces. This extension was made in [11] and [14], leading to various definitions for the so-called "controllability distributions". The concept was applied to solve special versions of the decoupling problem in [16] and [17]. In this paper, we will show that controllability distributions can be used to rederive a major result from the linear theory [24] in a nonlinear context. This continues a line of research [7]–[9], [11], [13]–[18] that is directed towards a systematic generalization of the "geometric approach" [23] to nonlinear systems, using the methods of differential geometry.

The organization of the paper is as follows. In § 2, the precise formulation is given of the decoupling problem that we consider here. The main result follows in § 3, which is largely devoted to lemmas that are needed in the sufficiency part of the proof. Some remarks on the structure of a decoupled system are given in § 4.

2. Problem formulation. Consider the affine nonlinear control system

$$\dot{x}(t) = A(x(t)) + \sum_{i=1}^{m} B_i(x(t))u_i(t) \tag{2.1}$$

where x are local coordinates of a smooth n-dimensional manifold M, A, $B_1, \cdots, B_m$ are smooth vector fields on M and $u_i : \mathbb{R}_+ \to \mathbb{R}$ is a piecewise smooth input function, $i \in \underline{m}$. Together with the dynamics (2.1) we consider as output functions

$$z_i(t) = C_i(x(t)), \qquad i \in \underline{m} \tag{2.2}$$

where $C_i : M \to N_i$ is a smooth map from M to a smooth p_i-dimensional manifold N_i, $p_i \geqq 1$, $i \in \underline{m}$. We assume that each C_i, $i \in \underline{m}$, is a surjective submersion. We will observe later on that the output functions $C_i : M \to N_i$ play no role beyond specification of the distributions $\operatorname{Ker} C_{i*}$, $i \in \underline{m}$. To rule out obvious unsolvable problems, we will assume that the maps C_i are mutually independent, i.e. the rank of the map $C = (C_1, \cdots, C_m) : M \to N_1 \times \cdots \times N_m$ equals $p_1 + \cdots + p_m$. Also we will assume—as is usual in the differential geometric approach—that the distribution generated by the input vector fields has no singularities, so

$$\dim \Delta_0 := \dim \{B_1, \cdots, B_m\} = m.$$

Furthermore we will assume that for the system (2.1) the *accessibility* distribution, see [14], equals TM. That is, the system (2.1) is *strongly accessible*, cf. [21], [22], i.e. the set of reachable points at time $t > 0$ from $x_0 \in M$ has a nonempty interior in M. In this

paper we allow for static state feedback. Thus an admissible control law is of the form

$$u = \alpha(x) + \beta(x)v, \tag{2.3}$$

where $\alpha : M \to \mathbb{R}^m$, $\beta : M \to \mathbb{R}^{m \times m}$ are smooth maps, and $v = (v_1, \cdots, v_m)' \in \mathbb{R}^m$ represents a new input. To keep as much open-loop control as possible, we seek β such that $\beta(x) = (\beta_{ij}(x))$ is nonsingular for all $x \in M$. Applying the feedback law (2.3) to (2.1), we obtain as the new dynamics

$$\dot{x}(t) = \tilde{A}(x(t)) + \sum_{i=1}^{m} \tilde{B}_i(x(t))v_i(t), \tag{2.4}$$

where

$$\tilde{A}(x) = A(x) + \sum_{i=1}^{m} B_i(x)\alpha_i(x), \tag{2.5a}$$

$$\tilde{B}_i(x) = \sum_{j=1}^{m} B_j(x)\beta_{ji}(x), \qquad i \in \underline{m}. \tag{2.5b}$$

In the *static state feedback noninteracting control problem* we seek a control law (2.3) such that in the modified dynamics the control $v_i(\cdot)$ does not affect the outputs $z_j(\cdot)$ for $j \neq i$; moreover we want the scalar input $v_i(\cdot)$ to "control" the corresponding (vector-valued) output $z_i(\cdot)$, $i \in \underline{m}$. This problem can be nicely formulated in a differential geometric framework, see also [8]. For doing so we need some terminology. Consider the set $V(M)$ of all smooth vector fields on M as a Lie algebra with Lie product $[X_1, X_2]$ for $X_1, X_2 \in V(M)$. For any set of vector fields $S \subset V(M)$ we denote by $\{S\}_{LA}$ the Lie subalgebra generated by S. Furthermore for $X_1, X_2 \in V(M)$ define $ad^0_{X_1}X_2 = X_2$, $ad^1_{X_1}X_2 = [X_1, X_2]$ and recursively $ad^k_{X_1}X_2 = [X_1, ad^{k-1}_{X_1}]$, $k = 2, 3, \cdots$. Associated with the system (2.4) we define the following Lie algebras, see also [19]:

$$L_0 = \{ad^k_{\tilde{A}}\tilde{B}_j,\ k \in \mathbb{Z}_+, j \in \underline{m}\}_{LA}, \tag{2.6a}$$

$$L_{0i} = \{ad^k_{\tilde{A}}\tilde{B}_i,\ k \in \mathbb{Z}_+\}_{LA}, \qquad i \in \underline{m}, \tag{2.6b}$$

and the Lie ideal generated by L_{0i} in L_0 which will be denoted by $\hat{L}_{0i}$, $i \in \underline{m}$. Clearly $L_{0i} \subset \hat{L}_{0i}$, $i \in \underline{m}$. Also we introduce the corresponding distributions

$$R_0 = \text{Span}\,\{L_0\}, \qquad R_i = \text{Span}\,\{L_{0i}\}, \quad \hat{R}_i = \text{Span}\,\{\hat{L}_{0i}\}, \quad i \in \underline{m}. \tag{2.7}$$

To take care that in the new dynamics (2.4) the input $v_i(\cdot)$ has no interaction with $z_j(\cdot)$, $j \neq i$, we must have

$$\hat{R}_i \subset \bigcap_{j \neq i} \text{Ker}\, C_{j*}, \qquad i \in \underline{m}. \tag{2.8}$$

To achieve that $v_i(\cdot)$ "controls" the corresponding output $z_i(\cdot)$, $i \in \underline{m}$, we need the nonlinear version of *output controllability*, that is

$$C_{i*}R_i = TN_i, \qquad i \in \underline{m} \tag{2.9}$$

which is equivalent to the fact that the set of reachable output values has nonempty interior in N_i, $i \in \underline{m}$, see [16], [17]. Because $R_i \subset \hat{R}_i$, we see that (2.9) implies

$$C_{i*}\hat{R}_i = TN_i, \qquad i \in \underline{m}. \tag{2.10}$$

There is a nice and in our context useful interpretation of (2.8), (2.9) and (2.10) in geometric terms. Recall the following definitions, see [7], [9], [13], [18]. An involutive

distribution D on M is called *controlled invariant* if there exists a feedback (2.3) such that the modified dynamics (2.4)-(2.5) satisfies

$$\begin{aligned} &[\tilde{A}, D] \subset D, \\ &[\tilde{B}_i, D] \subset D, \qquad i \in \underline{m}. \end{aligned} \tag{2.11}$$

An involutive distribution D on M is called a (*degenerate*) *controllability distribution* if there exists a feedback (2.3) and a subset $I \subset \underline{m}$ such that

$$D = \text{Span}\,(\{ad_{\tilde{A}}^k \tilde{B}_i \mid i \in I, k \in \mathbb{Z}_+\}_{LA}). \tag{2.12}$$

Finally an involutive distribution D on M is called a *regular controllability distribution* if there exists a feedback (2.3) and a subset $I \subset \underline{m}$ such that

$$\begin{aligned} &[\tilde{A}, D] \subset D, \\ &[\tilde{B}_i, D] \subset D, \qquad i \in \underline{m} \end{aligned} \tag{2.13}$$

and

$$D = \text{Span}\,(\{ad_{\tilde{A}}^k \tilde{B}_i, ad_{\tilde{B}_j}^k \tilde{B}_i \mid i \in I, k \in \mathbb{Z}_+\}_{LA}). \tag{2.14}$$

Returning to the noninteracting control problem, we see, by definition of the distributions $\hat{R}_i$, $i \in \underline{m}$, that for all $i \in \underline{m}$

$$\begin{aligned} &[\tilde{A}, \hat{R}_i] \subset \hat{R}_i, \\ &[\tilde{B}_j, \hat{R}_i] \subset \hat{R}_i, \qquad j \in \underline{m}. \end{aligned} \tag{2.15}$$

That is, $\hat{R}_i$ is controlled invariant and moreover it is a regular controllability distribution, $i \in \underline{m}$, whereas R_i is a degenerate controllability distribution, $i \in m$.

In this way the static state feedback noninteracting control problem can be stated as follows:

> Given the system (2.1)-(2.2), find, if possible, a feedback law (2.3) such that the distributions $\hat{R}_i$ defined by (2.6)-(2.7) satisfy (2.8) and (2.10).

In this paper we will solve a *regular local version* of this problem, to be called the *regular local noninteracting control problem.* In this context "local" means, that given an arbitrary initial state $x_0 \in M$, we are interested in the existence of a local feedback (2.3), i.e. the maps α and β in (2.3) are only well defined on a neighborhood of x_0. Working locally, we are able to fully exploit the differential geometric approach set up in [7], [8] and worked out in a series of papers [9], [13]-[18], [22], [25], [26]. It is a logical first step, and a common practice in the literature just cited, to exclude singularities. In this paper, too, we shall work under a series of regularity assumptions. In particular, we look for a specific set of "regular" distributions R_i^*, $i \in \underline{m}$, that satisfy the conditions (2.8) and (2.10). The exact definition of the word "regular" will be given in § 3; among its implications is that we demand that the distributions $\cup_{i \in I} R_i^*$ and $(\cup_{i \in I} R_i^*) \cap \Delta_0$ have constant dimension for all $I \subset \underline{m}$. This surely is a restrictive assumption, but we feel that it is necessary to complete the "regular" analysis before one can hope to successfully attack the singular cases. Moreover, one should realize that for analytic systems (2.1)-(2.2) the regularity assumptions will hold on an open and dense submanifold M' of M. Therefore, a (global) solution to the general noninteracting control problem has to satisfy the necessary and sufficient conditions of the next section on this submanifold M'.

3. Necessary and sufficient conditions for the regular local noninteracting control problem. We now come to the necessary and sufficient conditions for the solvability of the regular local noninteracting control problem. For this we need some notation and assumptions.

A1. The distribution $\Delta_0 = \text{Span}\{B_1, \cdots, B_m\}$ has dimension m on M.

A2. The strong accessibility distribution of (2.1), $R_0 = \text{Span}\{ad_A^k B_i \mid k \in \mathbb{Z}_+\}_{LA}$ has dimension n.

A3. The output maps (2.2) are mutually independent, i.e. the rank of the map $C = (C_1, \cdots, C_m): M \to N_1 \times \cdots \times N_m$ equals $p_1 + \cdots + p_m$.

For each subset $I \subset \underline{m}$ let us define R_I^* as the maximal regular local controllability distribution contained in $\bigcap_{j \notin I} \text{Ker } C_{j*}$. The involutive distributions R_I^*, $I \subset \underline{m}$, are well defined (see [11], [14]) but their dimension may vary on M. (We set $R_{\underline{m}}^* = TM$, as a consequence of the rule from logic that the empty intersection of parts is the whole.)

A4. The distributions R_I^*, $\sum_{i \in I} R_i^*$, $(\sum_{i \in I} R_i^*) \cap \Delta_0$, $I \subset \underline{m}$, as well as $\overline{\sum_{i \in I} R_i^*}$ and $\overline{\sum_{i \in I} R_i^* \cap \Delta_0}$, $I \subset \underline{m}$, all have constant dimension on M (here the bar denotes involutive closure).

THEOREM 3.1. *Consider the system* (2.1)-(2.2) *and assume* A1-A4 *hold. Then the regular local static state feedback noninteracting control problem is solvable if and only if*

$$\Delta_0 = \Delta_0 \cap R_1^* + \cdots + \Delta_0 \cap R_m^*. \tag{3.1}$$

Furthermore, if these conditions hold, then $\{R_i^*\}_{i=1}^m$ *is the only solution of the noninteracting control problem.*

Before we are able to prove this theorem, we need some preliminary results on the distributions R_I^*, $I \subset \underline{m}$. The following lemmas are also of independent interest, and give, even in the linear case, additional information on the structure of the distributions R_I^*, $I \subset \underline{m}$. Everywhere below, we shall consider the system (2.1)-(2.2), assuming that A1-A4 hold.

LEMMA 3.2. *Suppose that* (3.1) *holds. Then, for all* $i \in \underline{m}$,

$$\Delta_0 \cap R_i^* \not\subset \sum_{j \neq i} R_j^*. \tag{3.2}$$

Proof. Suppose that (3.1) were true and that (3.2) would not hold. It would then follow that

$$\Delta_0 \subset \sum_{j \neq i} R_j^* \subset \overline{\sum_{j \neq i} R_j^*}. \tag{3.3}$$

Note that the last distribution in (3.3) is controlled invariant, since it is the sum of controlled invariant distributions. Under the condition of strong accessibility, any controlled invariant distribution containing Δ_0 must be equal to TM (see [22]). Since it is clear from the definition of the distributions R_i^* that

$$\sum_{j \neq i} R_j^* \subset \overline{\sum_{j \neq i} R_j^*} \subset \text{Ker } C_{i*}, \tag{3.4}$$

it would follow that $\text{Ker } C_{i*} = TM$, or the map C_i is trivial. This contradicts our assumptions. □

We can use the above lemma in some counting arguments. Set $\gamma_1 = \dim(\Delta_0 \cap R_1^*)$, and define

$$\gamma_k = \dim \sum_{i=1}^{k} (\Delta_0 \cap R_i^*) - \dim \sum_{i=1}^{k-1} (\Delta_0 \cap R_i^*) \tag{3.5}$$

for $k=2, \cdots, m$. Note that the γ_k's are constant by A4. It is obviously true that $\gamma_k \geqq 0$ for all k, but if (3.1) holds then it follows from the above lemma that even $\gamma_k \geqq 1$ for $k \in \underline{m}$. For, suppose that $\gamma_i = 0$ for some i; then we would have

$$\Delta_0 \cap R_i^* \subset \sum_{j=1}^{i-1} (\Delta_0 \cap R_j^*) \subset \sum_{j \neq i} (\Delta_0 \cap R_j^*), \tag{3.6}$$

a situation which has been excluded by Lemma 3.2. On the other hand, still under the assumption that (3.1) holds, it is clear that

$$\sum_{k=1}^{m} \gamma_k = \dim \sum_{i=1}^{m} (\Delta_0 \cap R_i^*) = \dim \Delta_0 = m. \tag{3.7}$$

So the γ_k's form a set of m integers $\geqq 1$ which add up to m. It follows, of course, that $\gamma_k = 1$ for all k. As a consequence,

$$\dim \sum_{i=1}^{j} (\Delta_0 \cap R_i^*) = \sum_{i=1}^{j} \gamma_j = j \tag{3.8}$$

for all $j \in \underline{m}$. Since the order in which we numbered the output functions is arbitrary, the conclusion can be formulated as follows.

LEMMA 3.3. *Suppose that* (3.1) *holds. Then, for all* $I \subset \underline{m}$,

$$\dim \sum_{i \in I} (\Delta_0 \cap R_i^*) = |I|. \tag{3.9}$$

An obvious consequence of this is that the distributions $\Delta_0 \cap R_i^*$ are independent. Now, set $\theta_1 = \dim (\Delta_0 \cap R_1^*)$ and define the constants (see A4)

$$\theta_k = \dim \left(\Delta_0 \cap \sum_{i=1}^{k} R_i^* \right) - \dim \left(\Delta_0 \cap \sum_{i=1}^{k-1} R_i^* \right) \tag{3.10}$$

for $k=2, \cdots, m$. If we reason in the same way as above, now using the fact that $\sum_{i=1}^{m} R_i^* = TM$, then the conclusion we obtain is the following.

LEMMA 3.4. *Suppose that* (3.1) *holds. Then, for all* $I \subset \underline{m}$,

$$\dim \left(\Delta_0 \cap \sum_{i \in I} R_i^* \right) = |I|. \tag{3.11}$$

Since it is clear that

$$\sum_{i \in I} (\Delta_0 \cap R_i^*) \subset \Delta_0 \cap \sum_{i \in I} R_i^* \tag{3.12}$$

for all $I \subset \underline{m}$, we obtain as a corollary:

LEMMA 3.5. *Suppose that* (3.1) *holds. Then, for all* $I \subset \underline{m}$,

$$\Delta_0 \cap \sum_{i \in I} R_i^* = \sum_{i \in I} (\Delta_0 \cap R_i^*). \tag{3.13}$$

Set $\rho_1 = \dim (R_1^* \cap \Delta_0)$, and define

$$\rho_k = \dim (R_{\underline{k}}^* \cap \Delta_0) - \dim (R_{\underline{k-1}}^* \cap \Delta_0) \tag{3.14}$$

for $k=2, \cdots, m$. These numbers are constants by Assumption A4. We can use them to prove:

LEMMA 3.6. *Suppose that* (3.1) *holds. Then, for all* $I \subset \underline{m}$,

$$\dim (\Delta_0 \cap R_I^*) = |I|. \tag{3.15}$$

Proof. It is sufficient to show that $\rho_k \neq 0$ for all $k \in \underline{m}$. So, suppose that $\rho_k = 0$ for some $k \in \underline{m}$. Then we have

$$R_{\underline{k}}^* \cap \Delta_0 = R_{\underline{k-1}}^* \cap \Delta_0. \tag{3.16}$$

The distributions $R_{\underline{k-1}}^*$ and $R_{\underline{k}}^*$ are both regular local controllability distributions, and, of course, $R_{\underline{k-1}}^* \subset R_{\underline{k}}^*$. It follows from [17] that there exists a feedback $u = \alpha(x) + \beta(x)v$ for the system (2.1) such that both $R_{\underline{k}}^*$ and $R_{\underline{k-1}}^*$ are invariant for the modified dynamics. It then follows from the characterization of regular local controllability distributions given in [11] (Lemma 4.1) that

$$R_{\underline{k}}^* = R_{\underline{k-1}}^*. \tag{3.17}$$

Since $R_k^* \subset R_{\underline{k}}^*$ and $R_{\underline{k-1}}^* \subset \ker C_{k*}$, we obtain

$$R_k^* \subset \operatorname{Ker} C_{k*}. \tag{3.18}$$

Since we also have, by definition,

$$\sum_{i \neq k} R_k^* \subset \operatorname{Ker} C_{k*} \tag{3.19}$$

it follows from the strong accessibility condition that the map C_k is trivial, and we have reached a contradiction. □

We can use this to establish the following lemma, which will be instrumental in proving that, under the assumption (3.1), the distribution $\sum_{i \in I} R_i^*$ is involutive.

LEMMA 3.7. *Suppose that* (3.1) *holds. Then, for any* $I \subset \underline{m}$,

$$R_I^* = \bar{R}_I := \text{inv. clos.} \left[\sum_{i \in I} R_i^* \right]. \tag{3.20}$$

Proof. Let $I \subset \underline{m}$. Since R_I^* is involutive, and $R_i^* \subset R_I^*$ for all $i \in I$, we have

$$\sum_{i \in I} R_i^* \subset \bar{R}_I \subset R_I^*. \tag{3.21}$$

By Lemmas 3.4 and 3.6, it follows that

$$\Delta_0 \cap \bar{R}_I = \Delta_0 \cap R_I^*. \tag{3.22}$$

From [14], we know that $\bar{R}_I$ is a regular local controllability distribution. By the same argument as was used in the proof of Lemma 3.6, it follows that (3.20) holds. □

For linear systems, the above result already implies that, under the given circumstances, the subspace corresponding to the distribution $\sum_{i \in I} R_i^*$ is equal to the maximal controllability subspace contained in the intersection of the kernels of the output mappings C_i ($i \in \underline{m} \backslash I$)—a result which doesn't seem to have been formulated explicitly before. In the nonlinear context, we have to worry about involutiveness. Somewhat surprisingly, it turns out that this is not a problem. The key lemma is the following.

LEMMA 3.8. *Let* R_1 *and* R_2 *be regular local controllability distributions such that* $\overline{R_1 + R_2}$ *and* $\Delta_0 \cap \overline{R_1 + R_2}$ *have fixed dimension, and*

$$\Delta_0 \cap \overline{R_1 + R_2} = \Delta_0 \cap R_1 + \Delta_0 \cap R_2. \tag{3.23}$$

Then $R_1 + R_2$ *is involutive* (*and hence—see* [14]—$R_1 + R_2$ *is itself a regular local controllability distribution*).

Proof. Locally, we can choose sets of independent vector fields $\{B_{11}, \cdots, B_{1k}\}$ and $\{B_{21}, \cdots, B_{2l}\}$ such that

$$\Delta_0 \cap R_1 = \text{Span}\{B_{11}, \cdots, B_{1k}\}, \tag{3.24}$$

$$\Delta_0 \cap R_2 = \text{Span}\{B_{21}, \cdots, B_{2l}\}. \tag{3.25}$$

Note that

$$[B_{1i}, R_2] \subset (R_2 + \Delta_0) \cap \overline{R_1 + R_2} = R_2 + (\Delta_0 \cap \overline{R_1 + R_2}) = R_2 + \text{Span}\{B_{11}, \cdots, B_{1k}\} \tag{3.26}$$

for all $i \in \underline{k}$. Using [13], we see that there exist vector fields $\tilde{B}_{11}, \cdots, \tilde{B}_{1k}$ such that

$$\text{Span}\{\tilde{B}_{11}, \cdots, \tilde{B}_{1k}\} = \text{Span}\{B_{11}, \cdots, B_{1k}\} = \Delta_0 \cap R_1, \tag{3.27}$$

$$[\tilde{B}_{1i}, R_2] \subset R_2, \qquad i \in \underline{k}. \tag{3.28}$$

Likewise, one can find vector fields $\tilde{B}_{21}, \cdots, \tilde{B}_{2l}$ such that

$$\text{Span}\{\tilde{B}_{21}, \cdots, \tilde{B}_{2l}\} = \text{Span}\{B_{21}, \cdots, B_{2l}\} = \Delta_0 \cap R_2, \tag{3.29}$$

$$[\tilde{B}_{2i}, R_1] \subset R_1, \qquad i \in \underline{l}. \tag{3.30}$$

Since $\overline{R_1 + R_2}$ is controlled invariant (see [14]) and $R_1 \subset \overline{R_1 + R_2}$, we can find, as in [17], a closed-loop mapping $\tilde{A}$ such that

$$[\tilde{A}, R_1] \subset R_1, \tag{3.31}$$

$$[\tilde{A}, \overline{R_1 + R_2}] \subset \overline{R_1 + R_2}. \tag{3.32}$$

Now, note that

$$[\tilde{A}, R_2] \subset (R_2 + \Delta_0) \cap (\overline{R_1 + R_2}) = R_2 + \text{Span}\{\tilde{B}_{11}, \cdots, \tilde{B}_{1k}\}. \tag{3.33}$$

It follows from (3.33) and (3.28) (see [7]) that there exist functions γ_i ($i \in \underline{k}$) such that $\bar{A} = \tilde{A} + \sum_{i=1}^{k} \tilde{B}_{1i}\gamma_i$ satisfies

$$[\bar{A}, R_2] \subset R_2. \tag{3.34}$$

Using the general formula

$$[X \cdot c, Y] = [X, Y] \cdot c - X \cdot Y(c), \tag{3.35}$$

we find that also

$$[\bar{A}, R_1] \subset R_1, \tag{3.36}$$

$$[\bar{A}, \overline{R_1 + R_2}] \subset \overline{R_1 + R_2}. \tag{3.37}$$

By the characterization of regular local controllability distributions in [14], we know that, if we define

$$\hat{R}_i = \text{Span}\{ad_{\bar{A}}^j \tilde{B}_{is} \mid s \in \underline{k}, j \in \mathbb{Z}_+\}, \qquad i = 1, 2, \tag{3.38}$$

then

$$R_i = \text{inv.clos. } \hat{R}_i. \tag{3.39}$$

Using the Jacobi identity, one can easily prove by induction that

$$[ad_{\bar{A}}^j \tilde{B}_{1i}, R_2] \subset R_2 \tag{3.40}$$

for all $j \in \mathbb{Z}_+$ and $i \in \underline{k}$. Using (3.35) again, we find that

$$[\hat{R}_1, R_2] \subset R_1 + R_2. \tag{3.41}$$

Another induction argument based on the Jacobi identity then shows that

$$[R_1, R_2] \subset R_1 + R_2. \tag{3.42}$$

Hence, $R_1 + R_2$ is involutive. □

The result that we were after is now an easy consequence.

LEMMA 3.9. *Suppose that* (3.1) *holds. Then, for any subset* $I \subset \underline{m}$, *the distribution* $\sum_{i \in I} R_i^*$ *is involutive.*

Proof. It is sufficient to show that $\sum_{i=1}^{k} R_i^*$ is involutive for every $k \in \underline{m}$, and we do this by induction. Of course, R_1^* is involutive by definition. Now assume that $\sum_{i=1}^{k-1} R_i^*$ is involutive. Then $\sum_{i=1}^{k-1} R_i^* = R_{k-1}^*$ according to Lemma 3.7, and so it follows that $\sum_{i=1}^{k-1} R_i^*$ is a regular local controllability distribution. Moreover, the results of Lemmas 3.7, 3.6 and 3.5 show that

$$\Delta_0 \cap \text{inv.clos.} \left(\sum_{i=1}^{k} R_i^* \right) = \Delta_0 \cap \left(\sum_{i=1}^{k-1} R_i^* \right) + \Delta_0 \cap R_k^*. \tag{3.43}$$

An application of the preceding lemma now completes the proof. □

One notes that it is now immediate that, under the assumption (3.1), we have

$$R_I^* = \sum_{i \in I} R_i^* \tag{3.44}$$

just as in the linear case.

Now, let us proceed to an argument that will be directly needed in the proof of the main theorem. The issue is the "compatibility" of the distributions R_i^*.

LEMMA 3.10. *Suppose that* (3.1) *holds. Then, locally, there exists a basis of vector fields* $\{\tilde{B}_1, \cdots, \tilde{B}_m\}$ *for* Δ_0 *such that*

$$\text{Span}\{\tilde{B}_i\} = \Delta_0 \cap R_i^*, \qquad i \in \underline{m}, \tag{3.45}$$

$$[\tilde{B}_i^*, R_j^*] \subset R_j^*, \qquad i, j \in \underline{m}. \tag{3.46}$$

Proof. Take $i \in \underline{m}$, and let $\hat{B}_i$ be a vector field such that

$$\text{Span}\{\hat{B}_i\} = \Delta_0 \cap R_i^*. \tag{3.47}$$

From the fact that the R_j^*'s are controlled invariant, it follows that

$$[\hat{B}_i, R_j^*] \subset R_j^* + \Delta_0, \qquad j \in \underline{m}. \tag{3.48}$$

Since $\hat{B}_i \in R_i^*$ and $R_i^* + R_j^*$ is involutive (Lemma 3.9), we also have

$$[\hat{B}_i, R_j^*] \subset R_i^* + R_j^*, \qquad j \in \underline{m}. \tag{3.49}$$

Consequently, we have (using Lemma 3.5)

$$[\hat{B}_i, R_j^*] \subset (R_j^* + \Delta_0) \cap (R_i^* + R_j^*) = R_j^* + \Delta_0 \cap (R_i^* + R_j^*) = R_j^* + \text{Span}\{\hat{B}_i\}. \tag{3.50}$$

It is also clear that

$$\left[\hat{B}_i, \sum_{j \neq i} R_j^* \right] \subset \sum_{j \neq i} R_j^* + \text{Span}\{\hat{B}_i\}. \tag{3.51}$$

From this, it follows (see [13]) that there exists a function β_i such that $\tilde{B}_i := \hat{B}_i \beta_i$ satisfies

$$\left[\tilde{B}_i, \sum_{j \neq i} R_j^* \right] \subset \sum_{j \neq i} R_j^*. \tag{3.52}$$

But then we also have (from (3.50) and (3.52); cf. also [18]), for $i \neq j$

$$[\tilde{B}_i, R_j^*] \subset \sum_{j \neq i} R_j^* \cap (R_j^* + \text{Span}\{\tilde{B}_i\}) = R_j^* + \left(\sum_{j \neq i} R_j^* \cap \Delta_0 \cap R_i^* \right) = R_j^*. \tag{3.53}$$

Of course, the equality (3.53) also holds for $i=j$ since R_j^* is involutive. Going through this construction for each $i \in \underline{m}$, we obtain a set of vector fields $\{\tilde{B}_1, \cdots, \tilde{B}_m\}$ which satisfies (3.46), and which is then automatically a basis for Δ_0. □

We now proceed to the proof of the main theorem.

Proof (of Theorem 3.1). For sufficiency, we assume that (3.1) holds. First of all, note that $\sum_{i=1}^m R_i^*$ contains Δ_0 and is controlled invariant for the system (2.1), so that, as noted in the proof of Lemma 3.2,

$$\sum_{i=1}^{m} R_i^* = TM. \tag{3.54}$$

Since we have, for each i,

$$\sum_{j \neq i} R_j^* \subset \operatorname{Ker} C_{i*}, \tag{3.55}$$

it follows that

$$R_i^* + \operatorname{Ker} C_{i*} = TM \qquad (i \in \underline{m}) \tag{3.56}$$

i.e., (2.14) is satisfied.

Next, we have to show that there exists a local feedback of the form $u = \alpha(x) + \beta(x)v$ that leaves each of the distributions R_i^* $(i \in \underline{m})$ invariant; i.e., the R_i^*'s are "compatible". An appropriate $\beta := \operatorname{diag}(\beta_1, \cdots, \beta_m)$ was already shown to exist in Lemma 3.10. Vector fields $\tilde{B}_1, \cdots, \tilde{B}_m$ can be constructed that satisfy $[\tilde{B}_i, R_j^*] \subset R_j^*$ $(i, j \in \underline{m})$, and there exists a unique nonsingular map $\beta : M \to \mathbb{R}^{m \times m}$ (locally defined) such that $\tilde{B}_i(x) = \hat{B}_i(x)\beta_i(x)$, where the vector fields $\hat{B}_i$ have been selected to satisfy (3.47); clearly, these vector fields are obtained from a nonsingular transformation of the original input vector fields appearing in (2.1).

We have shown in Lemma 3.9 that $\sum_{j \neq i} R_j^*$ is locally controlled invariant, so

$$\left[A, \sum_{j \neq i} R_j^*\right] \subset \sum_{j \neq i} R_j^* + \Delta_0 = \sum_{j \neq i} R_j^* + \operatorname{Span}\{\hat{B}_i\}. \tag{3.57}$$

Also, we have

$$\left[\tilde{B}_i, \sum_{j \neq i} R_j^*\right] \subset \sum_{j \neq i} R_j^*. \tag{3.58}$$

According to [7], it follows from this that we can, locally, define a function α_i such that the vector field $A + \tilde{B}_i\alpha_i$ leaves $\sum_{j \neq i} R_j^*$ invariant:

$$\left[A + \tilde{B}_i\alpha_i, \sum_{j \neq i} R_j^*\right] \subset \sum_{j \neq i} R_j^*. \tag{3.59}$$

Having done this for each $i \in \underline{m}$, we next consider the vector field $A + \sum_{j=1}^m \tilde{B}_j\alpha_j$. The following holds:

$$\begin{aligned} \left[A + \sum_{j=1}^{m} \tilde{B}_j\alpha_j, \sum_{j \neq i} R_j^*\right] &\subset \left[A + \tilde{B}_i\alpha_i, \sum_{j \neq i} R_j^*\right] + \left[\sum_{j \neq i} \tilde{B}_j\alpha_j, \sum_{j \neq i} R_j^*\right] \\ &\subset \sum_{j \neq i} R_j^* + \sum_{j \neq i} \left[\tilde{B}_j\alpha_j, \sum_{j \neq i} R_j^*\right] \subset \sum_{j \neq i} R_j^*. \end{aligned} \tag{3.60}$$

This shows that the distributions $\sum_{j \neq i} R_j^*$ are compatible. Now, define

$$\tilde{R}_i = \bigcap_{j \neq i} \sum_{k \neq j} R_k^*. \tag{3.61}$$

Clearly, we have $R_i^* \subset \tilde{R}_i$ for all $i \in \underline{m}$. It is immediate from (3.60) that

$$\left[A + \sum_{j=1}^{m} \tilde{B}_j \alpha_j, \tilde{R}_i\right] \subset \tilde{R}_i, \qquad i \in \underline{m}. \tag{3.62}$$

Furthermore, one has

$$\tilde{R}_i \cap \Delta_0 = \bigcap_{j \neq i} \sum_{k \neq j} (R_k^* \cap \Delta_0) = R_i^* \cap \Delta_0. \tag{3.63}$$

It now follows from Lemma 4.1 of [11] that R_i^* is the maximal regular local controllability distribution in $\tilde{R}_i$, and that R_i^* is also invariant under $A + \sum_{j=1}^{m} \tilde{B}_j \alpha_j$, i.e.,

$$\left[A + \sum_{j=1}^{m} \tilde{B}_j \alpha_j, R_i^*\right] \subset R_i^*, \qquad i \in \underline{m}. \tag{3.64}$$

With the properties (3.45), (3.46) and (3.64) all fulfilled, we see that the first part of the proof is complete. To show that (3.1) is necessary, let $\{R_i\}_{i \in \underline{m}}$ be a set of regular local controllability distributions that gives a solution of the decoupling problem (see § 2). Since

$$\begin{aligned} \Delta_0 = \operatorname{Span}\{B_1, \cdots, B_m\} = \operatorname{Span}\{\tilde{B}_1, \cdots, \tilde{B}_m\} &\subset \sum_{i=1}^{m} (\Delta_0 \cap R_i) \\ &\subset \sum_{i=1}^{m} (\Delta_0 \cap R_i^*) \subset \Delta_0, \end{aligned} \tag{3.65}$$

we see immediately that (3.1) must hold. Finally, it is also clear that, for any solution of the decoupling problem, we must have

$$\Delta_0 \cap R_i = \Delta_0 \cap R_i^*. \tag{3.66}$$

By the same argument as was used in the proof of Lemma 3.6, it follows from (3.66) that $R_i = R_i^*$. This completes the proof of the theorem. □

Remarks. (i) In connection with the solution of the noninteracting control problem given in [8] we note the following. The distributions $\Delta_i = \sum_{j \neq i} R_j^*$, $i \in \underline{m}$, are locally compatible, i.e. there locally exists a feedback which leaves each of the Δ_i's, $i \in \underline{m}$, invariant. Moreover we have that $\tilde{B}_j \subset \Delta_i \subset \operatorname{Ker} C_{i*}$, $j \neq i$, and for any two disjoint subsets $I, J \subset \underline{m}$ we have $(\bigcap_{i \in I} \Delta_i) + (\bigcap_{j \in J} \Delta_j) = TM$. In other words we have shown that the Δ_i's, $i \in \underline{m}$, are compatible and satisfy the conditions of Theorem 5.1 of [8]; so we have produced a *constructive* local solution of the input-output decoupling problem. For the case of scalar outputs, such a constructive solution has already been provided in [8, Thm. 5.2].

(ii) In general the feedback $u = \alpha(x) + \beta(x)v$ is only locally well defined. Without any further requirements on the state space M nothing can be said about the global solution of the noninteracting control problem. For instance, the question whether or not the manifold M is simply connected will probably be of importance (see [9]). From [18] comes the suggestion that the holonomy group of a certain integrable connection plays a crucial role. An approach to global problems via singularity theory is advocated by C.I. Byrnes in [27], [28]. Much further work will be needed to fully develop a theory of global nonlinear decoupling.

(iii) Combining (2.10) (or (3.56)) with the result on nonlinear systems invertibility from [15], we see that each subsystem

$$\begin{aligned} \dot{x}(t) &= \tilde{A}(x(t)) + \tilde{B}_i(x(t))v_i(t), \qquad x(0) = x_0, \\ z_i(t) &= C_i(x(t)) \end{aligned} \tag{3.67}$$

is strongly invertible at x_0. In the case that each output function is 1-dimensional, so $p_1 = \cdots = p_m = 1$, our results coincide with the results obtained in [20], which is the nonlinear version of a result of [4], see also [8], [2].

(iv) As we also assume that $p_1 + \cdots + p_m = n$ the total output map $C = (C_1, \cdots, C_m)$ is a (local) diffeomorphism, we have that $\bigcap_{i \in \underline{m}} \operatorname{Ker} C_{i*} = 0$ and our result follows by the fact that in this case the conditions $\Delta_0 = \Delta_0 \cap R_1^* + \cdots + \Delta_0 \cap R_m^*$ and $R_i^* + \operatorname{Ker} C_{i*} = TM$, $i \in \underline{m}$, are equivalent. By using some easy dimension arguments one can show that $R_i^* = \bigcap_{j \neq i} \operatorname{Ker} C_{j*}$ provided that the preceding condition holds, see also [16] and (3.44).

(v) The strong accessibility assumption for (2.1) is not needed. What is necessary in Theorem 3.1 is that the strongly accessible distribution (see [14], [22]) R_0 satisfies $C_*(R_0) = T(N_1 \times \cdots \times N_m)$, where $C = (C_1, \cdots, C_m)$.

(vi) If the number of input channels exceeds the number of outputs, we obtain by doing similar computations as in Lemma 3.2-Lemma 3.7 that a *sufficient* condition for the solvability of the regular local noninteracting control problem is $\Delta_0 = \bigoplus_{i \in \underline{m}} \Delta_0 \cap R_i^*$, provided that $A_1 - A_4$ hold. This serves as a starting point for a more general decoupling problem in [26].

4. The structure of a decoupled system. Next we want to discuss the (local) structure of an input-output decoupled system. We will show that as in the linear theory, see [12], our decoupled system possesses a natural local canonical form which can be built up from the distributions R_i^*. An alternative derivation is given in [8], and in [25] one can find a more algebraically oriented approach to the same problem. We consider a nonlinear system (2.1)-(2.2) satisfying the conditions and assumptions for the regular local noninteracting control problem, see § 3. For simplicity we will take $m = 2$; the general case follows by an easy induction argument. For $m = 2$ consider the nested sequence of involutive distributions

$$R_1^* \cap R_2^* \subset R_1^* \subset R_1^* + R_2^* = TM \tag{4.1}$$

(here the last equality follows from the strong accessibility assumption). By [19] we know that around $x_0 \in M$ there exists a coordinate system such that

$$R_1^* \cap R_2^* = \operatorname{Span}\left\{\frac{\partial}{\partial x_3}\right\}, \tag{4.2a}$$

$$R_1^* = \operatorname{Span}\left\{\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_3}\right\}, \tag{4.2b}$$

$$R_1^* + R_2^* = \operatorname{Span}\left\{\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \frac{\partial}{\partial x_3}\right\} \tag{4.2c}$$

with each x_i, $i \in \underline{3}$, possibly being a vector.

Let us write the (locally) decoupled system as

$$\begin{aligned} \dot{x} &= \tilde{A}(x) + \tilde{B}_1(x)v_1 + \tilde{B}_2(x)v_2, \\ y_1 &= C_1(x), \\ y_2 &= C_2(x); \end{aligned} \tag{4.3}$$

then we have by definition of the regular local controllability distributions R_1^* and R_2^* that for $i = 1, 2$

$$R_i = \operatorname{Span}\{ad_{\tilde{A}}^k \tilde{B}_i,\ ad_{\tilde{B}_j}^k \tilde{B}_i \mid k \in \mathbb{Z}_+, j = 1, 2\}_{LA} \tag{4.4}$$

and so

$$
\begin{aligned}
&[\tilde{A}, R_i^*] \subset R_i^*, \\
&[\tilde{B}_j, R_i^*] \subset R_i^*, \qquad j = 1, 2.
\end{aligned} \tag{4.5}
$$

From (4.4) and (4.5) one can easily verify, by using the Jacobi identity extensively, that there exist sets of vector fields $\{X_\alpha\}_{\alpha \in I}$ and $\{Y_\beta\}_{\beta \in J}$ which span R_1^* and R_2^* respectively, such that

$$
[X_\alpha, Y_\beta] \in R_1^* \cap R_2^* = \operatorname{Span}\left\{\frac{\partial}{\partial x_3}\right\}, \qquad \alpha \in I, \quad \beta \in J. \tag{4.6}
$$

Here we recognize the ideal property of § 2. But (4.6) exactly implies, e.g. [19], [22], that the distribution R_2^* is locally given as

$$
R_2^* = \operatorname{Span}\left\{\frac{\partial}{\partial x_2}, \frac{\partial}{\partial x_3}\right\}. \tag{4.7}
$$

Let us now see what this implies for our system. Using (4.2)–(4.7) and the fact that $R_1^* \subset \operatorname{Ker} C_{2*}$ and $R_2^* \subset \operatorname{Ker} C_{1*}$ we obtain that (4.3) reads in our local coordinates as

$$
\begin{aligned}
\dot{x}_1 &= \tilde{A}_1(x_1) + \tilde{B}_1(x_1)v_1, \\
\dot{x}_2 &= \tilde{A}_2(x_2) + \tilde{B}_2(x_2)v_2, \\
\dot{x}_3 &= \tilde{A}_3(x_1, x_2, x_3) + \sum_{i=1}^{2} \tilde{B}_{3i}(x_1, x_2, x_3)v_i, \\
y_1 &= C(x_1), \\
y_2 &= C(x_2).
\end{aligned} \tag{4.8}
$$

This is what one might call a *local canonical form* for the input-output decoupled system. This is a direct generalization of the well-known linear result of [12]. In the same way one has for $m > 2$ the following *local canonical form*:

$$
\begin{aligned}
\dot{x}_1 &= \tilde{A}_1(x_1) + \tilde{B}_1(x_1)v_1, \\
&\vdots \\
\dot{x}_m &= \tilde{A}_m(x_m) + \tilde{B}_m(x_m)v_m, \\
\dot{x}_{m+1} &= \tilde{A}_{m+1}(x_1, \cdots, x_{m+1}) + \sum_{i=1}^{m} \tilde{B}_{(m+1)i}(x_1, \cdots, x_{m+1})v_i, \\
y_1 &= C(x_1), \\
&\vdots \\
y_m &= C(x_m).
\end{aligned} \tag{4.9}
$$

In this coordinate system one has $R_i^* = \operatorname{Span}\{\partial/\partial x_i, \partial/\partial x_{m+1}\}$, $i = 1, \cdots, m$. It is immediately seen from (4.9) that we have output controllability in each channel; this is an aspect that is not covered in [8].

Acknowledgment. The developments of § 4 benefited from suggestions by a reviewer.

REFERENCES

[1] G. BASILE AND G. MARRO, *A state space approach to noninteracting controls*, Ricerche di Automatica, 1 (1970), pp. 68-77.

[2] D. CLAUDE, *Decoupling of nonlinear systems*, Syst. Contr. Lett., 1 (1982), pp. 242-248.

[3] P. E. CROUCH AND N. CARMICHAEL, *Application of linear analytic systems theory to attitude control*, Report to ESTEC, Applied Systems Studies, Coventry, UK, 1981.

[4] P. L. FALB AND W. A. WOLOVICH, *Decoupling in the design and synthesis of multivariable control systems*, IEEE Trans. Automat. Contr., AC-12 (1967), pp. 651-659.

[5] H. FREEMAN, *Stability and physical realizability considerations in the synthesis of multipole control systems*, AIEE Trans. (Appl. Ind.) 77 (1958), pp. 1-5.

[6] M. L. J. HAUTUS AND M. HEYMANN, *Linear feedback decoupling-transfer function analysis*, preprint, Feb. 1980.

[7] R. M. HIRSCHORN, *(A, B)-Invariant distributions and disturbance decoupling of nonlinear systems*, this Journal, 19 (1981), pp. 1-19.

[8] A. ISIDORI, A. J. KRENER, C. GORI-GIORGI AND S. MONACO, *Nonlinear decoupling via feedback: a differential geometric approach*, IEEE Trans. Automat. Contr. AC-26 (1981), pp. 331-345.

[9] ———, *Locally (f, g)-invariant distributions*, Syst. Contr. Lett., 1 (1981), pp. 12-15.

[10] R. J. KAVANAGH, *Multivariable control system synthesis*, AIEE Trans. (Appl. Ind.) 77 (1958), pp. 425-429.

[11] A. J. KRENER AND A. ISIDORI, *(ad f, G)-Invariant and controllability distributions in feedback control of linear and nonlinear systems*, Lecture Notes in Control and Information Sciences, 39, Springer-Verlag, Berlin, 1982, pp. 157-164.

[12] A. S. MORSE AND W. M. WONHAM, *Status of noninteracting control*, IEEE Trans. Automat. Contr., AC-16 (1971), pp. 568-581.

[13] H. NIJMEIJER, *Controlled invariance for affine control systems*, Int. J. Contr., 34 (1981), pp. 825-833.

[14] ———, *Controllability distributions for nonlinear systems*, Syst. Contr. Lett., 2 (1982), pp. 122-129.

[15] ———, *Invertibility of affine nonlinear control systems: a geometric approach*, Syst. Contr. Lett., 2 (1982), pp. 163-168.

[16] ———, *Feedback decomposition of nonlinear control systems*, IEEE Trans. Automat. Contr., AC-28 (1983), pp. 861-863.

[17] ———, *The triangular decoupling problem for nonlinear control systems*, Nonlinear Anal. Appl., 8 (1984), pp. 273-279.

[18] H. NIJMEIJER AND A. J. VAN DER SCHAFT, *Controlled invariance for nonlinear systems*, IEEE Trans. Automat. Contr., AC-27 (1982), pp. 904-914.

[19] W. RESPONDEK, *On decomposition of nonlinear control systems*, Syst. Contr. Lett., 1 (1982), pp. 301-308.

[20] P. K. SINHA, *State feedback decoupling of nonlinear systems*, IEEE Trans. Automat. Contr., AC-22 (1977), pp. 487-489.

[21] H. J. SUSSMANN AND V. JURDJEVIC, *Controllability of nonlinear systems*, J. Differential Equations, 12 (1972), pp. 95-116.

[22] A. J. VAN DER SCHAFT, *System theoretic descriptions of physical systems*, Ph.D. dissertation, Univ. Groningen, the Netherlands, 1983.

[23] W. M. WONHAM, *Linear Multivariable Control: a Geometric Approach*, Springer, New York, 1979.

[24] W. M. WONHAM AND A. S. MORSE, *Decoupling and pole assignment in linear multivariable systems: A geometric approach*, this Journal, 8 (1970), pp. 1-18.

[25] H. NIJMEIJER, *Noninteracting control for nonlinear systems*, Proc. 22nd IEEE Conference on Decision & Control, 1983, pp. 131-133.

[26] H. NIJMEIJER AND J. M. SCHUMACHER, *Zeros at infinity for affine nonlinear control systems*, IEEE Trans. Automat. Contr., AC-30 (1985), pp. 566-573.

[27] C. I. BYRNES, *Toward a global theory of (f, g)-invariant distributions with singularities*, in Mathematical Theory of Networks and Systems (Proc. MTNS '83, Beer Sheva), P. A. Fuhrmann, ed., Lecture Notes in Control and Information Science 58, Springer, New York, 1984, pp. 149-165.

[28] ———, *Feedback decoupling of rotational disturbances for spherically constrained systems*, Proc. 23rd IEEE Conference on Decision and Control, Las Vegas, 1984, pp. 421-426.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 6, November 1986

010

INFINITE HORIZON STOCHASTIC PROGRAMS*

RICHARD C. GRINOLD†

Abstract. This paper considers a discrete time infinite horizon stochastic program with the objective of minimizing the expected present value of a sum of convex cost functions. In any finite time we assume that only a finite number of outcomes can occur. The paper presents a sequence of solvable finite horizon problems that converge in value and decision to the value and decision of the infinite problem.

Key words. mathematical programming, stochastic programming, infinite horizon, convergence, convexity

AMS(MOS) subject classifications. 60G, 65K

1. Introduction. This paper shows how an infinite horizon stochastic program that minimizes a convex objective, subject to linear constraints, with finite uncertainty at each stage can be approximated by a finite horizon stochastic program. The infinite horizon stochastic program can be considered as a decision tree starting from a single node at time 0. Each path through the tree corresponds to a realization of both the decision process and the stochastic process. At each time T we assume there is a finite collection of paths representing system evolution to that time. The state of the system at time t corresponds to a node in the decision tree; from any of the time t nodes there is a unique path back to the time 0 node. Since the problem has an infinite horizon, there can be an uncountable number of nodes.

The difficulty in solving such a problem is roughly proportional to the number of nodes. Our solution procedure looks at the time T nodes, and, by taking expectations, aggregates the paths emanating from each of the time T nodes into a single path. Then an end effects correction procedure is used to terminate that single path in a loop, at time T or at a later time. The paper shows that this procedure works. As $T \to \infty$, we get objective values that converge monotonically to the optimal value of the stochastic program, and the decisions of the T period problems converge to the optimal decisions of the infinite horizon problems.

This paper is organized as follows. Section 2 describes the stochastic process. Section 3 presents the decision problem along with our assumptions. The problem is a multistage stochastic program with linear constraints and a convex objective. The uncertainty is restricted to the right-hand side (constraining) vector. In § 4 we show how to aggregate the stochastic program into a deterministic program that will give a lower bound on the stochastic program. This idea is used to develop a T period approximation and in § 5 we prove that the values and solutions of the T period problems converge to their long-term counterparts. In § 6 we extend the model to include infinite time lags in the carry over of decisions; thus a time t decision can have an impact at all future times $t+s$ for $s \geqq 1$. In § 7 we consider the case in which the stochastic process is a finite homogeneous Markov chain. In that case we can allow the constraint matrix to depend on the state of the Markov chain. A final section describes how these ideas can be implemented.

* Received by the editors April 26, 1983, and in final revised form July 4, 1985.
† School of Business Administration, University of California, Berkeley, California 94720.

The paper is based on two earlier treatments on infinite horizon problems. Grinold (1977) treats the deterministic linear case, and Grinold (1983) treats the deterministic convex case. Flam and Wets (1984) have recently extended these results. See the references in those papers for background on infinite horizon programs. This paper uses the result of the 1977 and 1983 papers to extend the treatment to the stochastic case in §§ 3, 4, and 5. This calls for two insights: first correctly building the model, so that results of the 1977 and 1983 papers apply and second, using the method of proof in the 1983 paper twice-aggregating over outcomes and over time. In § 6, we extend the basic results of 1977 and 1983 to allow for infinite time lags. Finally in § 7, we synthesize the new results of this paper with an earlier paper, Grinold (1976), on a Markovian multi-stage system.

In some ways our procedure is an example of the successive use of partitioning the sample space, see Pfanzangl (1974). Huang, Ziemba and Ben-Tal (1977), Hausch and Ziemba (1983), Marti (1977) and Birge (1985) have used this idea in relation to stochastic programs. More recently Birge (1984) has applied a technique similar to the one proposed here to a finite horizon stochastic production planning problem.

Some basic work on multi-stage stochastic programs can be found in Rockafellar and Wets (1974), (1976), Eisner and Olsen (1975) and Olsen (1976). In these papers the complicating element is the nonfinite range of the random variables. In our model stochastic process has finite range over any finite time span. The complicating factor is the infinite horizon and the compounding of uncertainty.

There is a large literature on infinite horizon problems in the Markov programming literature. However, those results do not carry over to our case. There are two main approaches to infinite horizon problems in the Markov programming literature. The first is called value iteration. In value iteration one ignores the effect of period T decisions on all rewards and constraints in later periods $t > T$. Grinold (1977) considered the analogous procedure for the class of infinite horizon multistage programming problems. The procedure will not work in the general case. An example is given in Grinold (1977). The basic reason is an ability to borrow against the future; if the model allows us to borrow (either money or backlogged orders) then we can build up a huge debt that is ignored in period $T+1$. The other solution procedure advanced in the Markov programming literature is called policy iteration. In that case we focus on a policy, and test for possible improvements. If there are none the policy is optimal; if improvements exist then a better policy is obtained. This procedure may be possible for the class of problems considered here, but it seems far more complicated than the solution procedure we suggest. A policy in our context is a function that depends on the stochastic process of right-hand side (constraint) vector. This is not a promising approach unless we can prove that the optimal policy has a very special form; e.g. if the matrices A in what follows are Leontief substitution matrices and K and the sequence $b(t)$ are nonnegative, and the objective $f[x]$ is linear.

In particular, Schal (1975) considers a very general class of decision problems. Despite its generality there are differences between our model and Schal's. First, our model is essentially nonstationary. A complete state description at time t calls for knowledge of the subset of the partition $\Omega(t)$ that was obtained as well as the decisions $x(0), x(1), \cdots, x(t-1)$ taken at earlier times. In addition, we do not impose any strong boundedness assumptions on our objective function or compactness assumption on our set of decisions. The boundedness assumption we use, see § 3, is the natural analogue of the boundedness assumptions in linear and nonlinear programming. The boundedness assumption is as weak as possible, if the problem is to have a finite optimal value and a proper optimal value function.

2. The stochastic process. Consider a set Ω of possible evolutions of the underlying stochastic process. Knowledge of $\omega \in \Omega$ gives us a complete history of all random events. As the stochastic process gradually unfolds, we learn more and more about the actual ω. This is made precise by considering a sequence of partitions $\Omega(t)$ of Ω where $\Omega(t+1)$ is a refinement of $\Omega(t)$. At time t we know that ω is in some particular member, say Q, of the partition $\Omega(t)$; at time $t+1$ we can be sure ω is in some subset of Q that is a member of the partition $\Omega(t+1)$. Since we are interested in practical solution methods, we will assume that $\Omega(t)$ is finite for all t.

Let $\mathcal{F}(t)$ be the smallest Borel field containing $\Omega(t)$. Since $\Omega(t+1)$ has finite cardinality $N(t)$, $\mathcal{F}(t)$ will consist of $2^{N(t)}$ possible unions of sets in $\Omega(t)$.

A measurable function $x(t):(\Omega, \mathcal{F}(t)) \to \mathbb{R}^n$ will be called t-measurable. Such an $x(t)$ will be constant on each of the sets in the partition $\Omega(t)$. A t-measurable function $x(t)$ will also be s measurable for $s \geqq t$ since $\Omega(s)$ refines $\Omega(t)$. A t-measurable function cannot anticipate events that occur after time t. It is constant on each element of $\Omega(t)$; it cannot exploit subsequent refinements of $\Omega(t)$. This property is called a *nonanticipation* condition.

Let $\mathcal{F}$ be the smallest Borel field including all of the $\mathcal{F}(t)$ and π be a probability measure defined on F.

The expectations operation E is defined by

$$E[z] = \int_{\Omega} z(\omega)\, d\pi(\omega). \tag{2.1}$$

Let E^T denote the conditional expectations operator with respect to $\mathcal{F}(t)$. For any $t > T$ and t-measurable function $x(t)$, $E^T[x(t)]$ is the expectation of $x(t)$ conditional on knowledge up to time T. As such $E^T[x(t)]$ will be T-measurable; i.e. it will be constant on each set in $\Omega(T)$.

3. The decision problem. For $t = 0, 1, 2, \cdots$, let $b(t):(\Omega, \mathcal{F}(t)) \to \mathbb{R}^n$ be a t-measurable function; $b(t)$ will depend only on the partition $\Omega(t)$; b satisfies the regularity condition

$$\sum_{t=0}^{\infty} \alpha^t E\{\|b(t)\|\} < \infty.$$

At time t the decision maker will know which set in $\Omega(t)$ has obtained and a history of past decision $x(s)$ for $s = 0, 1, \cdots, t-1$.

The decision maker's policy at time t is a t-measurable function $x(t)$ that must satisfy

$$Ax(t) = b(t) + Kx(t-1), \qquad x(t) \geqq 0. \tag{3.1}$$

This section considers only a one period decision carry over; the decision $x(t-1)$ effects the constraints at time t but earlier decisions $x(t-s)$ for $s \geqq 2$ have no effect. The one period lag is theoretically equivalent to any finite lag. In § 6 the model is extended to include infinite lags. The relation (3.1) holds almost surely; consider any set Q in the partition $\Omega(t)$ where the probability of Q obtaining is positive. The functions $x(t)$, $b(t)$, and $x(t-1)$ are all t-measurable and therefore all constant for $\omega \in Q$. Since $\Omega(t)$ has finite cardinality $N(t)$, then (3.1) is just a set of $N(t)$ linear equations, one for each possible history up to time t. This notation suppresses the dependence of the decision function $x(t)$ and right-hand side $b(t)$ on the stochastic process. If the partition $\Omega(t)$ has cardinality $N(t)$, then $x(t)$ is a vector with $n \cdot N(t)$ elements; i.e. an n element vector for each member of $\Omega(t)$. Similarly $b(t)$ is an $m \cdot N(t)$ element vector.

A policy x is the sequence $x(t)$ for $t=0,1,\cdots$, of random variables satisfying (3.1). The value of that policy is given by

$$F[x]=\limsup_{T\to\infty}\sum_{t=0}^{T}\alpha^t E\{f[x(t)]\} \tag{3.2}$$

where α is a discount factor lying between zero and one, and f is a closed (lower semi-continuous), proper, convex function.

Let $X(b)$ be the set of policies that satisfy (3.1). Our problem then is to find

Problem P

$$V(b)=\operatorname{Inf}\{F[x]\,|\,x\in X(b)\}. \tag{3.3}$$

If $X(b)$ is empty, then $V(b)$ is defined as $+\infty$.

We will make two assumptions about Problem P.

A1: *Feasibility.* A solution $x\in X(b)$ is said to be *α-convergent* if

$$\sum_{t=0}^{\infty}\alpha^t E\{ex(t)\}<\infty, \tag{3.4}$$

e is a summation vector. We assume there is an α-convergent solution.

A2: *Boundedness.* Let $f0^+[y]$ be the recession function of f. From Rockafellar (1970, pp. 66–71), we see that the recession function is closed, convex and homogeneous of degree one.

Let $0\leqq\lambda\leqq\alpha$ and y satisfy

$$[A-\lambda K]y=0,\quad y\geqq 0,\qquad ey=1. \tag{3.5}$$

Then A2 requires

$$f0^+[y]>0.$$

These assumptions are discussed in detail in Grinold (1977) and (1983). We shall prove below that if A2 is satisfied and A1 is not, then $V(b)=+\infty$. It is also true, and easy to show (see Grinold (1977)), that if A1 is satisfied, and if $f0^+[y]<0$ for some λ and y satisfying (3.5), then V is an improper convex function; i.e. $V(b)=-\infty$ on the relative interior of its domain. With the exception of a borderline case, A1 and A2 only rule out problems with infinite optimal values. Theorem 1 shows the importance of A1.

THEOREM 1. *Under* A2. *If $x\in X(b)$ is not α-convergent, then $F[x]=+\infty$.*

Proof. Let

$$z(t)=E[x(t)] \tag{3.6}$$

and

$$d(t)=E[b(t)]. \tag{3.7}$$

Then, by taking expectations on the constraints (3.1) we get

$$Az(t)=d(t)+Kz(t-1),\qquad z(t)\geqq 0. \tag{3.8}$$

By Jensen's inequality

$$E\{f[x(t)]\}\geqq f[z(t)]. \tag{3.9}$$

Thus

$$F[x]\geqq\limsup_{T\to\infty}\sum_{t=0}^{T}\alpha^t f[z(t)]=F[z]. \tag{3.10}$$

Consider the deterministic problem of minimizing the right-hand side of (3.10) subject to the constraints (3.8). If x is not α-convergent, then z will not be α-convergent since

$$\sum_{t=0}^{\infty} \alpha^t E\{ex(t)\} = \sum_{t=0}^{\infty} \alpha^t ez(t) = \infty. \tag{3.11}$$

By Theorem 4.1 of Grinold (1983), we have

$$\limsup_{T\to\infty} \sum_{t=0}^{T} \alpha^t f[z(t)] = +\infty;$$

thus from (3.10) $F[x] = +\infty$. Q.E.D.

4. The T-period problem. Theorem 1 focuses attention on α-convergent solutions. Solutions that are not α-convergent have infinite value and are not candidates for optimality when the objective value $V(b)$ is finite. The T period approximating problem, called $P(T)$, is motivated by showing how an α-convergent policy x maps into a feasible policy x^T for $P(T)$, and that the objective value of x^T in $P(T)$ is less than its value, $F[x]$, in P. Assumption A1 guarantees the existence of an α-convergent policy so $P(T)$ is feasible for all T, and the optimal values $V^T(b)$ of $P(T)$ converge monotonically upwards to $V(b)$.

Let $E^T[\ ]$ be the expectations operator conditional on $\mathcal{F}(T)$. For b and any policy x define

$$\begin{aligned} &\text{(i)} && b^T(t) = E^T[b(t)], \\ &\text{(ii)} && x^T(t) = E^T[x(t)]. \end{aligned} \tag{4.1}$$

The functions $b^T(t)$ and $x^T(t)$ are T-measurable for all $t \geqq T$. In particular, $b^T(t) = b(t)$ and $x^T(t) = x(t)$ for $t \leqq T$. Moreover, $b^T(t)$ and $x^T(t)$ satisfy the constraints

$$Ax^T(t) = b^T(t) + Kx^T(t-1), \qquad x^T(t) \geqq 0 \tag{4.2}$$

for all t.

We will aggregate the constraints (4.2) for $t \geqq T$ by multiplying the tth constraint by $(1-\alpha)\alpha^{t-T}$ and summing. This gives

$$(A - \alpha K)\tilde{x}(T) = \tilde{b}(T) + (1-\alpha)Kx^T(T-1), \qquad \tilde{x}(T) \geqq 0, \tag{4.3}$$

where

$$\begin{aligned} &\text{(i)} && \tilde{x}(T) = (1-\alpha) \sum_{t=T}^{\infty} \alpha^{t-T} x^T(t), \\ &\text{(ii)} && \tilde{b}(T) = (1-\alpha) \sum_{t=T}^{\infty} \alpha^{t-T} b^T(t). \end{aligned} \tag{4.4}$$

The constraints for $P(T)$ are equations (4.2) for $0 \leqq t < T$, and equation (4.3). Recall A is an m by n matrix; $P(T)$ will have $m * \sum_{t=0}^{T} N(T)$ constraints with $n * \sum_{t=0}^{T} N(t)$ variables where $N(t)$ is the cardinality of the partition $\Omega(t)$. A vector $x^T = (x^T(0), x^T(1), \cdots, x^T(T-1), \tilde{x}(T))$ is feasible for $P(T)$ if x^T satisfies (4.2) for $t < T$ and (4.3). Let $X^T(b)$ be the set of feasible solutions for $P(T)$; each α-convergent solution in $X(b)$ maps into a solution in $X^T(b)$.

The objective for $P(T)$ is

$$F^T[x^T] = \sum_{t=0}^{T-1} \alpha^t E[f\{x^T(t)\}] + \frac{\alpha^T}{(1-\alpha)} E[f\{\tilde{x}(T)\}]. \tag{4.5}$$

Problem $P(T)$ and its optimal value are thus defined by

Problem $P(T)$

$$V^T(b) = \text{Inf}[F^T[x^T] \mid x^T \in X^T(b)]. \tag{4.6}$$

For an α-convergent solution, one can show by Jensen's inequality, lower semi-continuity, and the definition of $\tilde{x}(T)$ that an α-convergent x must satisfy

$$\sum_{t=T}^{\infty} \alpha^t E[f[x(t)]] \geqq \frac{\alpha^T}{(1-\alpha)} E[f[\tilde{x}(T)]]. \tag{4.7}$$

Therefore $F^T[x^T] \leqq F[x]$, and $V^T(b) \leqq V(b)$.

The properties of $P(T)$ are summarized in Proposition 1.

PROPOSITION 1. *Under* A1 *and* A2.

(i) $$-\infty < V^0(b) \leqq \cdots \leqq V^T(b) \leqq V^{T+1}(b) \leqq V(b),$$

and

(ii) *An optimal solution to $P(T)$ exists.*

Proof. $V^T(b) \leqq V(b)$ has already been demonstrated. To show $V^T(b) \leqq V^{T+1}(b)$ let x^{T+1} be feasible for $P(T+1)$. Then x^T defined by

$$\begin{aligned} &\text{(i)} && x^T(t) = x^{T+1}(t) \quad \text{for } t < T, \\ &\text{(ii)} && \tilde{x}(T) = (1-\alpha)x^{T+1}(T) + \alpha\tilde{x}(T+1) \end{aligned} \tag{4.8}$$

will be feasible for $P(T)$ and by Jensen's inequality will have a lower objective value; i.e. $F^T[x^T] \leqq F^{T+1}[x^{T+1}]$.

Assume $V^0(b)$ is not bounded below. Then there will exist a sequence of feasible solutions z^k for $P(0)$ with

$$F^0[z^k] = \frac{1}{1-\alpha} f[z^k]. \tag{4.9}$$

The z^k are not bounded so $ez^k \to \infty$.

Let $y^k = z^k / ez^k$, and let y be a limit point of the y^k; y must satisfy

$$[A - \alpha K]y = 0, \quad y \geqq 0, \qquad ey = 1. \tag{4.10}$$

However, Lemma 3 of Grinold (1983), and the presumption that $F^0[z^k] \to -\infty$, imply $f0^+[y] \leqq 0$ in violation of A2. Thus $V^0(b)$ must be bounded below.

To prove (ii) use A2 and the results on existence in Rockafellar (1970, pp. 77, 267). Q.E.D.

5. Convergence. The optimal values and solutions of the T-period problems $P(T)$ converge, in a sense described exactly below, to the optimal value and solution of the infinite horizon problem P.

Before we state the convergence theorem, we need a definition; the sequence x^T *converges diagonally* to x if for every τ there exists a subsequence of integers $S(\tau)$ with $S(\tau)$ a subsequence of $S(\tau-1)$ such that for all $t \leqq \tau$

$$x^T(t) \to x(t)$$

as $T \in S(\tau)$ goes to infinity.

THEOREM 2. *Under* A1 *and* A2.

(i) $\lim_{T\to\infty} V^T(b) = V(b)$.

(ii) *Suppose $V(b)$ is finite. If x^T is optimal for $P(T)$, then there exists an x such that x is optimal for P and x^T converges diagonally to x.*

Proof. Let k be the limit of the $V^T(b)$. If k is infinite, then (i) is true and (ii) is not relevant.

Let x^T be optimal for $P(T)$. Define:

$$\text{(i)} \qquad z^T(t) = E\{x^T(t)\},$$

$$\text{(ii)} \qquad \tilde{z}(T) = E\{\tilde{x}(T)\},$$

$$\text{(iii)} \qquad d^T(t) = E\{b^T(t)\},$$

$$\text{(iv)} \qquad \tilde{d}(T) = E\{\tilde{b}(T)\},$$

$$\text{(5.1)} \quad \text{(v)} \qquad \tilde{z}(T,0) = (1-\alpha)\sum_{t=0}^{T-1} \alpha^t z^T(t) + \alpha^T \tilde{z}(T),$$

$$\text{(vi)} \qquad \tilde{d} = (1-\alpha)\sum_{t=0}^{T-1} \alpha^t d^T(t) + \alpha^T \tilde{d}(T),$$

$$\text{(vii)} \qquad M(T) = e\tilde{z}(T,0),$$

$$\text{(viii)} \qquad y(T) = \tilde{z}(T,0)/M(T).$$

Take the constraints of $P(T)$, take expectations, aggregate by multiplying constraints 0 through $T-1$ by $(1-\alpha)\alpha^T$ and constraint T by α^T and summing, divide the result by $M(T)$; this yields

$$\text{(5.2)} \qquad [A-\alpha K]y[T] = \tilde{d}/M(T) \quad \text{where } ey(T) = 1,\ y(T) \geqq 0.$$

We will show by contradiction that $M(T)$ is bounded above.

An application of Jensen's inequality to (4.5) gives

$$\text{(5.3)} \qquad (1-\alpha)V(b) \geqq (1-\alpha)V^T(b) = (1-\alpha)F^T[x^T] \geqq f[\tilde{z}(T,0)].$$

If $V(b)$ is finite and $M(T)$ is unbounded on some subsequence, then there exists a refinement of that subsequence and a vector y such that $y(T) \to y$, and $f[\tilde{z}(T,0)]/M(T) \to \mu \leqq 0$. Lemma 3 of Grinold (1983) tells us that $f0^+[y] \leqq 0$. However, from (5.2) we have

$$\text{(5.4)} \qquad [A-\alpha K]y = 0, \quad ey = 1, \quad y \geqq 0.$$

Thus, by assumption A2, $f0^+[y] > 0$. This contradiction shows $M(T)$ is bounded above; there exists a finite M such that $M(T) \leqq M$ for all T.

Since $\Omega(t)$ is finite for any t and we can assume, with no loss of generality, that each set in the partition $\Omega(t)$ has positive probability of occurrence, so $x^T(t)$ is a bounded sequence in $N(t)^*n$ space.

Therefore we can first find $x(0)$ and subsequence $S(0)$ of the integers such that $x^T(0) \to x(0)$ for $t \in S(0)$, then successively refine $S(\tau)$ and produce the limiting policy x such that x^T diagonally converges to x. It is easy to show x is feasible for P and that x is α-convergent, in fact

$$\text{(5.5)} \qquad \sum_{t=0}^{\infty} \alpha^T E[ex(t)] \leqq M.$$

It remains to show x is optimal.

For fixed τ and $T > \tau$ define

$$\text{(5.6)} \qquad \tilde{x}(T,\tau) = \sum_{t=\tau}^{T-1} (1-\alpha)\alpha^{t-\tau} x^T(t) + \alpha^{T-\tau}\tilde{x}(T).$$

The expected value of $\tilde{x}(T, \tau)$ conditional on events up to time T is

$$\tilde{z}(T, \tau) = \sum_{t=\tau}^{T-1} (1-\alpha)\alpha^{t-\tau} z^T(t) + \alpha^{T-\tau}\tilde{z}(T). \tag{5.7}$$

Several applications of Jensen's inequality yield

$$(1-\alpha)V^T(b) \geqq \sum_{t=0}^{\tau-1} (1-\alpha)\alpha^T E\{f[x^T(t)]\} + \alpha^\tau f[\tilde{z}(T, \tau)]. \tag{5.8}$$

For all $T > \tau$ we have

$$\alpha^\tau e\tilde{z}(T, \tau) \leqq M. \tag{5.9}$$

Thus there will be a subsequence $\hat{S}(\tau)$ of $S(\tau)$ and a vector $\tilde{y}(\tau)$ such that $x^T(t) \to x(t)$ for $t \leqq \tau$ and $z(T, \tau) \to \tilde{y}(\tau)$ as $T \in \hat{S}(\tau)$ goes to infinity.

Take the lim inf of (5.8) for $T \in S(\tau)$, and the result is

$$k(1-\alpha) \geqq (1-\alpha) \sum_{t=0}^{\tau-1} \alpha^T E\{f[x(t)]\} + \alpha^\tau f[\tilde{y}(\tau)]. \tag{5.10}$$

Now take the lim sup of (5.10) as $\tau \to \infty$; this yields

$$(1-\alpha)k \geqq (1-\alpha)F[x] + \limsup_{\tau\to\infty} \alpha^\tau f[\tilde{y}(\tau)]. \tag{5.11}$$

Note, by definition, $F[x] \geqq k$ so

$$\limsup_{\tau\to\infty} \alpha^\tau f[\tilde{y}(\tau)] \leqq 0. \tag{5.12}$$

From (5.9), $\alpha^\tau \tilde{y}(\tau)$ is bounded. In fact $\alpha^\tau \tilde{y}(\tau) \to 0$. If not, suppose $\alpha^\tau y(\tau) \to y \neq 0$.

By the usual aggregating of equations one can show that $z = y/ey$ satisfies (5.4) and therefore, by A2 and the homogeneity of $f0^+$, that

$$f0^+[z] = \frac{1}{ey} f0^+[y] > 0. \tag{5.13}$$

However, (5.12) and Lemma 3 of Grinold (1983) imply that

$$f0^+[y] \leqq \limsup_{\tau\to\infty} \alpha^\tau f[\tilde{y}(\tau)] \leqq 0. \tag{5.14}$$

The assumption that $y \neq 0$ has produced a contradiction. When $y = 0$, then $f0^+[y] = 0$, so (5.14) implies the lim sup is zero and, by (5.11), $F[x] = k$. Q.E.D.

6. Infinite time lags. The model presented in earlier sections had a one period lag; the time t constraints were only influenced by time $t-1$ decisions. For theoretical purposes that result covers lags of any finite length. In this section the one period lag model is extended to cover infinite time lags and thus allow for many realistic situations in which new capital stock decays exponentially or remains indefinitely.

Let $K(t-s)$ be the $t-s$ period lag. The constraints of P become

$$Ax(t) = b(t) + \sum_{s=0}^{t-1} K(t-s)x(s), \qquad x(t) \geqq 0. \tag{6.1}$$

Since $x(t)$ and $x(s)$ for $s < t$ are t-measurable, the constraint (6.1) holds for all sets in the partition $\Omega(t)$.

Now define

$$\tilde{K}(t, \theta) = \sum_{j=t}^{\infty} \theta^j K(j). \tag{6.2}$$

The constraints in approximating problem $P(T)$ coincide with (6.1) for $t < T$. The Tth constraint is

$$[A - \tilde{K}(1, \alpha)]\tilde{x}(T) = \tilde{b}(T) + (1-\alpha)\sum_{s=0}^{T-1} \tilde{K}(T-s, \alpha)\alpha^s x(s), \qquad \tilde{x}(T) \geqq 0. \tag{6.3}$$

The boundedness condition, A2, becomes: for $0 \leqq \theta \leqq \alpha$ and y satisfying

$$[A - \tilde{K}(1, \theta)]y = 0, \qquad ey = 1, \quad y \geqq 0, \tag{6.4}$$

we must have $f0^+[y] > 0$.

For any $z \in R^n$, let the norm of z be defined as

$$\|z\| = \sum_{i=1}^{n} |z_i|, \tag{6.5}$$

and for K an m by n matrix

$$\|K\| = \text{Max}\,\{\|Kz\| \mid \|z\| = 1\}; \tag{6.6}$$

therefore

$$\|Kz\| \leqq \|K\|\,\|z\| \quad \text{for all } z. \tag{6.7}$$

The assumption we make on the matrices $K(t)$ is

$$\|K(t)\| \leqq \psi^t k \tag{6.8}$$

where $\psi\alpha < 1$.

PROPOSITION 2. *Under* (6.8), *Theorems* 1 *and* 2 *are valid with infinite time lags.*

Proof. We need only extend Theorem 4.1 of Grinold (1983) to cover infinite lags. That requires a different method of proof: We wish to show that a non-α-convergent solution $z(t)$ of the deterministic constraints

$$Az(t) = d(t) + \sum_{s=0}^{t-1} K(t-s)z(s) \tag{6.9}$$

will have objective value $F(z)$ (see (3.10)) equal to plus infinity.

Let ρ be the radius of convergence of the series $\sum_{t=0}^{\infty} \theta^t ez(t)$. Suppose first that $0 < \rho \leqq \alpha$. Choose $\theta < \rho$, multiply constraint t of (6.9) by $(1-\theta)\theta^t$ and sum. We get

$$[A - \tilde{K}(1, \theta)]\tilde{z}(\theta) = \tilde{d}(\theta), \qquad \tilde{z}(\theta) \geqq 0, \tag{6.10}$$

where

$$\begin{aligned} &\text{(i)} && \tilde{z}(\theta) = (1-\theta)\sum_{t=0}^{\infty} \theta^t z(t), \\ &\text{(ii)} && \tilde{d}(\theta) = (1-\theta)\sum_{t=0}^{\infty} \theta^t d(t). \end{aligned} \tag{6.11}$$

Now define

$$M(\theta) = e\tilde{z}(\theta), \qquad \tilde{y}(\theta) = \tilde{z}(\theta)/M(\theta). \tag{6.12}$$

Recall that $\rho < 1$ is the radius of convergence of $\sum_{t=0}^{\infty} \theta^t ez(t)$. As θ increases to ρ, $M(\theta)$ increases to $+\infty$. Therefore $\tilde{d}(\theta)/M(\theta)$ converges to zero. Consider a sequence of θ increasing to ρ and a limit point y of the sequence $\tilde{y}(\theta)$.

The vector y will satisfy

$$[A-K(1,\rho)]y=0, \quad ey=1, \quad y\geqq 0. \tag{6.13}$$

By assumption A2, $f0^{+}[y]>0$.

If the

$$\limsup_{T\to\infty} \sum_{t=0}^{T} \rho^{t} f[z(t)] = +\infty, \tag{6.14}$$

then, by Lemma 2 of Grinold (1977),

$$F[z]=+\infty. \tag{6.15}$$

If (6.14) is false, then there exists a B such that

$$\sum_{t=0}^{T} \rho^{t} f[x(t)] \leqq B \quad \text{for all } T. \tag{6.16}$$

Now apply the logic of Lemma 2 of Grinold (1977) again with

$$a_t=\theta^{t} f[z(t)], \quad \mu=\frac{\theta}{\rho}<1, \quad r_T=\sum_{t=0}^{T} \rho^{t} f[z(t)]. \tag{6.17}$$

We have

$$\sum_{t=0}^{T} \theta^{t} f[z(t)] = (1-\mu) \sum_{t=0}^{T} \mu^{t} r_T + \mu^{T+1} r_T. \tag{6.18}$$

From (6.16), $r_T \leqq B$ for all T, so

$$\sum_{t=0}^{T} \theta^{t} f[z(t)] \leqq B \quad \text{for all } t. \tag{6.19}$$

By the convexity of f

$$\frac{(1-\theta)}{(1-\theta^{T+1})} \sum_{t=0}^{T} \theta^{t} f[z(t)] \geqq f[\tilde{z}(T,\theta)] \tag{6.20}$$

where

$$\tilde{z}(T,\theta)=\frac{1-\theta}{1-\theta^{T+1}} \sum_{t=0}^{T} \theta^{t} z(t). \tag{6.21}$$

As $T\to\infty$, $\tilde{z}(T,\theta)\to\tilde{z}(\theta)$.

Since f is lower semi-continuous

$$\limsup_{T\to\infty} f[\tilde{z}(T,\theta)] \geqq \liminf_{T\to\infty} f[\tilde{z}(T,\theta)] \geqq f[\tilde{z}(\theta)]. \tag{6.22}$$

When (6.22) is combined with (6.19), we have

$$B \geqq \frac{1}{1-\theta} f[\tilde{z}(\theta)] \quad \text{for } \theta<\rho. \tag{6.23}$$

Now $\tilde{z}(\theta)/M(\theta)\to y$, on some sequence of θ ascending to ρ. On a subsequence we must have

$$\frac{1}{M(\theta)} f[\tilde{z}(\theta)] \to \mu \leqq 0,$$

by (6.23) and $M(\theta)$ increasing to plus infinity. By Lemma 3 of Grinold (1983), $f0^+[y] \leqq \mu \leqq 0$. However, by A2, $f0^+[y]$ is positive. This contradiction arises when we accept (6.16) rather than (6.14). Thus (6.14) is valid and $F[z] = +\infty$.

Now consider the case $\rho = 0$. For T fixed and $0 < \theta \leqq \alpha$ use the weights $((1-\theta)/1-\theta^{T+1})\theta^t$ to multiply the tth constraint of (6.9), for $t \leqq T$, and to multiply the identity.

$$-\sum_{s=0}^{T-1} K(t-s)z(s) = -\sum_{s=1}^{T} K(s)z(T-s) \tag{6.24}$$

for $t > T$. Sum the terms; this yields

$$[A-\tilde{K}(1,\theta)]\tilde{z}(T,\theta) = \tilde{d}(T,\theta) - \frac{1-\theta}{1-\theta^{T+1}} \sum_{s=1}^{T} \tilde{K}(s,\theta)\theta^{T-s}z(T-s) \tag{6.25}$$

where $\tilde{K}(s,\theta)$ is defined by (6.2), $\tilde{z}(T,\theta)$ by (6.21) and $\tilde{d}(T,\theta)$ by

$$\tilde{d}(T,\theta) = \frac{1-\theta}{1-\theta^{T+1}} \sum_{t=0}^{T} \theta^t d(t). \tag{6.26}$$

Now let

$$\begin{aligned} &\text{(i)} && M(T,\theta) = e\tilde{z}(T,\theta), \quad \text{and} \\ &\text{(ii)} && L(T,\theta) = (1-\theta^{T+1})M(T,\theta)/(1-\theta). \end{aligned} \tag{6.27}$$

For θ fixed, as $T \to \infty$, we can find a convergent subsequence $S(0)$ of $\tilde{z}(T,\theta)/M(T,\theta)$ converging to $y(\theta)$. Let $S(1)$ be a subsequence of $S(0)$ such that

$$\frac{\theta^{T-1}\tilde{z}(T-1,\theta)}{L(T,\theta)} \to y(1,\theta) \tag{6.28}$$

for $T \in S(0)$.

In general let $S(\tau)$ be a subsequence of $S(z-1)$ such that

$$\frac{\theta^{T-s}\tilde{z}(T-s,\theta)}{L(T,\theta)} \to y(s,\theta) \tag{6.29}$$

for $T \in S(\tau)$ where $s \leqq \tau$.

We wish to show that $y(\theta)$, and $y(s,\theta)$ for $s \geqq 1$ satisfy

$$[A-\tilde{K}(1,\theta)]y(\theta) = -\sum_{s=1}^{\infty} \tilde{K}(s,\theta)y(s,\theta). \tag{6.30}$$

For τ fixed and $T \in S(\tau)$,

$$\begin{aligned} &\lim_{T\in S(\tau)} \left\{ -\sum_{s=\tau+1}^{T} \tilde{K}(s,\theta)\frac{\theta^{T-s}z(T-s)}{L(T,\theta)} \right\} \\ &\quad = \lim_{T\in S(\theta)} \left\{ [A-\tilde{K}(1,\theta)]\frac{z(T,\theta)}{M(T,\theta)} + \sum_{s=1}^{\tau} \tilde{K}(s,\theta)\frac{\theta^{T-s}z(T-s)}{L(T,\theta)} \right\}. \end{aligned} \tag{6.31}$$

Note that assumption (6.8) implies

$$\begin{aligned} &\text{(i)} && \|K(s,\theta)\| \leqq \frac{(\theta\psi)^s k}{1-\theta\psi}, \quad \text{also that} \\ &\text{(ii)} && \left\| \frac{\theta^{T-s}z(T-s)}{L(T,\theta)} \right\| \leqq 1. \end{aligned} \tag{6.32}$$

Therefore the limit on the left of (6.31) satisfies

$$\left\| \lim_{T \in S(\theta)} \sum_{s=\tau+1}^{T} \tilde{K}(s, \theta) \frac{\theta^{T-s} z(T-s)}{L(T, \theta)} \right\| \leqq \frac{(\theta\psi)^{\tau+2} k}{(1-\theta\psi)^2}. \tag{6.33}$$

The limit on the right of (6.31) is

$$[A - \tilde{K}(1, \theta)] y(\theta) + \sum_{s=1}^{\tau} \tilde{K}(s, \theta) y(s, \theta). \tag{6.34}$$

As $\tau \to \infty$, we see, since $\theta\psi < 1$, that (6.30) holds.

Now consider a sequence converging to zero. Each $y(\theta)$ satisfies $ey(\theta) = 1$, $y(\theta) \geqq 0$, so there will be a y and sequence $y(\theta) \to y$. The $y(s, \theta)$ are all bounded, and $K(s, \theta) \to 0$ as $0 \to 0$; thus y will satisfy

$$Ay = 0, \qquad ey = 1, \quad y \geqq 0. \tag{6.35}$$

By assumption A2, $f0^+[y] > 0$. The lower semi-continuity $f0^+$ ensures

$$f0^+[y(\theta)] > 0 \quad \text{for } \theta \text{ small and positive.} \tag{6.36}$$

However, for this same θ

$$\frac{(1-\theta)}{1-\theta^{T+1}} \sum_{t=0}^{T} \theta^t f[z(t)] \geqq f[z(t, \theta)]. \tag{6.37}$$

The remainder of the proof is like the $\rho > 0$ case. If lim sup of $\sum_{t=0}^{T} \theta^t f[z(t)]$ is infinite we are done; if not the sequence $\sum_{t=0}^{T} \theta^t f[z(t)]$ is bounded. From (6.37) this leads us to conclude $f0^+[y(\theta)] \leqq 0$ in contradiction to (6.36). Q.E.D.

7. The Markovian case. This section presents a refined treatment of the expectations problem when the underlying stochastic process is a finite Markov chain. The Markovian assumption allows us to introduce some randomness into the technology matrix, A, and the carry over matrices, $K(t-s)$, and to still obtain the same strong results of earlier sections.

Let the realization ω be a vector with coordinates $\omega(t)$ that are states of a finite Markov chain. For each t, $\omega(t) \in \{1, 2, \cdots, N\}$ and the distribution of $\omega(t)$ given $\omega(t-1)$ is independent of t and of $\omega(s)$ for $s < t-1$. Thus the probabilities of moving from $\omega(s) = j$ to $\omega(t) = i$ depends only on j and $t-s$; i.e.

$$\pi[\omega(t) = i \,|\, \omega(s) = j] = \pi(i \,|\, j, t-s). \tag{7.1}$$

In this more restricted stochastic process the problem data, A and K can depend on the stochastic evolution of the system, although in a special way. Let $A(\omega, t)$ and $K(\omega, t, s)$ be the data conditional on outcome ω.

The assumption is

$$\begin{aligned} &\text{(i)} && A(\omega, t) = A(i) \quad \text{if } \omega(t) = i, \\ &\text{(ii)} && K(\omega, t-s) = K(i, j, t-s) \quad \text{if } \omega(t) = i \quad \text{and } \omega(s) = j. \end{aligned} \tag{7.2}$$

Let $z(i, t)$ be 1 if $\omega(t) = i$ and 0 if not, and $\pi^T(i, t)$ expectation of $z(i, t)$ given the evolution of the stochastic process up to time T; i.e. given $\omega(0), \omega(1), \cdots, \omega(T)$. For $t \leqq T$, $\pi^T(i, t)$ is 1 if $\omega(t) = i$, and 0 if not. For $t > T$ and $\omega(T) = j$, $\pi^T(i, t)$ equals $\pi(i \,|\, j, t-T)$.

Let

$$
\begin{aligned}
&\text{(i)} && z(i,t)=1 \quad \text{if } \omega(t)=i, 0 \quad \text{if not,}\\
&\text{(ii)} && \pi^T(i,t)=E^T[z(i,t)],\\
&\text{(iii)} && x^T(i,t)=E^T[x(t)\,|\,\omega(t)=i],\\
(7.3)\quad &\text{(iv)} && y^T(i,t)=\pi^T(i,t)x^T(i,t),\\
&\text{(v)} && b^T(i,t)=E^T[b(t)\,|\,\omega(t)=i],\\
&\text{(vi)} && d^T(i,t)=\pi^T(i,t)b^T(i,t),\\
&\text{(vii)} && H(i,j,t-s)=\pi(i\,|\,j,t-s)K(i,j,t-s),
\end{aligned}
$$

where E^T is expectation conditional on the partition $\Omega(T)$. Note that

$$
E^T[x(t)]=\sum_{i=1}^{N} y^T(i,t). \tag{7.4}
$$

PROPOSITION 3. *For any T and $x\in X(b)$ the variables y^T defined in* (7.3) *satisfy*

$$
\text{(i)}\quad A(i)y^T(i,t)=g^T(i,t)+\sum_{s=T+1}^{t-1}\sum_{j=1}^{N} H(i,j,t-s)y^T(j,s) \text{ and}
$$

$$
y^T(i,t)\geqq 0, \quad \textit{for all } t>T, \textit{ where} \tag{7.5}
$$

$$
\text{(ii)}\quad g^T(i,t)=d^T(i,t)+\pi^T(i,t)\sum_{s=0}^{T} E^T\{K(t-s)x(s)\,|\,\omega(t)=i\}.
$$

In addition

$$
\text{(iii)}\quad f\left[\sum_{i=1}^{N} y^T(i,t)\right]\leqq E^T\{f[x(t)]\} \quad \textit{for all } t>T.
$$

Proof. The first result is trivial if $\pi^T(i,t)=0$. That implies $y^T(i,t)$ and $d^T(i,t)$ are zero. Also for each j and $s<t$, either $\pi^T(j,s)=0$ or $\pi(i\,|\,j,t-s)=0$ since

$$
\pi^T(i,t)=\sum_{j=1}^{N}\pi^T(j,s)\pi(i\,|\,j,t-s)=0. \tag{7.6}
$$

So either $y^T(j,s)=0$ or $H(i,j,t-s)=0$ for all j and $s<t$.

Now consider $\pi^T(i,t)$ positive. The constraints defining $X(b)$ are

$$
Ax(t)=b(t)+\sum_{s=0}^{t-1} K(t-s)x(s), \qquad x(t)\geqq 0. \tag{7.7}
$$

We will take expectations of each term conditional on the history up to time T and the state at time t.

$$
E^T\{Ax(t)\,|\,\omega(t)=i\}=A(i)x^T(i,t), \tag{7.8}
$$

$$
E^T\{b(t)\,|\,\omega(t)=i\}=b^T(i,t). \tag{7.9}
$$

For any $T<s<t$, condition on $\omega(s)=j$

$$
\begin{aligned}
&E^T\{K(t-s)x(s)\,|\,\omega(t)=i\}\\
&\quad=\sum_{j=1}^{N}\frac{\pi(i\,|\,j,t-s)\pi^T}{\pi^T(i,t)}(j,s)K(i,j,t-s)E^T\{x(s)\,|\,\omega(s)=j,\omega(t)=i\}.
\end{aligned} \tag{7.10}
$$

Since $x(s)$ is s-measurable and $t > s$, the nonanticipation condition implies

$$E^T\{x(s) \mid \omega(s) = j, \psi(t) = i\} = E^T\{x(s) \mid \omega(s) = j\} = x^T(j, s). \tag{7.11}$$

Thus (7.10) becomes

$$\frac{1}{\pi^T(i, t)} \sum_{j=1}^{N} H(i, j, t-s) y^T(j, s). \tag{7.12}$$

When all expectations are multiplied by $\pi^T(i, t)$, we get item (i) of (7.5).

Item (iii) of (7.5) just is an application of Jensen's inequality using (7.4). Q.E.D.

We have derived an infinite horizon deterministic problem from our stochastic problem. If we fix T, then the vector $y^T(t)$ will have $n \times N^T$ elements. For each $i = 1, 2, \cdots, N$, $y^T(i, t)$ is an n element random variable, with a different value for each of the N^T possible evolutions in the period 0 to T.

From item (iii) of Proposition 2 we see that the deterministic problems will give a lower bound on the value of the stochastic problem. The finite horizon lower bounds will get arbitrarily close to the actual value when we impose a variant, described below, of assumption A2 on the problem.

Let A be the N block by N block diagonal matrix with $A(i)$, an m by n matrix, in diagonal block i for $i = 1, 2, \cdots, N$. Similarly, let $H(t-s)$ be an N block by N block matrix with $H(i, j, t-s)$ the m by n matrix in block (i, j) for i and j running from 1 to N. Let J be a one block by N block matrix with an n by n identity matrix in each block, (i, j). Now let y be an $N * n$ element vector $y(1), y(2), \cdots, Y(N)$ where $y(i)$ is an n element vector for each i. Then Jy is just the sum of the $y(i)$. The variant of A2 is:

For any $\lambda, 0 \leqq \lambda \leqq \alpha$ and y satisfying

$$[A - H(\lambda)]y = 0, \qquad ey = 1, \quad y \geqq 0, \tag{7.13}$$

we have

$$f0^+[Jy] > 0. \tag{7.14}$$

8. Practical considerations. Except in the Markov case of § 7 we have assumed that the technology (as described by A and matrices $K(t-s)$) is independent of the stochastic process and thus of time. We show by example that this does not necessarily have to be true. Suppose there is a technology, say electricity generated by fusion, that may or may not be available in the future. We could include the fusion power technology in the matrices A, $K(1), \cdots$. If fusion is not available in certain events at time t, then we can place an upper bound of zero on the amount of fusion power used in those events. If fusion power is available at time t, then we place a positive bound on the amount used. This moves the uncertainty to an upper bound, and thus to the right-hand side vector $b(t)$.

Notice that this is silly if the upper bound is either zero or five times the amount of electricity demand anticipated in period t. If, on the other hand, we assume a realistic phase-in of the new technology, e.g. the bound is 2% of demand in the first period, 5% in the second up to 15% after several periods, then we could expect reasonable results.

A second item is how to put the theory of this paper into practice. The idea is to break an infinite horizon problem into a three phase problem. The first phase, periods $t = 0, 1, 2, \cdots, T-1$ is a transition phase. During this phase the objectives and constraints depend on the stochastic process. The second phase is called the expectations phase. For $T \leqq t < \tau$ there will be one set of constraints and variables for each outcome

in $\Omega(T)$. The data in this phase can depend on what happened in $[0, T]$. The final phase is called the termination phase. In this phase we aggregate the remainder of the expectations problem for $t = \tau, \tau+1, \cdots$, into one period problem in the manner described for problem $P(T)$.

Some limited experience in the linear deterministic case, Grinold (1983), shows that we can make T and τ relatively small and still get good results.

Acknowledgments. My thanks to the editor and referees for their thorough reports, patience and prodding. Thanks also to Linda Lane who did a superb job of typing.

REFERENCES

J. R. BIRGE, *Aggregation bounds in stochastic programming production problems*, in System Modeling and Optimization, P. Thaft-Christensen, ed., Springer-Verlag Lecture Notes, Berlin, 1984.

———, *Aggregation bounds in stochastic linear programming*, Math. Programming, 31 (1985), pp. 25–41.

M. EISNER AND P. OLSEN, *Duality for stochastic programs interpreted as L. P. in L_p space*, SIAM. J. Appl. Math., 28 (1975), pp. 779–792.

S. D. FLAM AND R. WETS, *Existence results and finite horizon approximations for infinite horizon optimizations problems*, Working Paper CMI #842555-13, C. Michelsen Institute, Bergen, Norway, 1984.

R. GRINOLD, *Finite horizon approximations of infinite horizon linear programs*, Math. Programming, 12 (1977).

———, *A new approach to multi-stage stochastic linear programs*, Math. Programming Stud., 6 (1976), pp. 19–29.

———, *Convex infinite horizon programs*, Math. Programming, 25 (1983), pp. 64–82.

D. B. HAUSCH AND W. T. ZIEMBA, *Boards on the value of information in uncertain decision problems* II, Stochastics (1983); also to appear in Stochastic Optimization, M. A. H. Dempster, ed., M. J. Decker, New York.

C. C. HUANG, W. T. ZIEMBA AND A. BEN-TAL, *Bounds on the expectation of a convex function of a random variable: with applications to stochastic programming*, Oper. Res., 25 (1977), pp. 315–325.

K. MARTI, *Convex approximations to stochastic optimization problems*, in Methods of Operations Research, A. Hein, ed., Meisemheim, Germany, 1977.

P. OLSEN, *Discretizations of multistage stochastic programming problems*, Math. Programming Stud., 6 (1976), pp. 111–124.

J. PFANZANGL, *Convexity and conditional expectations*, Ann. Probab., 2 (1974), pp. 490–494.

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

R. T. ROCKAFELLAR AND R. J. B. WETS, *Continuous versus measurable recourse in N-stage stochastic programming*, J. Math. Anal. Appl., 48 (1974), pp. 836–859.

———, *Nonanticipativity and L_1 Martingales in stochastic optimization problems*, Math. Programming Stud., 6 (1976), pp. 170–187.

M. SCHAL, *Conditions for optimality in dynamic programming and for the limit of n-stage optimal policies to be optimal*, Z. Wahrsch., 32 (1975), pp. 179–196.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 6, November 1986

011

CONSTRAINED CONTROLLABILITY IN BANACH SPACES*

G. PEICHL† AND W. SCHAPPACHER†‡

Abstract. The aim of this paper is to study null-controllability of the linear infinite dimensional control problem $\dot{x} = Ax + Bu$ where the control u is constrained to lie in a convex, weakly compact subset Ω of the control space with $0 \in \Omega$. A necessary and sufficient condition for a particular initial state to be Ω-null-controllable within a fixed, finite time T is given. The result is extended to the case Ω = convex cone with vertex at 0. Applications to the one-dimensional heat- and wave-equation are given.

Key words. constrained null-controllability, infinite dimensional control problem

AMS(MOS) subject classifications. 93B05, 93C25

1. Introduction. Whereas there is an extensive literature on unconstrained controllability problems, little is known if the control is constrained to take on values in a preassigned subset Ω of the control space U. Classical results in this direction in the finite dimensional case, differing in the assumptions on Ω, can be found in [13], [17], [4]. These assert that controllability is equivalent, in a certain sense, to the famous Kalman rank condition plus an additional one due to the constraint. Only a few years ago W. E. Schmitendorf and B. R. Barmish presented in a series of papers [2], [3], [18] another point of view, which may already be found in a similar form in [11]. They assumed $\Omega \subset R^n$ compact. Their basic argument is as follows: if the origin were not in the reachable set one could find a strictly separating hyperplane. By contraposition and the use of Ekeland's selection theorem [8] they derived a characterization of the domain of Ω-null-controllability which amounts to a finite dimensional optimization problem. This forms the base for various controllability results. Basically, their conditions are integral conditions in terms of the nontrivial solutions of the adjoint equation.

In infinite dimensions there are only a few attacks on the constrained controllability problem, cf. [9], for the boundary control of the heat equation and [10], [15], [16] for the wave equation with distributed controls. In [10] Fattorini applied the theory of sine and cosine operators to show that any state in a suitably defined state space of a system governed by a second order differential equation in a Hilbert space may be controlled to the origin in finite time by means of controls whose values range in the unit ball of the control space. In [15] K. Narukawa presented a theorem which, roughly speaking, asserts that if there is a fixed time $T_0 > 0$, such that an initial state in the controlled space may be steered to the origin within $[0, T_0]$ by means of an unconstrained control (and other hypotheses concerning the solution semigroup are met) it may also be transferred to the origin in finite time T by a control $u(\cdot)$ such that $\|u(\cdot)\|_{p,[0,T]}$ is arbitrarily small. Observe that Narukawa bounds the controls in $L^p_{\text{loc}}(0, \infty; U)$. However, using this theorem, he succeeded in showing global controllability for certain subspaces for the wave equation with respect to appropriate Sobolev spaces even if one applies controls subject to rather arbitrary pointwise constraints [16]. Basically both authors reduce the controllability problem to a moment problem. But, when Ω is different from the unit ball in U, this procedure may cause some difficulties inasmuch as it is by no means clear how certain properties of the control, such as positivity, affect its Fourier coefficients. Therefore we tried, instead, to generalize the ideas of Schmitendorf and Barmish.

* Received by the editors December 4, 1984, and in revised form August 21, 1985.

† Institut für Mathematik, Universität Graz, A-8010 Graz, Austria.

‡ The work of this author was partially supported by Fonds zur Förderung der wissenschaftlichen Forschung Austria, Project S-3206.

Throughout this paper let X be a reflexive, U a separable reflexive Banach space, A be the infinitesimal generator of a C_0-semigroup $S(\cdot)$, $B \in L(U, X)$ and fix $T > 0$. Let X^*, U^* be the respective dual spaces and $\langle \cdot, \cdot \rangle$ be the duality pairing, the involved spaces being clear from the context. We consider the abstract Cauchy problem

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), \qquad t \geqq 0, \\ x(0) &= x_0 \end{aligned} \tag{1.1}$$

where $u(\cdot) \in L^p(0, T; U)$, $1 < p < \infty$. The mild solution of (1.1) is given by

$$x(t) = S(t)x_0 + \int_0^t S(t-s)Bu(s)\, ds, \qquad t \geqq 0.$$

Let Ω be a convex, weakly compact subset of U such that $0 \in \Omega$ and define the set of admissible controls

$$\mathcal{U}_{\text{ad}}(T) = \{u \in L^p(0, T; U) \mid u(t) \in \Omega \text{ a.e. on } [0, T]\}$$

and the corresponding reachable set

$$\mathcal{A}(T, x_0) = \{x(T) \mid x(\cdot) \text{ is a mild solution of (1.1), } x(0) = x_0,\ u \in \mathcal{U}_{\text{ad}}(T)\}.$$

DEFINITION 1.1. (a) $x_0 \in X$ is Ω-null-controllable (within T) iff $0 \in \mathcal{A}(T, x_0)$.

(b) $\{x \mid 0 \in \mathcal{A}(T, x)\}$ will be called the domain of Ω-null-controllability (with respect to T).

Since in infinite dimensions there is in general no analogue to the adjoint equation, we focus our attention on the characterization of the domain of Ω-null-controllability. This is accomplished in Theorem 2.3, which is the straightforward generalization of the corresponding result in finite dimensions [18, Thm. 2.3]. Though this characterization still involves—as should be expected—an infinite dimensional optimization problem, it is shown in the final section that in simple situations it yields results in a most elementary way. It should be pointed out that, in contrast to the aforementioned results, this paper deals with what can be achieved within a fixed (finite) time period. Also, rather than studying Ω-null-controllability of subspaces, our main result enables one to establish (at least theoretically) whether a given particular state may be controlled to the origin in fixed finite time (or not, which is much simpler to verify). For instance, we discuss the one-dimensional heat equation and show that, in case Ω = unitball in U every initial state ϕ with $\sup_{t \in [0,T]} |\phi(t)| < 1$ can be steered to the origin within time $T = 1$. In case $\Omega = [0, 1]$ an initial state ϕ cannot be controlled to the origin as long as the solution to the corresponding homogeneous system is positive on a set of positive measure. We also consider the one-dimensional wave equation and show that for every $T > 0$ the domain of Ω-null-controllability contains a ball with positive radius centered at the origin, and we thus sharpen an earlier result of Fattorini [10].

We would like to emphasize that the assumption $B \in L(U, X)$ rules out the application of our theory to boundary control problems, because in this situation B is typically unbounded. Since the connection between a boundary control problem and a Cauchy problem of the type (1.1) is by no means straightforward, see for example the recent article [6], Narukawa's remark that he also could handle boundary control problems in [15] is not evident (observe, that the proof of Theorem 1 in [15] heavily rests upon the boundedness of B).

Finally, we believe that the assumption $0 \in \Omega$ (but 0 need not be an interior point) is not a severe one from the practical point of view. Otherwise one could incorporate an additional inhomogeneity in the system (1.1).

2. Null controllability with constraints.

PROPOSITION 2.1. *$\mathcal{U}_{ad}(T)$ and $\mathcal{A}(T, x_0)$ are weakly compact.*

Proof. Since Ω is bounded, so is $\mathcal{U}_{ad}(T)$. Thus, by the Eberlein-Smulian theorem, it suffices to show that $\mathcal{U}_{ad}(T)$ is weakly closed. We have to show that the weak limit $u(\cdot)$ of any weakly convergent sequence in $\mathcal{U}_{ad}(T)$ satisfies $u(t) \in \Omega$ a.e. on $[0, T]$. This is followed immediately by a mimicry of the classical proof in [13, p. 157], once we observe that Ω is, by convexity, strongly closed and may therefore be represented as the intersection of a denumerable number of its supporting halfspaces by a theorem of Bishop and Phelps [14, p. 75]. Now, the weak compactness of $\mathcal{A}(T, x_0)$ is clear from

$$\begin{aligned} &\mathcal{T}: L^p(0, T; U) \to X, \\ &\mathcal{T}u = \int_0^T S(T-s)Bu(s)\, ds. \end{aligned} \tag{2.1}$$

Obviously $\mathcal{T} \in L(L^p(0, T; U), X)$ and hence also is continuous with respect to the corresponding weak topologies, which implies the weak compactness of $\mathcal{T}\mathcal{U}_{ad}(T)$. But $\mathcal{A}(T, x_0)$ is merely a translate of $\mathcal{T}\mathcal{U}_{ad}(T)$.

The next proposition gathers some technicalities.

PROPOSITION 2.2. (i) *For each $s \in [0, T]$, $x^* \to \max_{v \in \Omega} \langle x^*, S(s)Bv \rangle$ defines an equicontinuous family of mappings $X^* \to R_+$.*

(ii) *For each $x^* \in X^*$, $s \to \max_{v \in \Omega} \langle x^*, S(s)Bv \rangle$, as a mapping $[0, T] \to R_+$, is continuous.*

(iii) *For each $x^* \in X^*$ there is $u \in \mathcal{U}_{ad}(T)$, such that $\langle x^*, S(\cdot)Bu(\cdot) \rangle = \max_{v \in \Omega} \langle x^*, S(\cdot)Bv \rangle$, where equality is to hold a.e. on $[0, T]$.*

Proof. It should be noted that the maximum above really is obtained on Ω. Hence, for each $x^* \in X^*$, $u^* \in U^*$, $t \in [0, T]$ the following subsets of Ω are nontrivially defined:

$$\begin{aligned} &\Gamma(u^*) = \{\hat{v} \in \Omega \mid \max_{v \in \Omega} \langle u^*, v \rangle = \langle u^*, \hat{v} \rangle\}, \\ &\Gamma(x^*, t) = \{\hat{v} \in \Omega \mid \max_{v \in \Omega} \langle x^*, S(t)Bv \rangle = \langle x^*, S(t)B\hat{v} \rangle\}. \end{aligned} \tag{2.2}$$

First consider the time-independent version of (i), i.e.

$$u^* \to \max_{v \in \Omega} \langle u^*, v \rangle. \tag{2.3}$$

Choose u_0^* and any sequence $u_n^* \in U^*$, so that $\lim_{n \to \infty} u_n^* = u_0^*$ and correspondingly select $v_i \in \Gamma(u_i^*)$, $i = 0, 1, \cdots$. By weak compactness of Ω we may extract a weakly convergent subsequence $v_{i_k} \xrightarrow[k \to \infty]{} \tilde{v} \in \Omega$, which implies

$$\lim_{k \to \infty} \langle u_{n_k}^*, v_{n_k} \rangle = \langle u_0^*, \tilde{v} \rangle.$$

Now assume

$$\langle u_0^*, \tilde{v} \rangle < \langle u_0^*, v_0 \rangle.$$

Then by

$$\langle u_0^*, v_0 \rangle = \lim_{k \to \infty} \langle u_{n_k}^*, v_0 \rangle \leqq \lim_{k \to \infty} \langle u_{n_k}^*, v_{n_k} \rangle = \langle u_0^*, \tilde{v} \rangle$$

we arrive at a contradiction, which shows $\tilde{v} \in \Gamma(u_0^*)$, and hence (2.3) is continuous. From this it is easy to prove (i). Next fix $x^* \in X^*$, $t^* \in [0, T]$ and choose any sequence

t_n, $v_n \in \Gamma(x^*, t_n)$, such that $\lim_{k\to\infty} t_k = t^*$ and without loss of generality $w - \lim v_n = v_0 \in \Omega$.

$$\left|\max_{v\in\Omega} \langle x^*, S(t_n)Bv\rangle - \langle x^*, S(t^*)Bv_0\rangle\right| \leqq |\langle (S^*(t_n) - S^*(t^*))x^*, Bv_n\rangle| + |\langle x^*, S(t^*)B(v_n - v_0)\rangle|. \tag{2.4}$$

Since $S^*(\cdot)$ is a C_0-semigroup on X^* and Ω is bounded, the right-hand side of (2.4) tends to zero. As in part (i) we may show that $v_0 \in \Gamma(x^*, t^*)$, which completes the proof of (ii).

As to the last statement, we have to resort to a measurable selection theorem. For the reader's convenience we present here one which is perfectly suited for our purposes

MEASURABLE SELECTION THEOREM [1]. *Let U be a separable reflexive Banach space, $WC(U)$ be the class of nonempty weakly compact subsets of U, K a compact metric space with finite Lebesgue measure, and let $\Gamma: K \to WC(U)$ be such that*

(1) $\bigcup_{t\in K} \Gamma(t)$ *is bounded in U,*

(2) *for any sequence $\{t_i\}$, $t^* \in K$, $\lim_{i\to\infty} t_i = t^*$ one has*

$$\bigcap_{n=1}^{\infty} \overline{\bigcup_{i=n}^{\infty} \Gamma(t_i)}^{w} \subset \Gamma(t^*),$$

where $^{-w}$ denotes weak closure. Then there exists a strongly measurable selection $u: K \to U$ such that $u(t) \in \Gamma(t)$ a.e. on K.

Define for fixed $x^* \in X^*$ by (2.2) the set-valued mapping $\Gamma: [0, T] \to 2^U$, $\Gamma(t) = \Gamma(x^*, t)$. Clearly $\Gamma(t) \in WC(U)$ for $t \in [0, T]$ and $\bigcup_{t\in[0,T]} \Gamma(t) \subset \Omega$. Hence it remains to show (2). For this purpose assume that for some sequence $t_i \to t^*$

$$\hat{v} \in \bigcap_{n=1}^{\infty} \overline{\bigcup_{i=n}^{\infty} \Gamma(t_i)}^{w} \quad \text{and} \quad \hat{v} \notin \Gamma(t^*).$$

Consequently there is some $\rho > 0$, such that

$$\alpha = \max_{v\in\Omega} \langle x^*, S(t^*)Bv\rangle = \langle x^*, S(t^*)B\hat{v}\rangle + \rho.$$

Because of (ii) and the strong continuity of the dual semigroup we can find $N(\rho)$ so that for $n \geqq N(\rho)$ we have

$$\max_{v\in\Omega} \langle x^*, S(t_i)Bv\rangle \geqq \alpha - \rho/3$$

and

$$\|(S^*(t^*) - S^*(t_i))x^*\| \leqq \rho(3M\|B^*\|)^{-1},$$

M being a bound for Ω. Now consider the weak neighborhood of $\hat{v}$ given by

$$N = \{v \mid \langle B^*S^*(t^*)x^*, v - \hat{v}\rangle < \rho/3\}.$$

For any $v_i \in \Gamma(t_i)$, $i \geqq N(\rho)$ we find

$$\begin{aligned} |\langle B^*S^*(t^*)x^*, v_i - \hat{v}\rangle| &\geqq |\langle B^*S^*(t^*)x^*, v_i\rangle| - |\langle B^*S^*(t^*)x^*, \hat{v}\rangle| \\ &\geqq |\langle B^*S^*(t_i)x^*, v_i\rangle| \\ &\quad - |\langle B^*(S^*(t^*) - S^*(t_i))x^*, v_i\rangle| - \alpha + \rho \\ &\geqq \alpha - \rho/3 - \|B^*\| \cdot \|(S^*(t_i) - S^*(t^*))x^*\| \cdot \|v_i\| - \alpha + \rho \\ &\geqq \rho/3, \end{aligned}$$

which shows that $v_i \notin N$ in contradiction to

$$\hat{v} \in \overline{\bigcup_{i \geqq N} \Gamma(t_i)}^{w}.$$

Now we are prepared to present our main theorem, which gives a characterization of the domain of Ω-null-controllability:

THEOREM 2.3. *Let X, U be reflexive Banach spaces with U separable. Let $B \in L(U, X)$, A be the infinitesimal generator of a C_0-semigroup on X and $\Omega \subset U$ be weakly compact and convex with $0 \in \Omega$. Then for each $T > 0$ the following are equivalent:*

(2.5)
$$\text{(i)} \quad 0 \in \mathcal{A}(T, x_0),$$
$$\text{(ii)} \quad \langle x^*, S(T)x_0 \rangle + \int_0^T \max_{v \in \Omega} \langle x^*, S(t)Bv \rangle \, dt \geqq 0 \quad \textit{for all } x^* \in X^*.$$

Remark 2.4. Because of the positive homogeneity of the functional in (2.5), it suffices to check (2.5) for $x^* \in X^*$ with $\|x^*\| \leqq r$ $(= r)$ for some $r > 0$ and in view of Proposition 2.2(ii) we may further reduce to a dense subset of such a ball.

Proof of Theorem 2.3. Though the proof of Theorem 2.3 follows exactly the pattern given in [18], we sketch it here for the reader's convenience. Taking into account the strict separation theorem [14, p. 25], we find by contraposition

$$0 \in \mathcal{A}(T, x_0) \quad \text{iff } \sup \{x^*(x) \mid x \in \mathcal{A}(T, x_0)\} \geqq 0 \quad \text{for all } x^* \in X^*.$$

This amounts to

$$0 \leqq \langle x^*, S(T)x_0 \rangle + \sup \{\langle x^*, \mathcal{T}u \rangle \mid u \in \mathcal{U}_{ad}(T)\}$$

where $\mathcal{T}$ is the operator defined in (2.1). According to Proposition 2.1 we may find $u_n \in \mathcal{U}_{ad}(T)$, $n = 0, 1, \cdots$, so that

$$w - \lim_{n \to \infty} u_n = u_0,$$

$$\lim_{n \to \infty} \langle x^*, \mathcal{T}u_n \rangle = \sup \{\langle x^*, \mathcal{T}u \rangle \mid u \in \mathcal{U}_{ad}(T)\} \equiv \gamma.$$

Assuming

$$\gamma < \int_0^T \max_{v \in \Omega} \langle x^*, S(T-s)Bv \rangle \, ds,$$

we immediately arrive at a contradiction since by Proposition 2.2(iii) there is $u_0 \in \mathcal{U}_{ad}(T)$ with

$$\int_0^T \max_{v \in \Omega} \langle x^*, S(T-s)Bv \rangle \, ds = \langle x^*, \mathcal{T}u_0 \rangle.$$

We easily derive the following corollaries:

COROLLARY 2.5. *Choose $\Omega = \{u \mid \|u\| \leqq 1\} \equiv B_U$.*

Then the domain of Ω-null-controllability (at time T) contains a ball centered at the origin with radius γ iff for all $x^* \in X^*$

$$\|B^*S^*(\cdot)x^*\|_{L^1(0,T;U^*)} \geqq \gamma \|S^*(T)x^*\|. \tag{2.6}$$

Inequality (2.6) strongly resembles the well-known condition for exact controllability given in [5]. The difference in the norms is due to the fact that Curtain and Pritchard work with unconstrained controls in $L^p(0, T; U)$, whereas in our case the constraints imply that actually $u \in L^\infty(0, T; U)$; $u \in L^2(0, T; U)$ was assumed merely for technical reasons.

COROLLARY 2.6. *Let X be a Hilbert space, $S(\cdot)$ be a unitary group and $\Omega = B_U$. Then no initial datum with $\|x\| > T\|B\|$ can be steered to the origin within time T.*

COROLLARY 2.7. *Let $S(\cdot)$ be a group on X.*

Then necessary for global Ω-(null)-controllability is

$$V(x^*) = \int_{-\infty}^{0} \max_{v\in\Omega} \langle B^*S^*(t)x^*, v\rangle \, dt = \infty \quad \text{for all } x^* \in X^*\backslash\{0\} \tag{2.7}$$

and sufficient is

$$W = \sup_{t\geqq 0} \inf_{\|x^*\|=1} \int_{-t}^{0} \max_{v\in\Omega} \langle B^*S^*(s)x^*, v\rangle \, ds = \infty. \tag{2.8}$$

Equations (2.7) and (2.7) are the natural generalizations of the corresponding conditions in the finite dimensional case. We will give some comments on them in the final section. The proof of Corollary 2.7 is exactly the same as in finite dimensions [2].

Though the boundedness of Ω was crucial in the derivation of Theorem 2.3, we can get a similar result when Ω is a closed convex cone in U. This seems to be new even in finite dimensions. However, in this case $\mathcal{A}(T, x_0)$ need not be closed, which can be seen by the following simple example.

Example 2.8. Choose $X = R^2 = U$ and $\Omega = \{(u, v) \mid u \geqq 0,\ v \geqq 0\}$ and consider

$$\begin{aligned} \dot{x}_1 &= x_2 + u, & x_1(0) &= 0, \\ \dot{x}_2 &= -x_1 + v, & x_2(0) &= 0. \end{aligned} \tag{2.9}$$

We will show

$$\mathcal{A}(\pi/2, 0) = \{(x_1, x_2) \mid x_1 > 0,\ x_2 \in R\} \cup \{(0, 0)\}.$$

Obviously any $(0, x_2)$, $x_2 \in R\backslash\{0\}$, is not reachable in time $\pi/2$, any (x_1, x_2), $x_1 \geqq x_2 > 0$ may be reached by the constant control

$$\begin{pmatrix} u(\cdot) \\ v(\cdot) \end{pmatrix} = \frac{1}{2}\begin{pmatrix} x_1 - x_2 \\ x_1 + x_2 \end{pmatrix}$$

and for any (x_1, x_2), $0 < x_1 < x_2$, a successful control is given by

$$u(s) = \begin{cases} \alpha \sin s, & s \in [0, t^*), \\ \beta \cos s, & s \in [t^*, \pi/2], \end{cases} \qquad v(s) = \begin{cases} \alpha \cos s, & s \in [0, t^*), \\ \beta \sin s, & s \in [t^*, \pi/2] \end{cases}$$

with

$$\begin{aligned} \beta &= 2x_2 \cdot (\sin 2t^*)^{-1}, \\ \alpha &= t^{*-1} \cdot (x_1 - x_2(1 + \cos 2t^*)(\sin 2t^*)^{-1}) \end{aligned}$$

where $t^* \in (\pi/4, \pi/2)$ is chosen so that $\alpha > 0$. Similar controls with u and v interchanged work for $x_2 < 0$.

This suggests that in the following we replace $\mathcal{A}(T, x_0)$ by its closure. Next we recall the notion of the dual cone Ω^*

$$\Omega^* = \{x^* \in X^* \mid x^*(x) \geqq 0,\ x \in \Omega\}.$$

THEOREM 2.9. *Let Ω be a closed convex cone in U with apex at 0 and let the hypotheses of Theorem 2.3 hold. Then for $x_0 \in X$ and $T > 0$ the following are equivalent:*

(i) $0 \in \overline{\mathcal{A}(T, x_0)}$,

(ii) $\langle S^*(T)x^*, x_0\rangle \geqq 0$ *for all* $x^* \in \{x^* \in X^* \mid B^*S^*(t)x^* \in -\Omega^*$ *for all* $t \in [0, T]\}$. (2.10)

Proof. As in the proof of Theorem 2.3, we find

$$0 \in \overline{\mathcal{A}(T; x_0)}$$

iff for any $x^* \in X^*$

$$\langle x^*, S(T)x_0\rangle + \sup\left\{\int_0^T \langle x^*, S(t)Bu(t)\rangle\, dt \,\middle|\, u \in \mathcal{U}_{\mathrm{ad}}(T)\right\} \geqq 0. \tag{2.11}$$

Define

$$\gamma(x^*, T) = \sup\left\{\int_0^T \langle x^*, S(t)Bu(t)\rangle\, dt \,\middle|\, u \in \mathcal{U}_{\mathrm{ad}}(T)\right\}.$$

Since $\mathcal{U}_{\mathrm{ad}}(T)$ is a cone with vertex at 0, we readily see

$$\gamma(X^*, T) = \{0, \infty\}.$$

However, $\langle x^*, S(T)x_0\rangle$ does not matter for

$$x^* \in \gamma^{-1}(\infty),$$

so we may rewrite (2.11) as

$$x^* \in \gamma^{-1}(0) \Rightarrow \langle x^*, S(T)x_0\rangle \geqq 0.$$

Now we define, for $M > 0$,

$$\Omega_M = \Omega \cap \{u \,|\, \|u\| \leqq M\},$$

$$\mathcal{U}_{\mathrm{ad},M}(T) = \{u \in \mathcal{U}_{\mathrm{ad}}(T) \,|\, u(t) \in \Omega_M \text{ a.e. on } [0, T]\}.$$

It is easy to see that

$$x^* \in \gamma^{-1}(0) \quad \text{iff} \quad \sup_{M>0} \sup\left\{\int_0^T \langle x^*, S(t)Bu(t)\rangle\, dt \,\middle|\, u \in \mathcal{U}_{\mathrm{ad},M}(T)\right\} = 0. \tag{2.12}$$

But Ω_M is weakly compact and hence we infer, as for Theorem 2.3,

$$\sup\left\{\int_0^T \langle x^*, S(t)Bu(t)\rangle\, dt \,\middle|\, u \in \mathcal{U}_{\mathrm{ad},M}(T)\right\} = \int_0^T \max_{v \in \Omega_M} \langle x^*, S(t)Bv\rangle\, dt.$$

Therefore, by Proposition 2.2(ii), (2.12) is equivalent to

$$\max_{v \in \Omega_M} \langle x^*, S(t)Bv\rangle = 0 \quad \text{for all } M > 0 \text{ and for all } t \in [0, T]$$

or, equivalently, to

$$\sup_{v \in \Omega} \langle x^*, S(t)Bv\rangle = 0 \quad \text{for all } t \in [0, T].$$

This may be interpreted as

$$B^*S^*(t)x^* \in -\Omega^* \quad \text{for all } t \in [0, T].$$

Remark 2.10. In case $\Omega = U$ and consequently $\Omega^* = \{0\}$, we immediately read off the familiar result that, if $\overline{\mathrm{Im}\, B} = X$, any initial state may be transferred arbitrarily near zero in arbitrarily short time. If, furthermore, we assume that $S(\cdot)$ is a C_0-group, then from

$$\{x^* \in X^* \cap D(A^*) \,|\, B^*S^*(t)x^* = 0,\ t \in [0, T]\} = \bigcap_{n=0}^{\infty} \ker\,(B^*(A^*)^n)$$

we easily recover the well-known characterization of approximate controllability of R. Triggiani [19].

3. Examples. In this section we will show that, although (2.5) represents an infinite dimensional optimization problem, it may in simple cases yield a characterization of the domain of Ω-null-controllability in a most elementary way.

Example 3.1. We choose $X = U = l^2$ and consider

$$\dot{x} = x + u, \tag{3.1}$$

so

$$S(t)x_0 = e^t x_0.$$

First we take $\Omega_1 = B_U$. We easily see that x may be steered to the origin in finite time iff $\|x\| < 1$. Furthermore, in this case

$$V(x^*) = \|x^*\|, \qquad W = 1.$$

If we replace Ω_1 by

$$\Omega_2 = \{u \in l^2 \mid \|u\|_2 \leqq 1,\ u = (\omega_i),\ 0 \leqq \omega_i,\ i = 1, 2, \cdots\},$$

an easy calculation based on (2.5) yields x_0 controllable to the origin iff $\|x_0\| < 1$ and $-1 < \xi_i \leqq 0$ for $x = (\xi_i)$, which perfectly agrees with the scalar case [18]. It is easily seen that now

$$V(x^*) = \|x_+^*\|, \qquad W = 0,$$

x_+^* being the positive part of x^*. Finally let us consider $u_0 = (\eta_i) \in l^2$, $\eta_i \neq 0$, $i = 1, \cdots$,

$$\Omega_3 = \{u \in l^2 \mid u = (\omega_i),\ |\omega_i| \leqq |\eta_i|\}.$$

From (2.5) it is readily verified that $x_0 = 2u_0$ may not be controlled to zero in finite time and, accordingly, a simple calculation results in

$$V(x^*) = \sum_{i=1}^{\infty} |\xi_i^*||\eta_i| < \infty, \qquad W = 0$$

and again $V(x^*) \neq W$. If we modify (3.1) slightly to get an exponentially stable semigroup, i.e.

$$\dot{x} = -x + u, \tag{3.1'}$$

then in case $\Omega = \Omega_1$, (3.1') is globally Ω_1-null-controllable and

$$W = V(x^*) = \infty.$$

In case $\Omega = \Omega_2$, one has

$$V(x^*) = \begin{cases} 0, & x_+^* = 0, \\ \infty, & \text{else}, \end{cases} \qquad W = 0.$$

Hence the system is not globally Ω_2-null-controllable. Finally, if $\Omega = \Omega_3$ one has

$$V(x^*) = \infty, \qquad W = 0.$$

These simple examples show that (2.7) and (2.8) in general do not coincide and even that a controllability gap may arise, i.e. that the necessary but not the sufficient condition applies. Remember that in the finite dimensional case it can be shown that for linear

autonomous systems and $0 \in \Omega$ the controllability gap does not occur [3]. Its occurrence here is somewhat surprising since (3.1) is the simplest extension of the analogue scalar equation. Example (3.1) may seem somewhat artificial and therefore we present one closely linked to the one-dimensional wave equation which also shows this strange behavior.

Example 3.2. Choose $X = L^2(R)xL^2(R)$, $U = L^2(R)$, $\Omega = B_U$ and consider

$$\begin{aligned} &v_t = w_x + u(t, x), && t \geqq 0, \\ &w_t = v_x, \\ &v(x, 0) = \alpha(x), \\ &w(x, 0) = \beta(x), && x \in R. \end{aligned} \tag{3.2}$$

It is well known that the solution semigroup is given by

$$S(t)(\alpha, \beta) = \frac{1}{2}\begin{pmatrix} T(t)+T(-t) & T(t)-T(-t) \\ T(t)-T(-t) & T(t)+T(-t) \end{pmatrix}\begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

where $T(\cdot)$ denotes the translation semigroup in $L^2(R)$. For any $x^* = (x_1^*, x_2^*) \in X^* \sim X$, $\|x^*\| = 1$ we find, after some rearrangement,

$$\int_{-t}^{0} \max_{v \in \Omega} \langle x^*, S(\tau)(v, 0)\rangle \, d\tau = \frac{1}{2}\int_{-t}^{0} \| T(-\tau)(x_1^* + x_2^*) + T(\tau)(x_1^* - x_2^*)\| \, d\tau. \tag{3.3}$$

Squaring the integrand in (3.3) and applying the parallelogram law, we find

$$2(1 + \langle T(\tau)(x_1^* - x_2^*), T(-\tau)(x_1^* + x_2^*)\rangle).$$

Hence, in considering the sufficient condition (2.8) we are faced with

$$\begin{aligned} 0 &\leqq \inf_{\|x^*\|=1} \int_{-t}^{0} (1 + \langle T(\tau)(x_1^* - x_2^*), T(-\tau)(x_1^* + x_2^*)\rangle)^{1/2} \, d\tau \\ &\leqq \inf_{\|x^*\|=1} \sqrt{t}\left(\int_{-t}^{0} (1 + \langle T(\tau)(x_1^* - x_2^*), T(-\tau)(x_1^* + x_2^*)\rangle) \, d\tau\right)^{1/2}, \end{aligned} \tag{3.4}$$

the inequality at the utmost left of (3.4) being a consequence of the parallelogram law. This also shows that the integrals in (3.4) are well defined. Choosing

$$x_1^* = \begin{cases} -(2\sqrt{t})^{-1}, & x \in [-t, 0], \\ (2\sqrt{t})^{-1}, & x \in [0, t], \\ 0, & \text{else}, \end{cases} \qquad x_2^* = |x_1^*|,$$

we calculate

$$\int_{-t}^{0} \langle T(\tau)(x_1^* - x_2^*), T(-\tau)(x_1^* + x_2^*)\rangle \, d\tau = -t,$$

so that the right-hand side of (3.4) equals zero and consequently $W = 0$. Thus, the sufficient condition does not apply. As to the necessary condition (2.7), in view of (3.4) we consider for $x^* \in X^*$, $\|x^*\| = 1$,

$$\begin{aligned} &\int_{-\infty}^{0} \| T(-t)(x_1^* + x_2^*) + T(t)(x_1^* - x_2^*)\| \, dt \\ &\qquad \geqq \lim_{t \to \infty}\left(t - \int_{-t}^{0} |\langle T(\tau)(x_1^* + x_2^*), T(-\tau)(x_1^* + x_2^*)\rangle| \, d\tau\right). \end{aligned} \tag{3.5}$$

Let $u, v \in L^2(R)$, $\varepsilon > 0$, and choose $M > 0$ so that

$$\max\left(\int_{-\infty}^{-M} |u|^2\,dx, \int_M^{\infty} |v|^2\,dx\right) \leqq \varepsilon^2(2\|u\|)^{-2}.$$

We find that

$$\begin{aligned}\int_{-t}^{0} |\langle T(\tau)u, T(\tau)v\rangle|\,d\tau \leqq &\int_{-t}^{0}\int_{-\infty}^{-M} |u(x+2\tau)||v(x)|\,dx\,d\tau \\ &+ \int_{-t}^{0}\int_{-M}^{M} |u(x+2\tau)||v(x)|\,dx\,d\tau \\ &+ \int_{-t}^{0}\int_{M}^{\infty} |u(x+2\tau)||v(x)|\,dx\,d\tau.\end{aligned}$$

The first and the last term above are estimated by $\varepsilon/2t$, the middle one by $\sqrt{2Mt}\cdot\|u\|\,\|v\|$. Collecting things together and inserting into (3.5), we find a lower bound for $V(x^*)$:

$$V(x^*) \geqq \tfrac{1}{2}\lim_{t\to\infty}(t - \varepsilon t - \sqrt{t}\sqrt{2M}\,\|x_1^* + x_2^*\|\,\|x_1^* - x_2^*\|) = \infty$$

which shows the occurrence of the controllability gap in this example.

We believe these examples indicate that the corresponding generalizations of the necessary and sufficient conditions for global Ω-null-controllability in [3] are not the appropriate tool to handle this problem. In the next two examples, however, we will show that in case the solution semigroup of (1.1) is fairly well known, one can get a complete characterization of the domain of Ω-null-controllability in a very elementary way.

Example 3.3. Let us consider the one-dimensional heat equation. We take $X = U = L^2(0, 1)$.

$$\begin{aligned} &z_t = z_{xx} + u, && (t, x) \in (0, \infty)\times[0, 1], \\ &z(0, t) = z(1, t) = 0, && t \geqq 0, \\ &z(\cdot, 0) = \phi, && \phi \in X. \end{aligned} \tag{3.6}$$

The solution semigroup is given by

$$(S(t)\phi)(x) = \int_0^1 G(t, x, \xi)\phi(\xi)\,d\xi, \qquad t > 0,$$

where $G(\cdot, \cdot, \cdot)$ is the Green's function given by

$$G(t, x, \xi) = 2\sum_{k=1}^{\infty} e^{-(k\pi)^2 t}\sin k\pi x \cdot \sin k\pi\xi,$$

$$(t, x, \xi) \in (0, \infty)\times[0, 1]\times[0, 1].$$

At first we will discuss the case

$$\Omega_1 = \{v \in U \mid v(x) \in [0, 1] \text{ a.e. on } [0, 1]\}.$$

Choosing $x^* = -1$ it is obvious from (2.5) and the properties of G that no ϕ, with $\phi \geqq 0$ a.e., can be steered to the origin in finite time. This perfectly agrees with one's physical intuition. Assume $\phi \leqq 0$ a.e.; hence $S(t)\phi \leqq 0$ for $t > 0$. W.l.o.g. we may suppose

$-1<\phi\leqq 0$. In view of Proposition 2.2(ii) we may apply the classical mean value theorem to deduce existence of a $t^*\in(0, T)$ such that

$$\int_0^T \max_{v\in\Omega}\langle x^*, S(t)v\rangle\, dt = T\cdot \max_{v\in\Omega}\langle x^*, S(t^*)v\rangle$$
$$= T\cdot\langle S(t^*)x^*, v_0\rangle,$$

where v_0 is the characteristic function of the support of $(S(t^*)x^*)_+$, the positive part of $S(t^*)x^*$. Thus (2.5) is equivalent to

$$\langle S(T)x^*, \phi\rangle + T\cdot\langle S(t^*)x^*, v_0\rangle$$
$$= \langle S(t^*)x^*, S(T-t^*)\phi\rangle + T\cdot\langle S(t^*)x^*, v_0\rangle$$
$$= \langle (S(t^*)x^*)_+, S(T-t^*)\phi + T\cdot v_0\rangle + \langle (S(t^*)x^*)_-, S(T-t^*)\phi\rangle \geqq 0.$$

The last inequality is valid, because

$$S(T-t^*)\phi + Tv_0 \geqq 0$$

on the support of $(S(t^*)x^*)_+$ if $T\geqq 1$. This shows that any ϕ, $-1<\phi\leqq 0$, may be steered to rest in any time $T\geqq 1$. As to the general case, it is easy to deduce from the above that

$$0\notin \bigcup_{t>0}\mathscr{A}(t,\phi) \quad \text{if} \quad \text{meas}\,(E_+(t,\phi))>0, \quad t>0, \tag{3.7}$$

where we have defined

$$E_+(t,\phi)=\{x\in[0,1]\,|\,S(t)\phi(x)\geqq 0\}.$$

Note that meas $(E_+(T,\phi))>0$ implies meas $(E_+(t,\phi))>0$ for all $t\in[0,T]$. Unfortunately it is hard to characterize those ϕ satisfying (3.7), but we may give two readily available sufficient conditions:

$$\text{(a)} \quad \alpha \sin k\pi x \leqq \phi(x) \quad \text{for some } \alpha\in R, k\in N, \tag{3.8}$$

or

$$\text{(b)} \quad \phi_{2k+1}\geqq 0, \quad k=0,1,\cdots, \tag{3.9}$$

where $\phi_k=\int_0^1\phi(x)\sin k\pi x\,dx$. It is easily verified that the latter condition implies $\int_0^1 (S(t)\phi)(x)\,dx\geqq 0$ giving (3.7). Next we discuss the case $\Omega=\Omega_2=B_U$. Again without loss of generality we assume $\|\phi\|\leqq 1$ and, as above, we derive

$$\langle S(T)x^*, \phi\rangle + \int_0^T \|S(t)x^*\|\,dt \geqq -\|S(T)x^*\| + T\cdot\|S(t^*)x^*\|\geqq 0.$$

The last inequality is valid for $T\geqq 1$, which is seen from the representation of the semigroup. Therefore, in this case, (3.6) is globally Ω_2-null-controllable and if $\|\phi\|\leqq 1$, ϕ may be controlled to the origin in any time $T\geqq 1$.

Example 3.4. Consider the one-dimensional wave equation with distributed controls:

$$\begin{aligned} &u_{tt}=u_{xx}+f(t,x), && (x,t)\in[0,1]\times[0,\infty),\\ &u(0,\cdot)=\phi,\\ &u_t(0,\cdot)=\psi,\\ &u(t,0)=u(t,1)=0, && t\geqq 0. \end{aligned} \tag{3.10}$$

Let $W_0^{1,2}(0,1)$ be the usual Sobolev space endowed with the scalar product

$$((\phi,\psi))=\int_0^1 \frac{\partial\phi}{\partial x}\frac{\partial\psi}{\partial x}\,dx$$

and let X be the Hilbert space $W_0^{1,2}(0,1)\times L^2(0,1)$. It is well known [12] that the solution of (3.10) may be described in terms of the unitary group $S(\cdot)$ on X

$$S(t)(\phi,\psi)=(u(\cdot,t),u_t(\cdot,t)),\qquad t\geqq 0,$$

$$u(x,t)=\sum_{k=1}^{\infty}\{(k\pi)^{-1}\psi_k\sin k\pi t+\phi_k\cos k\pi t)\sin k\pi x\},\qquad x\in[0,1],$$

$$\phi_k=2\int_0^1\phi(s)\sin k\pi s\,ds,$$

$$\psi_k=2\int_0^1\psi(s)\sin k\pi s\,ds,\qquad k=1,\cdots.$$

In the following we will sharpen somewhat the result of [10]: we will show, that for fixed (but arbitrary time $T>0$) the domain of Ω-null-controllability contains a ball centered at the origin if controls are taken from B_U. In [10] it is shown that any state in such a ball may be controlled to the origin in finite time but no uniform bound on the consumed time may be given. According to Corollary 2.5 we have to show the existence of $\gamma>0$ such that for $x^*\in X^*$

$$\int_0^T\|B^*S(-t)x^*\|_2\,dt\geqq\gamma\|x^*\| \tag{3.11}$$

holds. In other words, the interior of the domain of Ω-null-controllability is nonempty iff

$$G: X^*\to L^1(0,T;L^2(0,1))$$

defined by

$$(Gx^*)(t)=B^*S(-t)x^*,\qquad t\in[0,T]$$

is continuously invertible or, equivalently (since G turns out to be 1-1), iff the range of G is closed in $L^1(0,T;L^2(0,1))$. Actually G maps into $C(0,T;L^2(0,1))$. If $z\in L^1(0,T;L^2(0,1))$ we may expand $z(t)$

$$z(t)=\sum_{k=1}^{\infty}z_k(t)\sin k\pi x$$

and

$$(z_k(\cdot))\in L^1(0,T;l^2).$$

Hence, for each $z=G(\sigma,\rho)$ the following identity holds:

$$z_k(t)=-\rho_k\cos k\pi t-\sigma_k k\pi\sin k\pi t. \tag{3.12}$$

Case 1. $T\geqq 1$. Now observe that there are k disjoint intervals A_{jk}, B_{jk}, $j=0,1,\cdots,k-1$, of length $(2k)^{-1}$ on $[0,1]$ on which $\sin k\pi t$, $\cos k\pi t$ do not change sign and such that

$$\left|\int_{A_{jk}}\cos k\pi x\,dx\right|=\sqrt{2}(k\pi)^{-1}=\left|\int_{B_{jk}}\sin k\pi x\,dx\right|,$$

$$\int_{A_{jk}}\sin k\pi x\,dx=0=\int_{B_{jk}}\cos k\pi x\,dx,\qquad j=0,\cdots,k-1.$$

Hence we may recover the Fourier coefficients of $x^* = (\sigma, \rho)$ from $z = Gx^*$ via

$$(3.13) \qquad \int_{A_{0k}} z_k(t)\,dt = \sqrt{2}(k\pi)^{-1}\rho_k, \qquad \int_{B_{0k}} z_k(t)\,dt = \sqrt{2}\sigma_k, \qquad k = 2, \cdots,$$

σ_1, ρ_1 may be calculated by integrating (3.12) over $[0, 1]$ and $[0, \frac{1}{2}]$, respectively. Adding the above integrals leads to

$$k|\sigma_k| \leqq \frac{\sqrt{2}}{2}\int_0^1 |z_k(t)|\,dt, \qquad |\rho_k| \leqq \frac{\sqrt{2}}{2}\int_0^1 |z_k(t)|\,dt, \quad k = 1, 2, \cdots,$$

and hence we infer that G is injective and that

$$\sum_{k=1}^{\infty} (|\rho_k|^2 + k^2|\sigma_k|^2) \leqq \tfrac{1}{2}(1+\pi^2)\int \sum_{k=1}^{\infty} |z_k(t)|^2\,dt,$$

which is finite since $z(\cdot) \in C(0, T; L^2(0, 1))$. Hence (σ, ρ), represented by (3.13), belongs to X. By Theorem VI.6.5 in [7], it suffices to show that G maps bounded closed sets into closed sets. Thus, let A be any bounded closed set in X, i.e.

$$(3.14) \qquad \sum_{k=1}^{\infty} (k^2|\sigma_k|^2 + |\rho_k|^2) \leqq M^2 \quad \text{for } (\sigma, \rho) \in A$$

for some $M > 0$, and let $z_n(\cdot)$ be a convergent sequence in GA. There exist (uniquely determined) $(\sigma_n, \rho_n) \in A$, such that

$$z_n(\cdot) = G(\sigma_n, \rho_n) \quad \text{and} \quad \lim_{n\to\infty} z_n(\cdot) = z(\cdot).$$

Let $z_k^n(t)$, σ_k^n, ρ_k^n denote the corresponding Fourier coefficients. By (3.14)

$$\sum_{k=1}^{\infty} |z_k(t)|^2 \leqq 2M^2 \quad \text{for } z \in GA$$

so that

$$(3.15) \qquad \left((2M)^{-2}\sum_{k=1}^{\infty} (z_k^n(t) - z_k^m(t))^2\right)^{1/2} \leqq 1$$

for every choice of n, m. Since $z_n(\cdot)$ is a Cauchy sequence in $L^1(0, T; L^2(0, 1))$, $(2M)^{-1}z_n(\cdot)$ also is. Consequently,

$$(2M)^{-1}\|z_n - z_m\| = \int_0^T \left((2M)^{-2}\sum_{k=1}^{\infty} (z_k^n(t) - z_k^m(t))^2\right)^{1/2} dt$$

$$\geqq (2M)^{-1}\sum_{k=1}^{\infty}\int_0^1 (z_k^n(t) - z_k^m(t))^2\,dt$$

(by (3.15)). However, by (3.12) we calculate

$$\int_0^1 (z_k^n(t) - z_k^m(t))^2\,dt = (\rho_k^n - \rho_k^m)^2 + k^2\pi^2(\sigma_k^n - \sigma_k^m)^2,$$

from which we deduce that (σ_n, ρ_n) is a Cauchy sequence in X. Therefore, there exists $(\sigma, \rho) \in X$ such that

$$\lim_{n\to\infty} (\sigma_n, \rho_n) = (\sigma, \rho)$$

and, by closedness of A, $(\sigma, \rho) \in A$. By continuity of G we conclude that

$$\lim_{n\to\infty} G(\sigma_n, \rho_n) = G(\sigma, \rho)$$

and by the uniqueness of the limit we infer

$$z(\cdot) = G(\sigma, \rho) \in GA.$$

Case 2. $0 < T < 1$. Choose n such that $n^{-1} < T$. In this case (σ_k, ρ_k), $k = 1, 2, \cdots, n$ may be determined by integrating (3.12) over $[0, 1/n]$ and $[0, 1/2n]$, respectively. (σ_k, ρ_k), $k > n$ are given by (3.13). Writing $k = n + m$ the interval $[0, 1/n]$ contains $[mn^{-1}]$ intervals A_{jk}, B_{jk}, where $[\cdot]$ denotes the largest integer function. Now the result follows from the preceding proof once we observe that

$$k[mn^{-1}]^{-1} \leqq 3n \quad \text{for all } k \geqq 2n.$$

Finally we want to present an example which illustrates the use of Theorem 2.9.

Example 3.5. Again consider the one-dimensional heat equation (3.6) but choose $X = U = L^2(R, R)$. In this case the solution semigroup is given by

$$(S(t)\phi)(x) = (4\pi t)^{-1/2} \int_R \exp(-(s-x)^2/4t)\phi(s)\, ds, \qquad t > 0.$$

Let

$$\Omega = \{u \in U \mid u(x) \geqq 0 \text{ a.e.}\}.$$

It is easily seen that

$$\Omega^* = \Omega.$$

Furthermore, we find that

$$S(t)^* x^* \in -\Omega^* \quad \text{for all } t \in [0, T] \quad \text{iff } x^* \in -\Omega^* = -\Omega.$$

This shows

$$\begin{aligned} x_0 \in \overline{\mathscr{A}(T, x_0)} \quad &\text{iff} \quad S(T)x_0 \in (-\Omega^*)^*, \\ &\text{iff} \quad S(T)x_0 \leqq 0, \end{aligned}$$

which is equivalent to (3.7).

Acknowledgment. We thank Prof. O. Carja (University of Iasi) for bringing to our attention his paper *Local controllability of nonlinear evolution equations in Banach spaces*, Anal. St. Univ. Al. I Cuza, Iasi, 25 (1979), pp. 117-125, which led to a considerable improvement of the original statement in Example 3.4.

REFERENCES

[1] N. U. AHMED AND K. L. TEO, *Optimal Control of Distributed Parameter Systems*, North-Holland, Amsterdam, 1981.

[2] B. R. BARMISH AND W. E. SCHMITENDORF, *A necessary and sufficient condition for local constrained controllability of a linear system*, IEEE Trans. Automat. Control, AC-25 (1980), pp. 97-100.

[3] ———, *New results on controllability of systems of the form* $\dot{x}(t) = A(t)x(t) + f(t, u(t))$, IEEE Trans. Automat. Control, AC-25 (1980), pp. 540-547.

[4] R. BRAMMER, *Controllability of linear autonomous systems with positive controls*, this Journal, 10 (1972), pp. 339-353.

[5] R. C. CURTAIN AND A. J. PRITCHARD, *Infinite Dimensional Linear Systems Theory*, Lecture Notes Control and Information Sciences 8, Springer, Berlin, 1978.

[6] W. DESCH, I. LASIECKA AND W. SCHAPPACHER, *Feedback boundary control problems for linear semigroups*, Israel J. Math., 51 (1985), pp. 177-207.

[7] N. DUNFORD AND J. SCHWARTZ, *Linear Operators*, I, Interscience, New York, 1966.

[8] I. EKELAND AND R. TEMAN, *Convex Analysis and Variational Problems*, North-Holland, Amsterdam, 1976.

[9] H. O. FATTORINI, *Boundary control of temperature distributions in a parallelepipedon*, this Journal, 13 (1975), pp. 1-13.

[10] ———, *The time optimal problem for distributed control of systems described by the wave equation*, in Control Theory of Systems Governed by Partial Differential Equations, A. K. Aziz, J. W. Wingate and A. J. Balas, eds., Academic Press, New York, 1977.

[11] R. GABASOV AND F. KIRILLOVA, *The Qualitative Theory of Optimal Processes*, Marcel Dekker, New York, 1976.

[12] F. KAPPEL AND W. SCHAPPACHER, *C_0-semigroups and where they come from*, preprint 9, Graz, 1982.

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

[14] J. MARTI, *Konvexe Analysis*, Birkhäuser, Basel, 1977.

[15] K. NARUKAWA, *Admissible null controllability and optimal time control*, Hiroshima Math. J., 11 (1981), pp. 533-551.

[16] ———, *Admissible controllability of vibrating systems with constrained controls*, this Journal, 20 (1982), pp. 770-782.

[17] ST. SAPERSTONE, *Global controllability of linear systems with positive controls*, this Journal, 11 (1973), pp. 417-423.

[18] W. E. SCHMITENDORF AND B. R. BARMISH, *Null controllability of linear systems with constrained Controls*, this Journal, 18 (1980), pp. 327-345.

[19] R. TRIGGIANI, *Controllability and observability in Banach spaces with bounded operators*, this Journal, 13 (1975), pp. 462-491.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 6, November 1986

012

FREQUENCY/PERIOD ESTIMATION AND ADAPTIVE REJECTION OF PERIODIC DISTURBANCES*

D. L. RUSSELL†

Abstract. We discuss a method for suppressing the oscillations of a linear system subject to an external periodic disturbance of fixed, but unknown, period. The method entails augmentation of the original plant with a compensator and parameter identifier. The near equilibrium dynamics of the resulting system are analyzed and shown to be related to a linear delay equation with infinite delay and periodic coefficients. A corresponding Floquet theory is indicated. A FORTRAN program approximately realizing the period identifier is included and numerical results obtained with this program are graphically displayed and analyzed.

Key words. adaptive control, decoupling, delay equations

AMS(MOS) subject classifications. 93B30, B40, C40, 45A05, D05

0. Introduction. In a wide variety of applications one encounters a system of the form

$$\dot{x} = Ax + Cu + v, \qquad \left(\dot{x} = \frac{dx}{dt}\right) \tag{0.1}$$

wherein x is the n-dimensional state vector, u is the m-dimensional control vector and v is a periodic n-dimensional vector disturbance function with least positive period T

$$v(t) = v(t+T). \tag{0.2}$$

In many cases $\dot{x} = Ax$ by itself represents the dynamics of an elastic system, the disturbance v arises from the environment in which the elastic system is placed, and the control u is used to mitigate the effects of this disturbance. Examples include sighting devices (cameras, telescopes, etc.), weapons, and machine tool arms, operated under conditions which involve significant oscillatory disturbances, such as would be the case for a telescope operated from an aircraft. Another important application arises in connection with the measurement and active suppression of aerodynamic flutter in aircraft wings, tail structures, etc.

The approach taken in this paper is to suppose that $v(t)$ can be adequately modelled by

$$v(t) = Bz(t), \qquad Bn \times 2t, \tag{0.3}$$

where $z(t)$ satisfies a linear system

$$\dot{z} = Fz, \tag{0.4}$$

* Received by the editors January 17, 1985, and in revised form June 29, 1985. This work was sponsored in part by the Air Force Office of Scientific Research under grant 84-0088 and in part by the Army Research Office under contract DAAG29-80-C-0041.

† Mathematics Research Center and Department of Mathematics, University of Wisconsin, Madison, Wisconsin 53706.

$$F=\begin{pmatrix} 0 & \alpha_1 & 0 & 0 & \cdots & 0 & 0 \\ -\alpha_1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \alpha_2 & \cdots & 0 & 0 \\ 0 & 0 & -\alpha_2 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & \alpha_r \\ 0 & 0 & 0 & 0 & \cdots & -\alpha_r & 0 \end{pmatrix}, \tag{0.5}$$

$$\alpha_j=\frac{2k_j\pi}{T}\equiv k_j\alpha, \qquad j=1,2,\cdots,r, \tag{0.6}$$

the k_j being positive integers. These need not necessarily be $1, 2, \cdots, r$; in some cases it is known, for example, that only odd order harmonics occur so that we would use $k_1=1$, $k_2=3, \cdots, k_r=2r-1$.

Assuming F known (this will bring us to the subject of frequency estimation later on), we can construct a compensator

$$\dot{y}=Sx+Fy \tag{0.7}$$

where y is the $2r$-dimensional compensator state, and consider the combined system

$$\begin{aligned} \dot{x}&=Ax+Bz+Cu, \\ \dot{y}&=Sx+Fy, \\ \dot{z}&=Fz. \end{aligned} \tag{0.8}$$

We will suppose that the range of C includes the range of B. This means that, in principle, one could solve

$$Cu=-Bz \tag{0.9}$$

and cancel the effect of the disturbance altogether. For a telescope operated from a moving vehicle, neglecting translational motion and considering only the angular displacements, this would be the case if the controls, acting through the mounting, have both azimuth and elevation correction capability. In practice the direct cancellation (0.9) is rarely feasible due to noise, measurement delays, limited measurement instrumentation, etc.

Let Φ be any nonsingular $2r\times 2r$ matrix which commutes with F; in most cases we would use the identity matrix. We may then find an $m\times 2r$ matrix L such that

$$CL=-B\Phi^{-1}. \tag{0.10}$$

Assuming additionally that (A, C) is stabilizable, let K be an $m\times n$ matrix such that $A+CK$ is a stability matrix and let u be generated by the feedback control law

$$u=Kx+Ly. \tag{0.11}$$

Using this in (0.8) we have

$$\frac{d}{dt}\begin{pmatrix} x \\ y \\ z \end{pmatrix}=\begin{pmatrix} A+CK & -B\Phi^{-1} & B \\ S & F & 0 \\ 0 & 0 & F \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix}. \tag{0.12}$$

We will see in § 1 that it is possible to select S in such a way that the control law (0.11) dynamically decouples the plant state x from the periodic disturbance $v(t)=Bz(t)$.

The foregoing scheme, to be developed more fully in the next section, clearly amounts to the construction of a reduced order observer for the disturbance state $z(t)$ (see [10]) and assumes that the plant state $x(t)$ is completely accessible. If this is not the case, dynamic decoupling is probably best realized with the construction of a full $(n+2r)$-dimensional state observer. Assuming an observation

$$w = H_0 x + H_1 z \tag{0.13}$$

is available such that the pair

$$(H_0, H_1)\begin{pmatrix} A+CK & B \\ 0 & F \end{pmatrix} \tag{0.14}$$

is observable, compatible matrices L_0, L_1 are selected (see, for example, [8]) such that

$$\begin{pmatrix} A+CK-L_0H_0 & B-L_0H_1 \\ -L_1H_0 & F-L_1H_1 \end{pmatrix} \tag{0.15}$$

is a stability matrix. We then adjoin to the plant disturbance system

$$\dot{x} = (A+CK)x + Bz + Cu, \tag{0.16}$$

$$\dot{z} = Fz, \tag{0.17}$$

the estimator system

$$\begin{aligned} \dot{\xi} &= (A+CK)\xi + L_0(H_0x - H_0\xi) + L_0(H_1z - H_1\zeta), \\ \dot{\zeta} &= Fz + L_1(H_0x - H_0\xi) + L_1(H_1z - H_1\zeta). \end{aligned} \tag{0.18}$$

Then, choosing u such that

$$Cu = -B\xi \tag{0.19}$$

and letting $e = x - \xi$, $f = z - \zeta$ we find that

$$\begin{pmatrix} \dot{e} \\ \dot{f} \end{pmatrix} = \begin{pmatrix} A+CK-L_0H_0 & B-L_0H_1 \\ -L_1H_0 & F-L_1H_1 \end{pmatrix}\begin{pmatrix} e \\ f \end{pmatrix}$$

and we conclude, since (0.15) is a stability matrix, that

$$\lim_{t\to\infty} e(t) = \lim_{t\to\infty} f(t) = 0.$$

Since, with (0.19), (0.16), (0.17) become

$$\dot{x} = (A+CK)x + Bf, \tag{0.20}$$

$$\dot{z} = Fz, \tag{0.21}$$

we conclude that

$$\lim_{t\to\infty} x(t) = 0$$

and thus $x(t)$ is decoupled from z. This is a standard procedure, such as described in [10], for example.

The subject of dynamic decoupling, which includes our rather specialized topic, has an extensive literature. Major current references are [10] and [11]; a large number

of additional references are at the end of Chapter 9 in [10] and at the end of Chapter 8 in [11]. In general one considers a system x, A, B, C not necessarily the same as earlier in this section

$$\dot{x} = \begin{pmatrix} \dot{x}^1 \\ \dot{x}^2 \\ \vdots \\ \dot{x}^r \end{pmatrix} = \begin{pmatrix} A_1^1 & A_2^1 & \cdots & A_r^1 \\ A_1^2 & A_2^2 & \cdots & A_r^2 \\ \vdots & \vdots & & \vdots \\ A_1^r & A_2^r & \cdots & A_r^r \end{pmatrix} \begin{pmatrix} x^1 \\ x^2 \\ \vdots \\ x^r \end{pmatrix} + \begin{pmatrix} B_1^1 & B_2^1 & \cdots & B_\rho^1 \\ B_1^2 & B_2^2 & \cdots & B_\rho^2 \\ \vdots & \vdots & & \vdots \\ B_1^\rho & B_2^\rho & \cdots & B_\rho^\rho \end{pmatrix} \begin{pmatrix} v^1 \\ v^2 \\ \vdots \\ v^\rho \end{pmatrix} + Cu = Ax + Bv + Cu. \tag{0.22}$$

It is required to find a state feedback relation $u = Kx$ such that, assuming $r \leqq \rho, x^1, \cdots, x^{r-\rho}$ are decoupled from $v^1, v^2, \cdots, v^\rho$ and for $k = r-\rho+1, \cdots, r$, only the output x^k is dynamically coupled to v^k. This means that the closed loop transfer matrix for (0.22), i.e.,

$$T(\lambda) \equiv (\lambda I - (A + CK))^{-1} B$$

is block diagonal with $T_k^k(\lambda) = 0$, $k = 1, 2, \cdots, r-\rho$, $T_k^k(\lambda)$ possibly nonzero for $k = r-\rho+1, \cdots, r$, $T_k^j = 0, j \neq k$. Our problem can be cast in this form with x^1, x^2, x^3 replaced by x, y, z and A replaced by

$$\begin{pmatrix} A & 0 & B \\ S & F & 0 \\ 0 & 0 & F \end{pmatrix} \quad (A, B \text{ here referring to the matrices in (0.8))},$$

B replaced by

$$\begin{pmatrix} 0 \\ 0 \\ I_r \end{pmatrix},$$

and C replaced by

$$\begin{pmatrix} C \\ 0 \\ 0 \end{pmatrix} \quad (C \text{ here referring to the matrix in (0.8))}.$$

It is not our goal to provide any general discussion of dynamic decoupling; our primary focus is on the period identification problem required for *adaptive* dynamic decoupling in the special context of (0.8). Nevertheless, we do present a simplified dynamic decoupling procedure for (0.8) in § 1 which we have used in conjunction with our period estimation procedure.

Whether decoupling is carried out as in § 1 to follow, or as in one of the cited references, it is clear that the estimator system requires knowledge of the matrix (0.5) and hence the parameter $\alpha = 2\pi/T$ in (0.6). When the period T, and hence α, is unknown it is necessary to adjoin a parameter estimator to supply the system with an estimate for T. Such a parameter estimator is described in § 2. Stability considerations in connection with the period estimator lead to examination of a related functional equation of retarded type in § 3. A numerical realization of the estimator of § 2 is developed in § 4 and examples of its use are presented in § 5.

1. Compensator design for a known disturbance frequency. If the period T, or equivalently, the frequency $\nu = 1/T$ of the disturbance v is known, then we may assume that F is known and the only problem in constructing the compensator (0.7) is the selection of the $2r \times n$ matrix S. Let us note that the matrix equation

$$\begin{pmatrix} A+CK & -B\Phi^{-1} \\ S & F \end{pmatrix}\begin{pmatrix} 0 \\ \Phi \end{pmatrix} - \begin{pmatrix} 0 \\ \Phi \end{pmatrix} F + \begin{pmatrix} B \\ 0 \end{pmatrix} = 0 \tag{1.1}$$

is clearly valid whatever S may be. This means that if we define ξ, η by

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ \Phi \end{pmatrix} z + \begin{pmatrix} \xi \\ \eta \end{pmatrix} \qquad (\dot{z} = Fz) \tag{1.2}$$

we shall have

$$\begin{pmatrix} \dot{\xi} \\ \dot{\eta} \end{pmatrix} = \begin{pmatrix} A+CK & -B\Phi^{-1} \\ S & F \end{pmatrix}\begin{pmatrix} \xi \\ \eta \end{pmatrix}, \tag{1.3}$$

as is easily checked. If the matrix in (1.3) is a stability matrix, then $x(t) = \xi(t)$ will have the property

$$\lim_{t\to\infty} \|x(t)\| = 0$$

so that the periodic disturbance $v(t) = Bz(t)$ has only a transient effect on $x(t)$; the range of the transfer function matrix from z to $\binom{x}{y}$ includes only vectors of the form $\binom{0}{y}$. Thus the plant state vector x is dynamically decoupled from z. If the matrix in (1.3) is not a stability matrix no such inferences are valid. Our proof that S can be selected so as to satisfy this stability requirement begins with

THEOREM 1. *Let F be antihermitian (as in (0.5)), so that*

$$F^* = -F.$$

Then the $n \times m$ linear matrix equation

$$(A+CK)P_0 - P_0F - B\Phi^{-1} = 0 \tag{1.4}$$

has a unique $n \times 2r$ solution P_0. If the pair (P_0, F) is observable, then the $2r \times n$ matrix S can be chosen in such a way that

$$M = \begin{pmatrix} A+CK & -B\Phi^{-1} \\ S & F \end{pmatrix} \tag{1.5}$$

is a stability matrix.

Proof. Since $F^* = -F$ implies that F has only purely imaginary eigenvalues, the existence of a unique solution P_0 of (1.4) is assured by a familiar theorem in matrix theory (see, for example, [2]). An easy application of the implicit function theorem then shows that the cubic matrix equation

$$(A+CK)P - PF - B\Phi^{-1} + \varepsilon PP^*P \equiv Q(P, \varepsilon) = 0 \tag{1.6}$$

has a unique solution $P = P(\varepsilon)$ defined for small $\varepsilon \to 0$ with

$$\lim_{\varepsilon\to 0} P(\varepsilon) = P_0. \tag{1.7}$$

Setting

$$S = S(\varepsilon) = -\varepsilon P(\varepsilon)^* \tag{1.8}$$

we note that M in (1.5) is similar to

$$\tilde{M}(\varepsilon)=\begin{pmatrix} I_n & -P(\varepsilon) \\ 0 & I_{2r} \end{pmatrix}\begin{pmatrix} A+CK & -B\Phi^{-1} \\ -\varepsilon P(\varepsilon)^* & F \end{pmatrix}\begin{pmatrix} I_n & P(\varepsilon) \\ 0 & I_{2r} \end{pmatrix}$$
$$=\begin{pmatrix} A+CK+\varepsilon P(\varepsilon)P(\varepsilon)^* & Q(P(\varepsilon),\varepsilon) \\ -\varepsilon P(\varepsilon)^* & F-\varepsilon P(\varepsilon)^*P(\varepsilon) \end{pmatrix}$$
$$=\begin{pmatrix} A+CK+\varepsilon P(\varepsilon)P(\varepsilon)^* & 0 \\ -\varepsilon P(\varepsilon)^* & F-\varepsilon P(\varepsilon)P(\varepsilon)^* \end{pmatrix}.$$

Since K has been chosen so that $A+CK$ is a stability matrix,

$$M_n(\varepsilon)\equiv A+CK+\varepsilon P(\varepsilon)P(\varepsilon)^*$$

is an $n\times n$ stability matrix for sufficiently small $\varepsilon>0$. From the antihermitian property of F we can see that

$$(F-\varepsilon P(\varepsilon)^*P(\varepsilon))^*I_{2r}+I_{2r}(F-\varepsilon P(\varepsilon)^*P(\varepsilon))+2\varepsilon P(\varepsilon)^*P(\varepsilon)=0.$$

Applying a well-known modification of Lyapunov's theorem (see, for example, [8]) we conclude that

$$M_{2r}(\varepsilon)\equiv F-\varepsilon P(\varepsilon)^*P(\varepsilon)$$

is a stability matrix for $\varepsilon>0$ if $(P(\varepsilon), F)$ is observable. Since we have assumed (P_0, F) observable and (1.7) is true, $(P(\varepsilon), F)$ is observable for $\varepsilon>0$ sufficiently small and $M_{2r}(\varepsilon)$ is thus a stability matrix for these values of ε, at least. Since $\tilde{M}(\varepsilon)$ is lower block triangular with blocks $M_n(\varepsilon)$, $M_{2r}(\varepsilon)$, its stability, and hence that of $M=M(\varepsilon)$ in (1.5), is assured with the choice (1.8) for $S=S(\varepsilon)$ for sufficiently small $\varepsilon>0$.

It will be noted that the choice of the feedback matrix K is important in at least two ways. Improvement of the convergence of $\Phi^{-1}y$ to z, i.e., reduction of the transient effect of the disturbance $v=Bz$, dictates choosing ε larger to improve the stability properties of $F-\varepsilon P(\varepsilon)^*P(\varepsilon)$. But, since $A+CK+\varepsilon P(\varepsilon)P(\varepsilon)^*$ suffers stability-wise as ε is increased, K must be used to offset this effect. In § 3 we will find even further considerations to take into account in the selection of ε and K.

Since $P(\varepsilon)$ satisfies a cubic equation, which may entail some difficulty of solution, the following corollary is useful in applications.

COROLLARY 2. *If ε is sufficiently small, then*

$$\hat{M}(\varepsilon)=\begin{pmatrix} A+CK & -B\Phi^{-1} \\ -\varepsilon P_0^* & F \end{pmatrix}, \tag{1.9}$$

corresponding to

$$S=\hat{S}(\varepsilon)=-\varepsilon P_0^* \tag{1.10}$$

in (1.5) *is also a stability matrix.*

Proof. With the indicated choice of S, the matrix $\hat{M}(\varepsilon)$ is similar to

$$\begin{pmatrix} I & -P_0 \\ 0 & I \end{pmatrix}\begin{pmatrix} A+CK & -B\Phi^{-1} \\ -\varepsilon P_0^* & F \end{pmatrix}\begin{pmatrix} I & P_0 \\ 0 & I \end{pmatrix}$$
$$=\begin{pmatrix} A+CK+\varepsilon P_0P_0^* & (A+CK)P_0-P_0F-B\Phi^{-1}+\varepsilon P_0P_0^*P_0 \\ -\varepsilon P_0^* & F-\varepsilon P_0^*P_0 \end{pmatrix} \tag{1.11}$$
$$=\text{cf. (1.4)}=\begin{pmatrix} A+CK+\varepsilon P_0P_0^* & \varepsilon P_0P_0^*P_0 \\ -\varepsilon P_0^* & F-\varepsilon P_0^*P_0 \end{pmatrix}.$$

Let $\mu = \varepsilon^{1/2}$ for $\varepsilon \geqq 0$. With $P_0(\mu) = \mu P_0$, the matrix (1.11) becomes

$$\begin{pmatrix} A + CK + P_0(\mu)P_0(\mu)^* & (1/\mu)P_0(\mu)P_0(\mu)^*P_0(\mu) \\ -\mu P_0(\mu)^* & F - P_0(\mu)^*P_0(\mu) \end{pmatrix}$$

which is similar to

$$\text{(1.12)} \quad \begin{pmatrix} A + CK + P_0(\mu)P_0(\mu)^* & P_0(\mu)P_0(\mu)^*P_0(\mu) \\ -P_0(\mu)^* & F - P_0(\mu)^*P_0(\mu) \end{pmatrix} \equiv \begin{pmatrix} A_1(\mu) & P_3(\mu) \\ -P_0(\mu)^* & F_1(\mu) \end{pmatrix}.$$

The corresponding lower triangular matrix

$$\begin{pmatrix} A_1(\mu) & 0 \\ -P_0(\mu)^* & F_1(\mu) \end{pmatrix}$$

is a stability matrix, using essentially the same argument as in Theorem 1, provided $\mu > 0$ is sufficiently small. Consider the equation

$$\text{(1.13)} \quad \begin{pmatrix} A_1(\mu)^* & -P_0(\mu) \\ 0 & F_1(\mu)^* \end{pmatrix}\begin{pmatrix} Q & R \\ R^* & T \end{pmatrix} + \begin{pmatrix} Q & R \\ R^* & T \end{pmatrix}\begin{pmatrix} A_1(\mu) & 0 \\ -P_0(\mu)^* & F_1(\mu) \end{pmatrix} + \begin{pmatrix} I & 0 \\ 0 & 2P_0(\mu)^*P_0(\mu) \end{pmatrix} = 0.$$

Solving this, we find that $T = I_{2r}$ and

$$\text{(1.14)} \quad A_1(\mu)^*Q - P_0(\mu)R^* + QA_1(\mu) - RP_0(\mu)^* + I = 0,$$

$$\text{(1.15)} \quad A_1(\mu)^*R - P_0(\mu) + RF_1(\mu) = 0.$$

For small μ (equivalently, small ε) the eigenvalues of $A_1(\mu)$ and $-F_1(\mu)$ are uniformly separated and the solution of (1.15) shows that

$$R = Q(\|P_0(\mu)\|) = (\mu)$$

and then a similar analysis of the first equation shows that

$$Q = Q_0(\mu) + (\mu^2)$$

where

$$A_1(\mu)^*Q_0(\mu) + Q_0(\mu)A_1(\mu) + I_n = 0.$$

Thus $Q_0(\mu)$, and hence Q, remains bounded for $\mu > 0$ small. Since (1.13) is satisfied, using the matrix of (1.12) instead, we have

$$\begin{pmatrix} A_1(\mu)^* & -P_0(\mu) \\ P_3(\mu)^* & F_1(\mu)^* \end{pmatrix}\begin{pmatrix} Q & R \\ R^* & T \end{pmatrix} + \begin{pmatrix} Q & R \\ R^* & T \end{pmatrix}\begin{pmatrix} A_1(\mu) & P_3(\mu) \\ -P_0(\mu)^* & F_1(\mu) \end{pmatrix}$$
$$= -\begin{pmatrix} I & (-\mu^2/\sqrt{2})QP_0P_0^* \\ 0 & \sqrt{2}\mu P_0^* \end{pmatrix}\begin{pmatrix} I & 0 \\ (-\mu^2/\sqrt{2})P_0P_0^*Q & \sqrt{2}\mu P_0 \end{pmatrix}$$
$$+ \begin{pmatrix} (\mu^4/2)QP_0P_0^*P_0P_0^*Q & 0 \\ 0 & 0 \end{pmatrix}.$$

From this it is easy to see that the matrix on the right-hand side is nonpositive for small $\mu > 0$ and the result follows by the familiar Lyapunov theorem, provided that whenever

$$\begin{pmatrix} \dot{\zeta} \\ \dot{\eta} \end{pmatrix} = \begin{pmatrix} A_1(\mu) & P_3(\mu) \\ -P_0(\mu)^* & F_1(\mu) \end{pmatrix}\begin{pmatrix} \hat{\zeta} \\ \hat{\eta} \end{pmatrix}$$

the quadratic form

$$(\hat{\zeta}(t)^*, \hat{\eta}(t)^*)\left[-\begin{pmatrix} I & (-\mu^2/\sqrt{2})QP_0P_0^* \\ O & \sqrt{2}\mu P_0^* \end{pmatrix}\begin{pmatrix} I & 0 \\ (-\mu^2/\sqrt{2})P_0P_0^*Q & \sqrt{2}\mu P_0 \end{pmatrix} + \begin{pmatrix} \mu^4 QP_0P_0^*P_0P_0^*Q & 0 \\ 0 & 0 \end{pmatrix}\right]\begin{pmatrix} \hat{\zeta}(t) \\ \hat{\eta}(t) \end{pmatrix}$$

cannot vanish on any interval of positive length. For small $\mu > 0$ this question reduces very quickly to the observability of the pair (P_0, F), which has already been assumed. This completes the proof.

2. Period estimation and a related stability problem. Whether the submatrix S in the definition (1.5) of M is selected as in Theorem 1, or as in Corollary 2, or by some other procedure, it is clear that the overall matrix M will depend on the period T, of the disturbance v so that, supposing now that the design parameter ε has been fixed, we have

$$M = M(\alpha) = \begin{pmatrix} A + CK & -B\Phi(\alpha)^{-1} \\ S(\alpha) & F(\alpha) \end{pmatrix} \tag{2.1}$$

where $\alpha = (2\pi/T)$ (cf. (0.6)).

It would be possible to estimate α directly using various well-known parameter estimation procedures [6], [9]. However, in these procedures one tends to encounter either instability or slow convergence, or other difficulties. For example, the model reference algorithm of [6] cannot be applied because in the complete system (0.12) the portion $\dot{z} = Fz$ of that system is not controllable with respect to u.

We have elected to use a very simple procedure to estimate the period T, directly. Assuming that an output, or observation,

$$\omega(t) = H_1x(t) + H_2y(t) = (H_1, H_2)\begin{pmatrix} x(t) \\ y(t) \end{pmatrix} \equiv Hw(t), \tag{2.2}$$

where (cf. (0.3), (0.12), (1.5))

$$\dot{w} = M(\alpha)w + \beta \qquad \left(\beta = \begin{pmatrix} v \\ 0 \end{pmatrix}\right), \tag{2.3}$$

is available, from the assumed stability property of the matrix $M(\alpha)$ it follows that a periodic disturbance input v will result in an output $\omega(t)$ which, except for transient behavior, is also periodic with the same period. It therefore makes sense, in the continuous framework which we use here for analysis, to consider the cost functional for $\gamma > 0$,

$$\begin{aligned} C_0(T, t) &= \int_0^t e^{\gamma(s-t)}(\omega(s) - \omega(s-T))^*(\omega(s) - \omega(s-T))\, ds \\ &= \int_0^t e^{\gamma(s-t)}(w(s) - w(s-T))^*H^*H(w(s) - w(s-T))\, ds \end{aligned} \tag{2.4}$$

and select as our estimate for the period T at time t, that value $T(t)$ which minimizes $C(T, t)$ within a given range $T_1 \leqq T \leqq T_2$. (The range must be restricted in order to avoid the trivial period $T = 0$ and multiples of the minimum period of the disturbance.) Then

$$\alpha(t) = \frac{2\pi}{T(t)} \tag{2.5}$$

is the estimate at time t for α in (0.6), (2.3). A numerical procedure approximating this optimization process is described in § 4 and is used to obtain the computational results of § 5. There it will be seen that certain steps do have to be taken in order to ensure the stability of the combined control/estimation system. Our purpose here is to provide a framework for the stability analysis by developing a linearized variational equation for that system about the nominal time trajectory in the case where the true period, which we will call T_0, lies in the interior of the interval $T_1 \leqq T \leqq T_2$.

For our analysis of the combined use of (2.3) and (2.4) we will consider, instead of $C_0(T, t)$, as given by (2.4), the cost

$$C(T, t) = \int_{-\infty}^{t} e^{\gamma(s-t)}(w(s) - w(s-T))^* H^* H(w(s) - w(s-T))\, ds, \tag{2.6}$$

wherein we assume that the trajectory $w(s)$ is defined in the indefinite past. The justification for this lies in the fact that if (2.3) and (2.6) together yield a stable process, the difference between the use of (2.6) and (2.4) will be transient.

To carry out this program we begin by supposing that when the correct value $\alpha_0 = (2\pi/T_0)$ is used in (2.3) the steady-state (T_0)-periodic solution resulting from the (T_0)-periodic input $\beta(t)$ is $w_0(t)$ and $\omega_0(t) = Hw_0(t)$. Since our estimate $\alpha(t)$ will vary from α_0, we suppose that the actual solution of (2.3) which we obtain is $w(t)$. Thus

$$\beta(t) = \beta(t - T_0), \tag{2.7}$$

$$w(t) = w_0(t) + \Delta w(t), \tag{2.8}$$

$$\alpha(t) = \alpha_0 + \Delta\alpha(t), \tag{2.9}$$

$$T(t) = T_0 + \Delta T(t). \tag{2.10}$$

A necessary condition in order that $T(t)$ should minimize $C(T, t)$ is obtained by differentiating $C(T, t)$ with respect to T and setting the result at $T = T(t)$ equal to zero. Thus (cf. (2.2), (2.3), (2.5))

$$\begin{aligned} 0 &= \frac{1}{2}\frac{\partial C(T, t)}{\partial T}\bigg|_{T=T(t)} \\ &= \int_{-\infty}^{t} e^{\gamma(s-t)}(w(s) - w(s - T(t)))^* H^* H\dot{w}(s - T(t))\, ds \\ &= \int_{-\infty}^{t} e^{\gamma(s-t)}(w(s) - w(s - T(t)))^* H^* H \\ &\quad \times [M(\alpha(s - T(t)))w(s - T(t)) + \beta(s - T(t))]\, ds. \end{aligned} \tag{2.11}$$

Noting (2.5), we see that (2.11) is implicitly an equation for $T(t)$ which is coupled with the system (2.3) satisfied by $w(t)$. The resulting coupled system is clearly a nonlinear functional equation of delay type. We are concerned with the (at least local with respect to α_0 and $w_0(t)$) existence, uniqueness and asymptotic stability of solutions.

LEMMA 3. *For fixed t and a trajectory $w(s) - \infty < s \leqq t$, for* (2.3), *corresponding to a continuous (T_0)-periodic $\beta(s)$, $-\infty < s \leqq t$, the equation* (2.11) *is solvable for $T(t)$ near T_0 if*

$$-(w(t) - w(t - T_0))^* H^* H\dot{w}(t - T_0) + \int_{-\infty}^{t} e^{\gamma(s-t)}\dot{w}(s)^* H^* H\dot{w}(s - T_0)\, ds$$

$$
\begin{aligned}
&+\gamma\int_{-\infty}^{t} e^{\gamma(s-t)}(w(s)-w(s-T_0))^*H^*H\dot{w}(s-T_0)\,ds\\
&=\int_{-\infty}^{t} e^{\gamma(s-t)}\dot{w}(s)^*H^*H\dot{w}(s-T_0)\,ds\\
&-\frac{\partial}{\partial t}\int_{-\infty}^{t} e^{\gamma(s-t)}(w(s)-w(s-T_0))^*H^*H\dot{w}(s-T_0)\,ds\neq 0.
\end{aligned}
\tag{2.12}
$$

Within the class of w which satisfy

$$\|w(s)\|\leqq M_0,\qquad -\infty<s\leqq t, \tag{2.13}$$

$$\|w(s)-w_0(s)\|\leqq\varepsilon,\qquad t-\tau\leqq s\leqq t, \tag{2.14}$$

this is true for sufficiently large τ and sufficiently small ε if

$$\int_{-\infty}^{t} e^{\gamma(s-t)}\dot{w}_0(s)H^*H\dot{w}_0(s)\,ds\neq 0 \tag{2.15}$$

and this, in turn, is true if the (T_0)-periodic function $Hw_0(s)=\omega_0(s)$ is not constant.

Proof. Assume for the moment that $\beta(t)$ is a function in C^1. Differentiating the second line of (2.11) with respect to T at $T=T_0$ and retaining only zero and first order terms in $\Delta T(t)$ we obtain the equation, linearized with respect to ΔT,

$$
\begin{aligned}
0=&\int_{-\infty}^{t} e^{\gamma(s-t)}\dot{w}(s-T_0)^*H^*H\dot{w}(s-T_0)\,ds\,\Delta T(t)\\
&-\int_{-\infty}^{t} e^{\gamma(s-t)}(w(s)-w(s-T_0))^*H^*H\ddot{w}(s-T_0)\,ds\,\Delta T(t)\\
&+\int_{-\infty}^{t} e^{\gamma(s-t)}(w(s)-w(s-T_0))^*H^*H\dot{w}(s-T_0)\,ds+\cdots.
\end{aligned}
\tag{2.16}
$$

Integrating the second term by parts we have

$$
\begin{aligned}
0=&\Big[-(w(t)-w(t-T_0))^*H^*H\dot{w}(t-T_0)+\int_{-\infty}^{t} e^{\gamma(s-t)}\dot{w}(s)^*H^*H\dot{w}(s-T_0)\,ds\\
&+\gamma\int_{-\infty}^{t} e^{\gamma(s-t)}(w(s)-w(s-T_0))^*H^*H\dot{w}(s-T_0)\,ds\Big]\Delta T(t)\\
&+\int_{-\infty}^{t} e^{\gamma(s-t)}(w(s)-w(s-T_0))^*H^*H\dot{w}(s-T_0)\,ds+\cdots.
\end{aligned}
\tag{2.17}
$$

This expression no longer depends on $\ddot{w}(s-T_0)$, hence we may relax the requirement $\beta\in C^1$ since $\beta\in C^0$ can be uniformly approximated in the C^0 norm by $\beta\in C^1$. Examination of the remainder term shows that the same argument applies there and we conclude that (2.16) is indeed valid to first order in $\Delta T(t)$. The first statement of our lemma then follows immediately from the implicit function theorem.

The last statement follows from the property

$$w_0(s)-w_0(s-T_0)\equiv 0,\qquad -\infty<s\leqq t, \tag{2.18}$$

and together with (2.3), (2.7), (2.17), this enables one to make the first term in (2.16) arbitrarily close to the left-hand side of (2.15) and the second term arbitrarily close to zero. This completes the proof of Lemma 3.

The linearization with respect to $\Delta T, \Delta w$ is obtained by using (2.18) in (2.16). Because $w_0(s) - w_0(s - T_0) \equiv 0$ and because only zero order terms are retained as coefficients of the first order term $\Delta T(t)$, the result is

$$\left[\int_{-\infty}^{t} e^{\gamma(s-t)} \dot{w}_0(s)^* H^* H \dot{w}_0(s - T_0)\, ds\right] \Delta T(t)$$

$$+ \int_{-\infty}^{t} e^{\gamma(s-t)} (\Delta w(s) - \Delta w(s - T_0))^* H^* \dot{w}_0(s - T_0)\, ds = 0,$$

and using (2.18), $\dot{w}_0(s - T_0) \equiv \dot{w}(s)$, and the assumption (2.15) we have

$$\Delta T(t) = \frac{-\int_{-\infty}^{t} e^{\gamma(s-t)} (\Delta w(s) - \Delta w(s - T_0))^* H^* H \dot{w}_0(s)\, ds}{\int_{-\infty}^{t} e^{\gamma(s-t)} \dot{w}_0(s)^* H^* H \dot{w}_0(s)\, ds}, \tag{2.19}$$

which is a delay type functional equation relating $\Delta T(t)$ and the time history of Δw. Then from (2.3) we have, to first order,

$$\Delta \dot{w} + \dot{w}_0 = M(\alpha_0) w_0 + \beta + M(\alpha_0) \Delta w + \left(\frac{\partial M}{\partial \alpha}(\alpha_0) \Delta \alpha\right) w_0 + \cdots$$

and, since $\dot{w}_0 = M(\alpha_0) w_0 + \beta$ and (2.5) applies, we have as our linearized system (2.19) and

$$\Delta \dot{w}(t) = M(\alpha_0) \Delta w(t) - \left(\frac{2\pi}{T_0^2} \frac{\partial M}{\partial \alpha}(\alpha_0) w_0(t)\right) \Delta T(t). \tag{2.20}$$

A single equation may be obtained by noting that for $\tau > T_0$

$$\Delta \dot{w}(s) - \Delta \dot{w}(s - T_0) = M(\alpha_0)(\Delta w(s) - \Delta w(s - T_0)) - \frac{2\pi}{T_0^2} \frac{\partial M}{\partial \alpha}(\alpha_0) w_0(s) \Delta T(s)$$

$$+ \frac{2\pi}{T_0^2} \frac{\partial M}{\partial \alpha}(\alpha_0) w_0(s - T_0) \Delta T(s - T_0)$$

$$= M(\alpha_0)(\Delta w(s) - \Delta w(s - T_0))$$

$$- \frac{2\pi}{T_0^2} \frac{\partial M}{\partial \alpha}(\alpha_0) w_0(s)(\Delta T(s) - \Delta T(s - T_0)),$$

where we have used $w_0(s) = w_0(s - T_0)$. Then

$$\Delta w(t) - \Delta w(t - T_0) = \frac{-2\pi}{T_0^2} \int_{-\infty}^{t} e^{M(\alpha_0)(t-\sigma)} \frac{\partial M}{\partial \alpha}(\alpha_0) w_0(\sigma)(\Delta T(\sigma) - \Delta T(\sigma - T_0))\, ds$$

and thus

$$\Delta T(t) = \frac{2\pi / T_0^2}{\int_{-\infty}^{t} e^{\gamma(s-t)} \dot{w}_0(s)^* H^* \dot{w}_0(s)\, ds}$$

$$\times \int_{-\infty}^{t} e^{\gamma(s-t)} \int_{-\infty}^{s} w_0(\sigma)^* \frac{\partial M}{\partial \alpha}(\alpha_0)^* e^{M(\alpha_0)^*(s-\sigma)}$$

$$\times (\Delta T(\sigma) - \Delta T(\sigma - T_0))\, d\sigma\, H^* H \dot{w}_0(s)\, ds.$$

Letting

$$W_0(\gamma, t) = \frac{2\pi / T_0^2}{\int_{-\infty}^{t} e^{\gamma(s-t)} \dot{w}_0(s)^* H^* H \dot{w}_0(s)\, ds} \tag{2.21}$$

we have

$$\Delta T(t) = W_0(\gamma, t) \int_{-\infty}^{t} w_0(\sigma)^* \frac{\partial M}{\partial \alpha}(\alpha_0)^* \int_{\sigma}^{t} e^{\gamma(s-t)}\, e^{M(\alpha_0)^*(s-\sigma)} H^* H \dot{w}_0(s)\, ds \times (\Delta T(\sigma) - \Delta T(\sigma - T_0))\, d\sigma,$$

which has the form

$$\Delta T(t) = \int_{-\infty}^{t} W_1(\gamma, t, \sigma, M(\alpha_0))(\Delta T(\sigma - T_0))\, d\sigma \tag{2.22}$$

with

$$W_1(\gamma, t, \sigma, M(\alpha_0)) = W_0(\gamma, t) w_0(\sigma)^* \frac{\partial M}{\partial \alpha}(\alpha_0)^* \times \int_{\sigma}^{t} e^{\gamma(s-t)}\, e^{M(\alpha_0)^*(s-\sigma)} H^* H \dot{w}_0(s)\, ds. \tag{2.23}$$

LEMMA 4. *$W_1(\gamma, t, \sigma, M(\alpha_0))$ is periodic in t and σ with period T_0, in the sense*

$$W_1(\gamma, t, \sigma, M(\alpha_0)) = W_1(\gamma, t + T_0, \sigma + T_0, M(\alpha_0)).$$

Proof. Using the formula (2.23) directly we see that

$$W_1(\gamma, t + T_0, \sigma + T_0, M(\alpha_0)) = W_0(\gamma, t + T_0) w_0(\sigma + T_0)^* \frac{\partial M}{\partial \alpha}(\alpha_0)^* \times \int_{\sigma + T_0}^{t + T_0} e^{\gamma(s - t - T_0)}\, e^{M(\alpha_0)^*(s - \sigma - T_0)} H^* H \dot{w}_0(s)\, ds.$$

From the (T_0)-periodicity of $w_0(t)$ the same periodicity of $W_0(\gamma, t)$ follows easily. Then with $r = s - T_0$

$$W_1(\gamma, t + T_0, \sigma + T_0, M(\alpha_0)) = W_0(\gamma, t) w_0(\sigma)^* \frac{\partial M}{\partial \alpha}(\alpha_0)^* \times \int_{\sigma}^{t} e^{\gamma(r-t)}\, e^{M(\alpha_0)^*(r-\sigma)} H^* H \dot{w}_0(r + T_0)\, dr$$

and, using the (T_0)-periodicity of $\dot{w}_0$ we have

$$W_1(\gamma, t + T_0, \sigma + T_0, M(\alpha_0)) = W_1(\gamma, t, \sigma, M(\alpha_0)),$$

as claimed.

The equation (2.22) can be rewritten as

$$\Delta T(t) = \int_{-\infty}^{t} W(\gamma, t, \sigma, M(\alpha_0)) \Delta T(\sigma)\, d\sigma \tag{2.24}$$

with

$$W(\gamma, t, \sigma, M(\alpha_0)) = W_1(\gamma, t, \sigma, M(\alpha_0)), \qquad t - T_0 < \sigma \leqq t,$$

$$W(\gamma, t, \sigma, M(\alpha_0)) = W_1(\gamma, t, \sigma, M(\alpha_0)) - W_1(\gamma, t, \sigma + T_0, M(\alpha_0)), \qquad -\infty < \sigma \leqq t - T_0.$$

Since, for $\sigma = t - T_0$, $W_1(\gamma, t, \sigma + T_0, M(\sigma_0)) = W_1(\gamma, t, t, M(\alpha_0)) = 0$ we see that $W(\gamma, t, \sigma, M(\alpha_0))$ is continuous as a function of σ and, clearly, $W(\gamma, t + T_0, \sigma + T_0, M(\alpha_0)) = W(\gamma, t, \sigma, M(\alpha_0))$.

We have shown $\gamma, M(\alpha_0)$ directly as arguments of W because γ, ε and K, the last two involved in the construction of $M(\alpha_0)$ (see (2.1)) are the parameters which we have to work with in order to influence W, and hence the solutions of (2.24). It is clear from (2.22) that W depends on T_0 as well.

The stability and rate of convergence of our estimation procedure depends on the behavior of solutions of the variational equation (2.24), for $T(t)$ in a neighborhood of the exact value T_0, i.e., for $\Delta T(t)$ small. In the next section we show that the asymptotic behavior of solutions of (2.24) is dominated by solutions of "Floquet type" which can be analyzed by standard procedures.

3. Analysis of Floquet type solutions. The fact that (2.24) involves an infinite time delay places it in a class of functional differential equations with periodic coefficients whose properties have not been fully explored. From the behavior of solutions of such equations with finite delays [3], [4] we expect that, with some restrictions on the kernel $W(\gamma, t, \sigma, M(\alpha_0))$, the dominant solutions should be solutions of "Floquet type," i.e., solutions of the form

$$\Delta T(t) = e^{\lambda t} P(t), \tag{3.1}$$

where $P(t)$ is a continuous (T_0)-periodic function

$$P(t+T_0) = P(t).$$

The main point of this section is to indicate that this is, indeed, the case for kernels satisfying a uniform decay condition

$$|W(\gamma, t, \sigma, M(\alpha_0))| \leqq C e^{-c(t-\sigma)}, \qquad \sigma \leqq t, \tag{3.2}$$

for positive C, c.

Before entering upon the proof of this, let us note some rather transparent results which, however superficial, give some indication of the factors which are likely to play a role in our analysis. Suppose an inequality (3.2) is satisfied for positive c, C. Supposing a solution of the form (3.1) to exist, we normalize $P(t)$ so that

$$\sup_{s \in [t, t-T_0]} |P(s)| = 1.$$

Then we let t be such that $|P(t)| = 1$. Multiplying by a constant, if need be, we may assume $P(t) = 1$. Then

$$e^{\lambda t} - \int_{-\infty}^{t} W(\gamma, t, \sigma, M(\alpha_0))\, e^{\lambda \sigma} P(\sigma)\, d\sigma = 0$$

or

$$\int_{-\infty}^{t} W(\gamma, t, \sigma, M(\alpha_0))\, e^{\lambda(\sigma - t)} P(\sigma)\, d\sigma = 1.$$

But

$$|W(\gamma, t, \sigma, M(\alpha_0))\, e^{\lambda(\sigma-t)} P(\sigma)| \leqq C e^{(c + \mathrm{Re}(\lambda))(\sigma - t)}$$

so that

$$C \int_{-\infty}^{t} e^{(c+\mathrm{Re}(\lambda))(\sigma-t)}\, d\sigma \geqq 1,$$

yielding an upper bound on $\mathrm{Re}(\lambda)$

$$\frac{C}{c + \mathrm{Re}(\lambda)} \geqq 1 \quad \Rightarrow \mathrm{Re}(\lambda) \leqq C - c.$$

Under what circumstances could a bound of the type (3.2) be expected? Recalling that

$$W_1(\gamma, t, \sigma, M(\alpha_0)) = W_0(\gamma, t) w_0(\sigma)^* \frac{\partial M}{\partial \alpha}(\alpha_0)^* \, e^{\gamma(\sigma - t)} \times \int_\sigma^t e^{\gamma(s-\sigma)} \, e^{M(\alpha_0)^*(s-\sigma)} H^* H \dot{w}_0(s) \, ds,$$

we note that with $r = s - \sigma$,

$$\int_\sigma^t e^{\gamma(s-\sigma)} \, e^{(M(\alpha_0)^* + \gamma I)r} H^* H \dot{w}(r+\sigma) \, ds = \int_0^{t-\sigma} e^{(M(\alpha_0)^* + \gamma I)r} H^* H \dot{w}_0(r+\sigma) \, dr.$$

Since $\dot{w}_0$ is periodic, if the eigenvalues μ of $M(\alpha_0)$ satisfy

$$\operatorname{Re}(\mu) \leqq -\delta$$

for some $\delta > 0$, we will have, for some $M_0 > 0$

$$\| e^{(M(\alpha_0)^* + \gamma I)r} H^* H \dot{w}_0(r+\sigma) \| \leqq M_0 \, e^{(\gamma - \delta)r},$$

so that

$$\left\| \int_0^{t-\sigma} e^{(M(\alpha_0)^* + \gamma I)r} H^* H \dot{w}_0(r+\sigma) \, dr \right\| \leqq M_0 \int_0^{t-\sigma} e^{(\gamma-\delta)r} \, dr = \frac{M_0}{\gamma - \delta} [e^{(\gamma-\delta)(t-\sigma)} 1]. \tag{3.3}$$

We expect $W_0(t, \gamma)$ to be (γ) from (2.21); write

$$|W_0(t, \gamma)| \leqq M_1 \gamma$$

and then, since $w_0(\sigma)$ is periodic,

$$\left\| W_0(t, \gamma) w_0(\sigma)^* \frac{\partial M}{\partial \alpha}(\alpha_0)^* \, e^{\gamma(\sigma - t)} \right\| \leqq M_2 \gamma \, e^{\gamma(\sigma - t)}. \tag{3.4}$$

Combining this with (3.3), (2.23)

$$|W_1(\gamma, t, \sigma, M(\alpha_0))| \leqq M_0 M_2 \frac{\gamma}{\gamma - \delta} \, e^{\delta(\sigma - t)}$$

giving $C = M_0 M_2 (\gamma/(\gamma - \delta))$, $c = \delta$. A comparable estimate will then apply to $W(\gamma, t, \sigma, M(\alpha_0))$.

From this we see that if we are to control the identifier stability properties, this must be done through γ and through the system matrix $M(\gamma_0)$, by choice of γ, ε and K (or through choice of γ, K, L_0, L_1 if we use the full system estimator as described in § 7). Further, we see from (2.21) that $W_0(\gamma, t)$, and hence $W_1(\gamma, t, \sigma, M(\alpha_0))$, $W(\gamma, t, \sigma, M(\alpha_0))$ increase rapidly as the frequency parameter $\alpha = 2\pi / T_0$ increases, i.e., as T_0 decreases. Thus, to be able to reject higher frequencies while maintaining stability we must expect to find it necessary to increase the damping in the system (2.3) by use of higher gains ε and K (of (1.5), (1.8)). We will also see in § 5 that this expectation is realized.

We proceed now to state a theorem to the effect that if a bound of the form (3.2) applies, then all solutions of (2.24) which do not satisfy

$$\|\Delta T(t)\| \leqq B \, e^{-\beta t}, \qquad 0 \leqq t < \infty, \tag{3.5}$$

where B is positive and $\beta > 0$ is less than c by an arbitrarily small amount, must be linear combinations of Floquet type solutions.

THEOREM 5. *Consider the vector functional equation*

$$z(t) = \int_{-\infty}^{t} W(t, s)z(s)\, ds, \qquad z \in R^m, \tag{3.6}$$

where $W(t, s)$ is a (piecewise continuous, at least) $m \times m$ matrix function satisfying

$$\| W(t, s)\| \leqq C e^{-c(t-s)}, \qquad -\infty < s \leqq t, \tag{3.7}$$

$$W(t+T, s+t) = W(t, s), \tag{3.8}$$

for positive numbers C, c, T. Then, given any $\beta < c$, and any solution $z(t)$ with locally square integrable initial function satisfying

$$\int_{-\infty}^{0} e^{2cs} \| \tilde{z}(s)\|_{R^m}^2 \, ds < \infty, \tag{3.9}$$

we can write

$$z(t) = z_F(t) + z_\beta(t), \qquad t \geqq 0 \tag{3.10}$$

where, for some positive B,

$$\| z_\beta(t)\| \leqq B e^{-\beta t}, \qquad t \geqq 0 \tag{3.11}$$

and $z_F(t)$ is a linear combination of Floquet type solutions, i.e., solutions of the form

$$\zeta(t) = e^{\lambda t} P(t), \quad P(t+T) = P(t), \quad P \in C([0, \infty], R^m), \tag{3.12}$$

or, in some cases (multiple "eigenvalues")

$$\zeta(t) = e^{\lambda t} t^p P(t), \tag{3.13}$$

where p is a positive integer and $P(t)$ is as in (3.12).

A complete proof of Theorem 5 is beyond the scope of the present work but a sketch of the proof will be given in the Appendix.

From this result we see that whenever an inequality of the type (3.2) is valid with $c > 0$, then all solutions of (2.24) decay at a uniform exponential rate unless there are actually solutions (3.1) of Floquet type for which $\mathrm{Re}\,(\lambda) > 0$. The question arises, of course, as to how such Floquet exponents might actually be computed. It seems almost certain that the most efficient procedure involves actual solution of (2.24) or (2.19), (2.20), assuming an adequate approximation procedure is available. The procedure is essentially the same one as is used to compute the dominant (pairs of) root(s) of an ordinary polynomial.

Returning to $\Delta T(t)$ as the name for the solution, we select a more or less arbitrary initial state $\Delta T(t)$ on some interval $[-\tau, 0]$ (in terms of § 6 this should be a $\tilde{z}$ such that the residue of $(I - Q(\lambda))^{-1} q(\lambda, \tilde{z})$ at $\lambda = \lambda z$ is not zero, which is generically true). The resulting solution $\Delta T(t)$, $t \geqq 0$, is computed and we examine successive segments of length T_0

$$\Delta T_k(s) = \Delta T(kT_0 + s), \qquad 0 \leqq s \leqq T_0, \quad k = 0, 1, 2, 3, \cdots.$$

If the largest multiplier

$$\mu_\nu = e^{\lambda_\nu T_0}$$

is a unique real number, then generically with respect to the choice of initial function $\Delta T(t)$, $t \in [-\tau, 0]$, we shall have (using the least squares approach)

$$\mu_\nu = \lim_{k \to \infty} \frac{\int_0^T \Delta T_k(s) \Delta T_{k-1}(s)\, ds}{\int_0^T (\Delta T_{k-1}(s))^2\, ds}.$$

In the case of a dominant complex conjugate pair the procedure is only slightly more complicated. We solve

$$\Delta T_k(s)+\alpha\Delta T_{k-1}(s)+\beta\Delta T_{k-2}(s)=0$$

for α and β in the least squares (least L^2 norm) sense, which amounts to

$$\begin{pmatrix} \int_0^T \Delta T_{k-1}(s)^2\,ds & \int_0^T \Delta T_{k-1}(s)\Delta T_{k-2}(s)\,ds \\ \int_0^T \Delta T_{k-2}(s)\Delta T_{k-1}(s)\,ds & \int_0^T \Delta T_{k-2}(s)^2\,ds \end{pmatrix} \begin{pmatrix} \alpha_k \\ \beta_k \end{pmatrix} + \begin{pmatrix} \int_0^T \Delta T_k(s)\Delta T_{k-1}(s)\,ds \\ \int_0^T \Delta T_k(s)\Delta T_{k-2}(s)\,ds \end{pmatrix} = 0.$$

The pair $\mu_\nu, \bar{\mu}_\nu$ is then approximated at the kth stage by the roots $\mu_k, \bar{\mu}_k$ of

$$\mu^2+\alpha_k\mu+\beta_k=0.$$

It seems likely that while (2.24) is nicer from the viewpoint of mathematical simplicity, it is better to solve (2.19), (2.20) rather than (2.24) because the formula for the kernel $W(\gamma, t, \sigma, M(\alpha_0))$ in (2.24) is rather complicated.

If a simulation routine combining the period estimator, compensator and a mathematical model of the plant to be controlled is already in hand, as was the case for the writer, approximate solutions of the variation equation can be obtained by running the simulator with slightly different initial conditions and forming the appropriate difference quotient of the resulting solutions. This does not test the validity of our derivation of the variational equation but, as we will see in § 5, it does provide results consistent with the proposed functional equation model for error propagation.

4. Numerical realization of the period estimator. If $x(t)$ is a solution of

$$\dot{x}(t)=(A+CK)x(t)+v(t), \quad t\geqq 0,$$

and the disturbance $v(t)$ is periodic with period T

$$v(t)=v(t+T),$$

then an observation on $x(t)$,

$$\omega(t)=Hx(t)$$

will tend exponentially to a periodic observation, i.e.,

$$\lim_{t\to\infty}(\omega(t)-\omega(t+T))=0.$$

In this section we develop a numerical procedure for estimation of T which is a realization of the continuous procedure described in § 2. We will take ω to be scalar here but the extension to vector observations is quite immediate.

We will suppose that $\omega(t)$ is not available continuously. Rather, we have discrete samples

$$\omega_k=\omega(t_k), \quad t_{k+1}-t_k=h>0, \quad k=0, 1, 2, \cdots.$$

For computational purposes we define the interpolated observation on $t_k\leqq t\leqq t_{k+1}$ by

$$\tilde{\omega}(t_k+\sigma h)\equiv\eta_k(\sigma)=\sigma\omega_{k+1}+(1-\sigma)\omega_k, \qquad 0\leqq\sigma\leqq 1. \tag{4.1}$$

We note that $\tilde{\omega}(t_k) = \omega_k$, $\tilde{\omega}(t_{k+1}) = \omega_{k+1}$. We define $\eta_k = \omega_k$, $k = 0, 1, 2, \cdots$. Our method for estimating T is to form, at each instant t_k, and for a range $L_{\min} \leqq l \leqq L_{\max}$ (we use L_m, L_M below), the functions

$$\rho_{k,l}(\sigma) = \eta_k - \eta_{k-l}(\sigma) \tag{4.2}$$

and determine values l_k, σ_k of l, σ which minimize

$$C_{k,l}(\sigma) = \sum_{j=0}^{k} \gamma^j \rho_{k-j,l}(\sigma),$$

which should be compared with (2.6). The functions (4.2), of course, require only the values $\eta_k = \omega_k$, $\eta_{k+1} = \omega_{k+1}$ for this description and the $\rho_{k,l}$ admit a comparable finite characterization. Once l_k, σ_k have been determined, the estimate for T at the instant t_k is

$$T_k = (l_k - \sigma_k)h. \tag{4.3}$$

If γ is close to one this estimate may be expected to change only slowly, as k varies, in response to varying periodic behavior of $\omega(t)$ while values of γ closer to zero provide more rapid updating capability. The use of the parameter σ, allowing for interpolation between recorded discrete data, permits one to obtain accurate results without an excessively fast sampling rate.

Let us now examine the computational considerations applying to the method. For $0 < \sigma < 1$ we have

$$\rho_{k,l}(\sigma) = \eta_k - [\sigma\eta_{k-l+1} + (1-\sigma)\eta_{k-l}]$$

and thus

$$\rho_{k,l}(\sigma)^2 = \eta_k^2 + \sigma^2\eta_{k-l+1}^2 + (1-\sigma)^2\eta_{k-l}^2 - 2\sigma\eta_{kk-l+1} - 2(1-\sigma)\eta_k\eta_{k-l} + 2\sigma(1-\sigma)\eta_{k-l+1}\eta_{k-l}.$$

Defining

$$S_k = \sum_{j=0}^{\infty} \gamma^j \eta_{k-j}^2, \quad S_{k-l+1} = \sum_{j=0}^{\infty} \gamma^j \eta_{k-l+1-j}^2, \quad S_{k-l} = \sum_{j=0}^{\infty} \gamma^j \eta_{k-l-j}^2,$$

$$P_{k,k-l} = \sum_{j=0}^{\infty} \gamma^j \eta_{k-j}\eta_{k-l-j}, \qquad P_{k,k-l+1} = \sum_{j=0}^{\infty} \gamma^j \eta_{k-j}\eta_{k-l+1-j},$$

$$P_{k-l+1,k-l} = \sum_{j=0}^{\infty} \gamma^j \eta_{k-l+1-j}\eta_{k-l-j},$$

we see that

$$C_{k,l}(\sigma) = \sigma^2[S_{k-l+1} - S_{k-l} - 2P_{k-l+1,k-l}] + 2\sigma[S_{k-l} - P_{k,k-l+1} + P_{k,k-l} + P_{k-l+1,k-l}] + [S_k + S_{k-l} - 2P_{k,k-l}].$$

The numbers S_k, S_{k-l}, S_{k-l+1} are included in $S_k, \cdots, S_{k-L_M}$, and these are stored in a "push-down" mode and updated via

$$S_{k+1} = \eta_{k+1}^2 + \gamma S_k, \qquad S_{(k+1)-1} = S_k, \cdots,$$

$$S_{(k+1)-L_M} = S_{k-(L_M-1)}.$$

Similarly $P_{k,k-l}$, $P_{k,k-l+1}$ are stored among $P_{k,k-1}, \cdots, P_{k,k-L_M}$ are are updated via

$$P_{k+1,k+1-l} = \eta_{k+1}\eta_{k+1-l} + \gamma P_{k,k-l}, \text{ etc.} \tag{4.4}$$

Finally, it is necessary to store

$$P_{k,k-1}, P_{k-1,k-2}, \cdots, P_{k-L_M+1,k-L_M}.$$

The numbers $P_{k,k-1}$ are also updated via (4.4) and

$$P_{(k+1)-l,(k+1)-(l+1)} = P_{k-(l-1),k-l}$$

defines the "push-down" operation.

With the above numbers available we clearly have

$$\frac{1}{2}\frac{\partial C_{k,l}}{\partial \sigma} = \sigma[S_{k-l+1} - S_{k-l} - 2P_{k-l+1,k-l}] + [S_{k-l} - P_{k,k-l+1} + P_{k,k-l} + P_{k-l+1,k-l}]. \tag{4.5}$$

In particular,

$$\frac{1}{2}\frac{\partial C_{k,l}}{\partial \sigma}\bigg|_{\sigma=0} = S_{k-l} - P_{k,k-l+1} + P_{k,k-l} + P_{k-l+1,k-l},$$

$$\frac{1}{2}\frac{\partial C_{k,l}}{\partial \sigma}\bigg|_{\sigma=1} = S_{k-l+1} - P_{k,k-l+1} + P_{k,k-l} - P_{k-l+1,k-l}.$$

Each pair l, σ corresponds to a delay $T(l, \sigma) = (l-\sigma)h$. Thus $C_{k,l}(\sigma)$ can be associated with a function $C_k(t)$ defined for $t_k - L_M h < t < t_k - L_m h$, the values of l corresponding, when $\sigma = 0$, to the points $t = t_k - lh$. As σ increases from zero to one we pass from $t = t_k - lh$ to $t = t_k - (l-1)h$. Thus we have

$$\frac{\partial C_{k,l}}{\partial \sigma}\bigg|_{\sigma=0} = \frac{\partial C_k}{\partial t}\bigg|_{t=(t_k-lh)+},$$

$$\frac{\partial C_{k,l}}{\partial \sigma}\bigg|_{\sigma=1} = \frac{\partial C_k}{\partial t}\bigg|_{t=(t_k-(l-1)h)-}.$$

It follows that $T(l, 0) = lh$ is a candidate for the minimizing value T_k just in case

$$\frac{1}{2}\frac{\partial C_{k,l}}{\partial \sigma}\bigg|_{\sigma=0} \geqq 0, \qquad \frac{1}{2}\frac{\partial C_{k,l+1}}{\partial \sigma}\bigg|_{\sigma=1} \leqq 0,$$

i.e.,

$$S_{k-l} - P_{k,k-l+1} + P_{k,k-l} + P_{k-l+1,k-l} \geqq 0,$$

$$S_{k-l} - P_{k,k-l} + P_{k,k-l+1} - P_{k-l,k-l-1} \leqq 0.$$

On the other hand, the interval $[(l-1)h, lh]$ is a candidate for containing the minimizing value of T_k just in case

$$\frac{1}{2}\frac{\partial C_{k,l}}{\partial \sigma}\bigg|_{\sigma=1} > 0, \qquad \frac{1}{2}\frac{\partial C_{k,l}}{\partial \sigma}\bigg|_{\sigma=0} < 0,$$

i.e.,

$$S_{k-l+1} - P_{k,k-l+1} + P_{k,k-l} - P_{k-l+1,k-l} > 0, \tag{4.6}$$

$$S_{k-l} - P_{k,k-l+1} + P_{k,k-l} + P_{k-l+1,k-l} < 0. \tag{4.7}$$

If (4.6), (4.7) are true for a given l, we compute the corresponding σ by setting (4.5) equal to zero, i.e.,

$$\sigma = -\frac{[S_{k-l} - P_{k,k-l+1} + P_{k,k-l} + P_{k-l+1,k-l}]}{[S_{k-l+1} - S_{k-l} - 2P_{k-l+1,k-l}]}. \tag{4.8}$$

Once the finitely many possible candidates for T_k have been selected by this process, T_k is chosen from these as the one yielding the smallest value of $C_{k,l}(\sigma)$.

It is possible to economize on memory space by using slightly modified quantities. With

$$\begin{aligned} &\rho_{k,l} = \eta_k - \eta_{k-l}, \qquad \hat{S}_{k,l} = \sum_{j=0}^{\infty} \gamma^j (\rho_{k-j,l})^2, \\ &\hat{P}_{k,l,l-1} = \sum_{j=0}^{\infty} \gamma^j \rho_{k-j,l}\rho_{k-j,l-1}, \qquad \hat{S}_{k,l-1} = \sum_{j=0}^{\infty} \gamma^j (\rho_{k-j,l-1})^2, \end{aligned} \tag{4.9}$$

updated via

$$\begin{aligned} &\rho_{k+1,l} = (\eta_{k+1} - \eta_k) + \rho_{k,l-1}, \qquad l = 1, 2, \cdots, L_M, \\ &\hat{S}_{k+1,l} = (\rho_{k+1,l})^2 + \gamma \hat{S}_{k,l}, \qquad l = 1, 2, \cdots, L_M, \\ &\hat{P}_{k+1,l,l-1} = \rho_{k+1,l}\rho_{k+1,l-1} + \gamma \hat{P}_{k,l,l-1}, \qquad l = 2, \cdots, L_M, \end{aligned}$$

it may be seen that we have

$$C_{k,l}(\sigma) = \sigma^2[\hat{S}_{k,l} - 2\hat{P}_{k,l,l-1} + \hat{S}_{k,l-1}] - 2\sigma[\hat{S}_{k,l} - \hat{P}_{k,l,l-1}] + \hat{S}_{k,l-1} \tag{4.10}$$

so

$$\frac{1}{2}\frac{\partial C_{k,l}}{\partial \sigma} = \sigma[\hat{S}_{k,l} - 2\hat{P}_{k,l,l-1} + \hat{S}_{k,l-1}] - [\hat{S}_{k,l} - \hat{P}_{k,l,l-1}]$$

and this vanishes when

$$\sigma = [\hat{S}_{k,l} - \hat{P}_{k,l,l-1}]/[\hat{S}_{k,l} - 2P_{k,l,l-1} + \hat{S}_{k,l-1}].$$

The other aspects of the analysis remain as above. This procedure is one actually used in FORTRAN SUBROUTINE PERIOD (LMAX, LMIN, GAMMA, H, PER, Y), whose listing is included here and which forms the basis for the numerical experiments carried out in § 5.

```
C SUBROUTINE PERIOD GIVES AN ESTIMATE OF THE PERIOD OF A
C SIGNAL, ETA(T), BASED ON DISCRETE SAMPLES, Y, WITH SAMPLING
C INTERVAL H.  THE PERIOD ESTIMATE, PER, IS PROVIDED AS EACH NEW
C SAMPLE IS OBTAINED AND APPROXIMATELY MINIMIZES THE SUM FROM
C K= -1 TO (-INFINITY) OF (GAMMA**(-K))*(Y(T(J-K))-Y(T(J-K)-TT))
C **2 FOR TT IN THE INTERVAL [LMIN*H, LMAX*H].  HERE GAMMA IS A
C DISCOUNTING FACTOR BETWEEN 0 AND 1 AND THE SAMPLING INSTANTS
C ARE T(J).

      SUBROUTINE PERIOD(LMAX, LMIN, GAMMA, H, PER, Y)
      DIMENSION S(40), P(40), RHO(40)
      INTEGER L, LMIN, LMAX, LMINP1, LM1
      LMAXM1 = LMAX - 1
```

```
C UPDATA SAMPLE
      ETAOLD = ETANEW
      ETANEW = Y
C COMPUTE S, P, RHO AS PER METHOD DESCRIPTION
      DO 15 L = 1,LMAXM1
      KL = LMAX - L + 1
      KLM1 = KL - 1
   15 RHO(KL) = ETANEW - ETAOLD + RHO(KLM1)
      RHO(1) = ETANEW - ETAOLD
      DO 16 L = 1,LMAX
   16 S(L) = GAMMA*S(L) + (RHO(L))**2
      DO 17 L = 1,LMAXM1
      LP1 = L+1
   17 P(L) = GAMMA*P(L) + RHO(L)*RHO(LP1)
C COMPUTE NEW PERIOD ESTIMATE, PER, ACCORDING TO OPTIMALITY CRITERION.
      PER = FLOAT(LMIN)*H
      SMINI = S(LMIN)
      LMINP1 = LMIN+1
      DO 26 L = LMINP1,LMAX
      LM1 = L-1
      IF(S(LM1).LT.SMINI)GO TO 21
      GO TO 22
   21 PER = FLOAT(LM1)*H
      SMINI = S(LM1)
   22 IF(P(LM1).GT.S(LM1))GO TO 24
      IF(P(LM1).GT.S(L))GO TO 24
      QUO = S(L)+S(LM1)-2.*P(LM1)
C THE NEXT STATEMENT IS MACHINE DEPENDENT (VAX 11-780). IT MAY
C NEED TO BE MODIFIED FOR GENERAL APPLICATION.
      IF(QUO.LT..000001)GO TO 26
C VALSIG IS THE NEW CANDIDATE MINIMAL VALUE FOR THE OBJECTIVE
C FUNCTION AT THE POINT (LMAX-(L-SIG))*H
      SIG = (S(L)-P(LM1))/QUO
      VALSIG = SIG*SIG*S(LM1)+(1.-SIG)*(1.-SIG)*S(L)+2.*SIG*
     1(1.-SIG)*P(LM1)
      IF(VALSIG.LT.SMINI)GO TO 23
      GO TO 24
   23 PER = (FLOAT(L)-SIG)*H
      SMINI = VALSIG
   24 IF(S(L).LT.SMINI)GO TO 25
      GO TO 26
   25 PER = FLOAT(L)*H
      SMINI = S(L)
   26 CONTINUE
      RETURN
      END
```

If there is some danger of confusing the minimal period T with one of its multiples $2T$, $3T$, etc., this can usually be overcome by specifying LMIN, LMAX correctly.

5. Some numerical experience with period. We have carried out extensive computer based simulations using SUBROUTINE PERIOD described in the preceding section. Here we will describe results obtained in connection with the plant

$$\begin{pmatrix} \dot{x}^1 \\ \dot{x}^2 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 0 & -.2 \end{pmatrix} \begin{pmatrix} x^1 \\ x^2 \end{pmatrix} + \begin{pmatrix} 0 \\ 28 \end{pmatrix} (u+v) \tag{5.1}$$

with

$$v(t) = z^1(t), \tag{5.2}$$

$$\begin{pmatrix} \dot{z}^1 \\ \dot{z}^2 \end{pmatrix} = \begin{pmatrix} 0 & \alpha \\ -\alpha & 0 \end{pmatrix} \begin{pmatrix} z^1 \\ z^2 \end{pmatrix}. \tag{5.3}$$

The rather nondescript parameters appearing in (5.1) result from the fact that this damped inertial system is a model for a certain physical plant of interest. A compensator was constructed in the form (0.7) using $S=-\varepsilon P_0^*$ with P_0 as in (1.10). To provide a more or less standard basis of comparison the feedback coefficients in all cases were chosen to achieve critical damping (i.e., multiple real eigenvalues) at various rates in the closed loop matrix

$$A+CK=\begin{pmatrix} 0 & 1 \\ 28k_1 & -.2+28k_2 \end{pmatrix}. \tag{5.4}$$

The feedback coefficients k_1 and k_2 are identified as FB1 and FB2, respectively, on the accompanying figures. In this very simple example

$$B=\begin{pmatrix} 0 & 0 \\ 28 & 0 \end{pmatrix}, \quad C=\begin{pmatrix} 0 \\ 28 \end{pmatrix}, \quad -L=(1,0)$$

and Φ is chosen to be the 2×2 identity matrix. For the values of k_1 and k_2 which were used P_0 turns out to be a very small matrix and we used $\varepsilon=1000$.

The output used for the period estimation was $w(t)=x^1(t)+y^1(t)$ and the output shown on the diagrams is $x^1(t)$. It will be seen that the initial estimates for the period $T(t)$ are wildly inaccurate but, in the cases when the complete plant/compensator/identifier system is asymptotically stable, the estimate $T(t)$ converges to the correct value T_0 (within the accuracy permitted by the approximations inherent in PERIOD, as described in the preceding section). In all cases we selected $z^1(0)=1$, $z^2(0)=0$, so that

$$z^1(t)=\cos(\alpha t), \qquad z^2(t)=-\alpha\sin(\alpha t).$$

In Figs. 1-16 the odd numbered figures show the output $x^1(t)$ while the next, even numbered figure in each case shows the period estimate $T(t)$ for the same run.

Figures 1 through 4 correspond to choices of k_1 and k_2 such that the matrix (5.4) has a double eigenvalue $\lambda=-5$. In Figs. 1 and 2 $\alpha_0=20\pi$ (10 Hertz) corresponding to $T_0=.1$, indicated by the dotted line. Figures 3 and 4 illustrate the corresponding experience for $\alpha_0=30\pi$ (15 Hertz), or $T_0=.0667$. Here the period identifier diverges from the correct value and, as seen in Fig. 3, no significant reduction of the oscillation of $x^1(t)$ is realized. We believe that this is accounted for by the fact that the term $2\pi/(T_0)^2$ in (2.21) changes from 200π in the 10 Hz case to 450π in the 15 Hz case.

Figures 5 through 8 show the 10 Hz case with k_1 and k_2 chosen so that (5.4) has a double eigenvalue $\lambda=-8$ (Figs. 5 and 6) and with k_1 and k_2 chosen so that $\lambda=-9$ (Figs. 7 and 8). These cases seem quite satisfactory with rapid attenuation of the oscillation in $x^1(t)$, better in the second case than in the first, and rapid convergence of the period estimate $T(t)$ to $T_0=.1$. The corresponding experience in the 15 Hz case is not nearly so satisfactory. Figures 9 and 10 show the performance for $\lambda=-8.5$ while Figs. 11 and 12 show $\lambda=-10$. We see from Figs. 10 and 12 that, although the value T_0 is unstable, the estimate $T(t)$ tends to undergo a self-excited oscillation about the equilibrium value indicated by the dotted lines. The evidence favors the conjecture that in these cases the nonlinear equation (2.11) may exhibit a Hopf type bifurcation as the parameter T_0 passes from .1 (10 Hz) to .0667 (15 Hz). Detailed analysis of this possibility must await later treatment.

Figures 13 through 16 show experience in the 15 Hz case with k_1 and k_2 selected so that $\lambda=-14$ (Figs. 13 and 14) and so that $\lambda=-20$ (Figs. 15 and 16). We see that the performance improves as (5.4) is made progressively more stable, in agreement with the conjectures of § 3. The small residual oscillation evident in Fig. 15 is probably

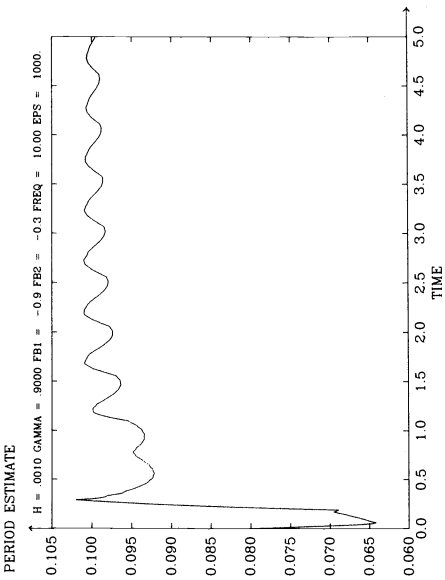


FIG. 2

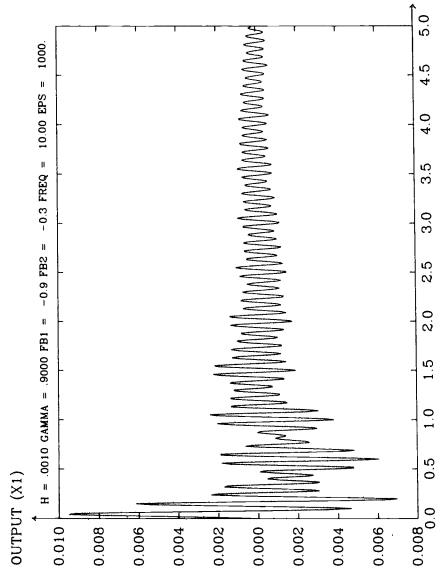


FIG. 1

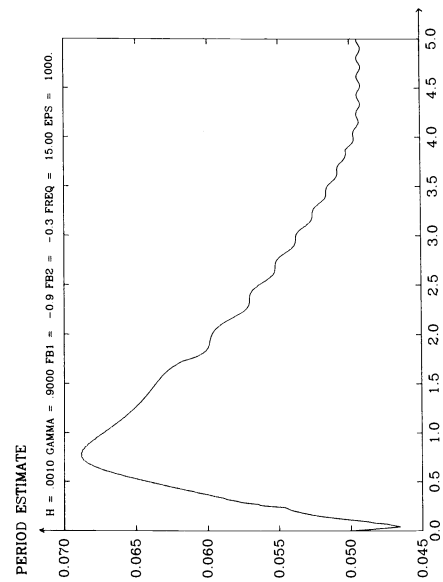


FIG. 4

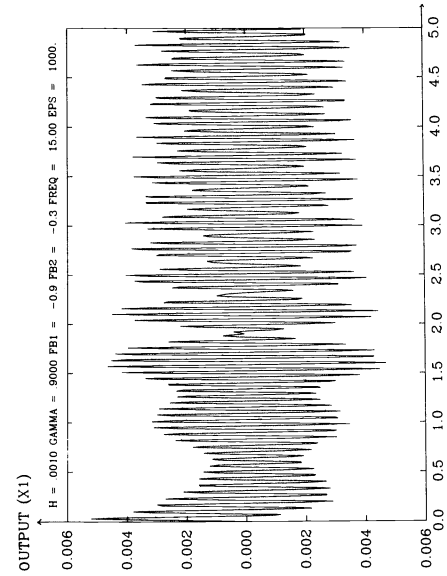


FIG. 3

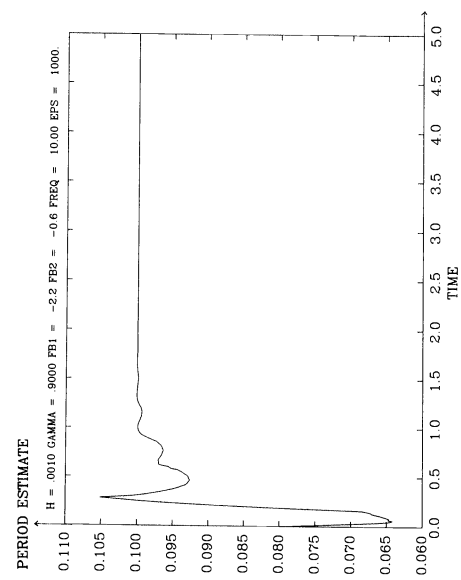


FIG. 6

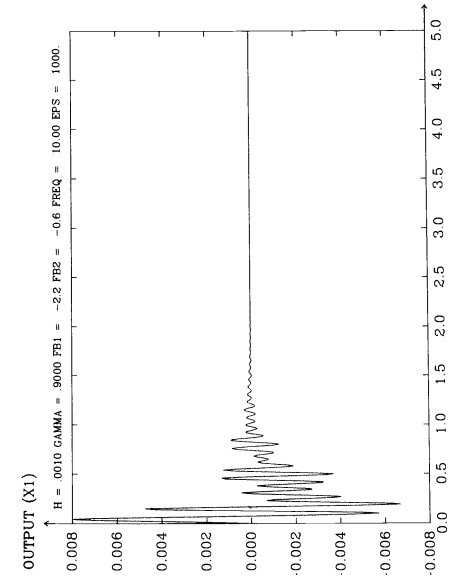


FIG. 5

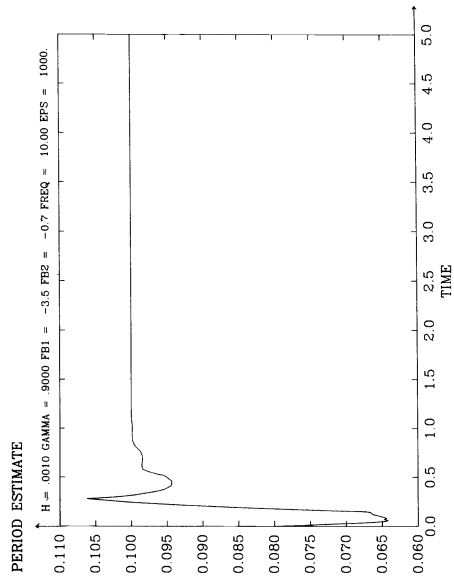


FIG. 8

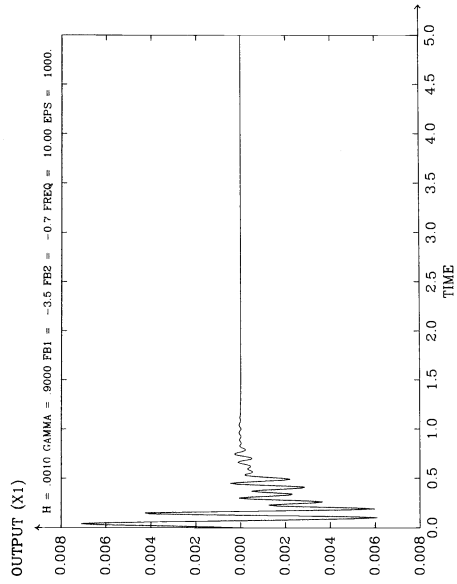


FIG. 7

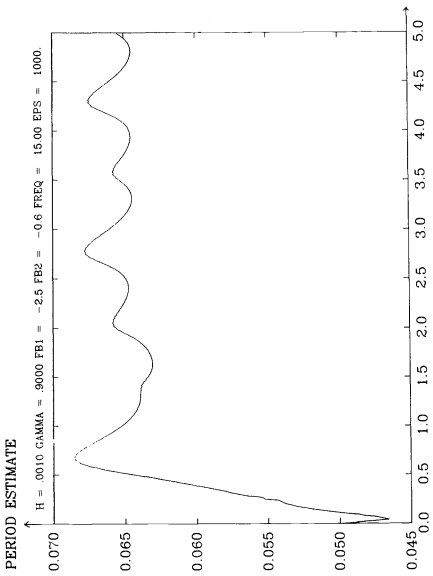


FIG. 10

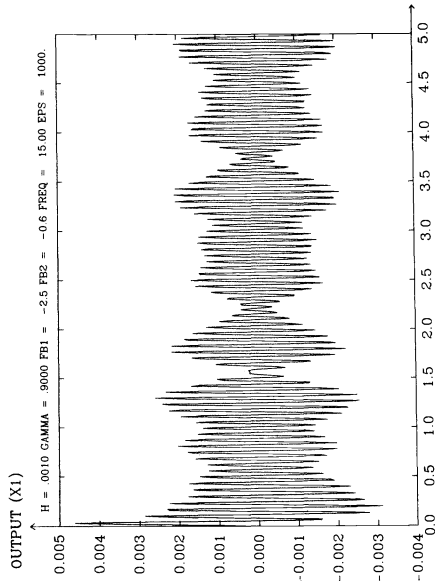


FIG. 9

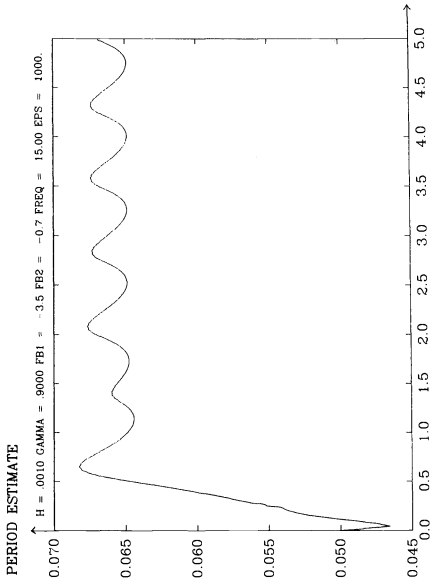


FIG. 12

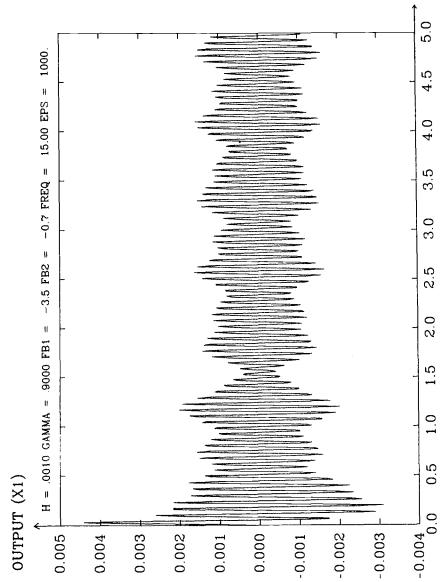


FIG. 11

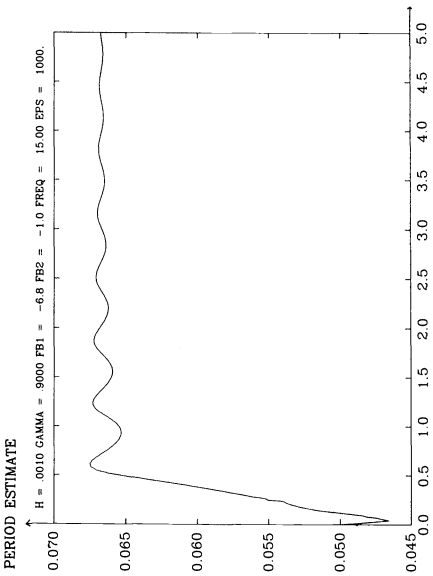


FIG. 14

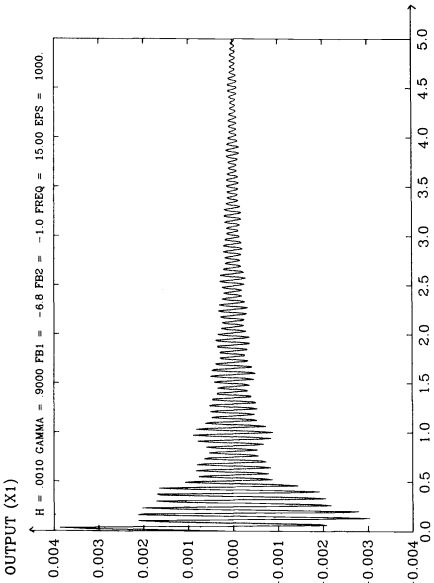


FIG. 13

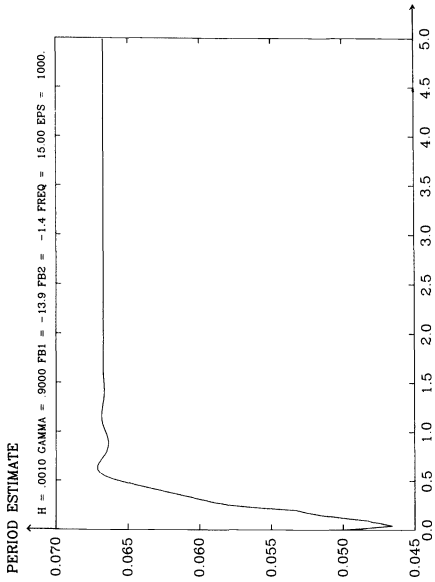


FIG. 16

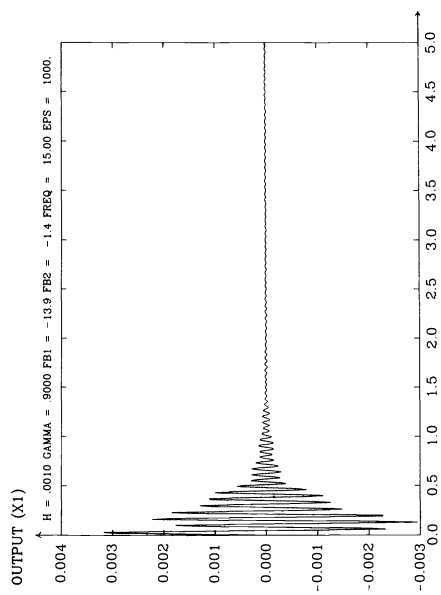


FIG. 15

due to the fact that PERIOD does not provide an exact estimate even when the corresponding continuous process associated with minimization of (2.4) or (2.6) is asymptotically stable.

Figures 17 and 18 show variational ΔT solutions, obtained in the manner described at the end of § 3, for the 15 Hz case with $\lambda = -9$ and $\lambda = -14$, respectively. Because $T(t)$ does not converge to T_0 in the first case (cf. Fig. 12), even the variational solutions are not sinusoidal.

In Fig. 18, corresponding to $\lambda = -14$, $T(t)$ converges to $T_0 = .0667$ (cf. Fig. 14) and the corresponding variational solution tends to zero in a convincing exponentially damped sinusoidal manner, this behavior becoming more convincing as t increases. It is of interest to estimate the frequencies and damping factors here and compare them with the eigenvalues of $M(\alpha_0)$, $F(\alpha_0)$. Analyzing the data plotted in Fig. 18 one obtains the estimate $T = .66$ and, comparing the successive amplitudes, we see that the oscillation there is approximated by

$$C_+ e^{(-.525+i9.52)t} + C_- e^{(-.525-i952)t}. \tag{5.5}$$

Here

$$M(\alpha_0) = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -190.12 & -27.188 & -28 & 0 \\ -2.966 & 82.417 & 0 & 1 \\ -.009 & -2.966 & -8873.8 & 0 \end{pmatrix}$$

and its eigenvalues may be computed to be

$$-2.77 \pm i105.32, \qquad -10.82 \pm i5.9.$$

None of these corresponds to (5.5) and we conclude that the dynamics exhibited in the identification process arise from a different source which, on the basis of our earlier investigations in this paper, we believe to be the functional equation (2.22) (equivalently (2.19), (2.20)).

In Figs. 19 through 26 we indicate the effect of varying the parameters γ and ε (cf. (1.9), (2.4)) while leaving the feedback parameters in K fixed at the values which produced Figs. 13 and 14 with $\gamma = .9$ and $\varepsilon = 1000$. This corresponds to the double eigenvalue $\lambda = -14$ for (5.4). Here it needs to be explained that the value "GAMMA" referred to on the figure heading corresponds to $e^{-\gamma h}$, where h is the length of the sampling interval used by PERIOD. For all cases studied here h is ten times the H value shown in the figure heading; thus $h = .01$ and

$$\text{GAMMA} = .8 = e^{-.01\gamma} \to \gamma = \frac{\log_e .8}{-.01} = -22.3,$$

$$\text{GAMMA} = .9 \to \gamma = -10.5,$$

$$\text{GAMMA} = .95 \to \gamma = -5.13.$$

As we see by comparison of Figs. 19 and 20 with 13 and 14, performance is substantially degraded by discounting past values too much; $\gamma = -5.13$ gives better performance than $\gamma = -10.5$.

For Figs. 23 and 24 we have set GAMMA = .9 again but have increased ε to 2000 rather than the earlier 1000. The improvement over the results in Figs. 13 and 24 is

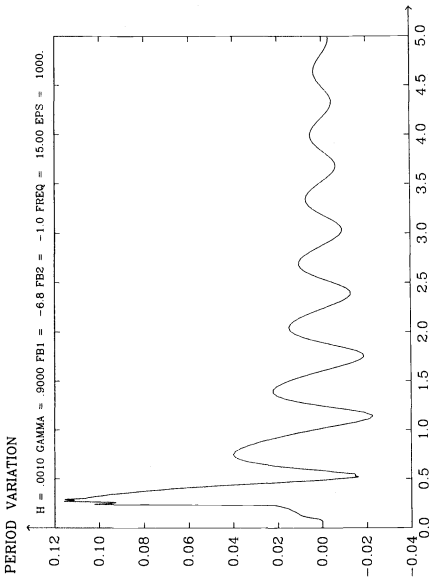


FIG. 18

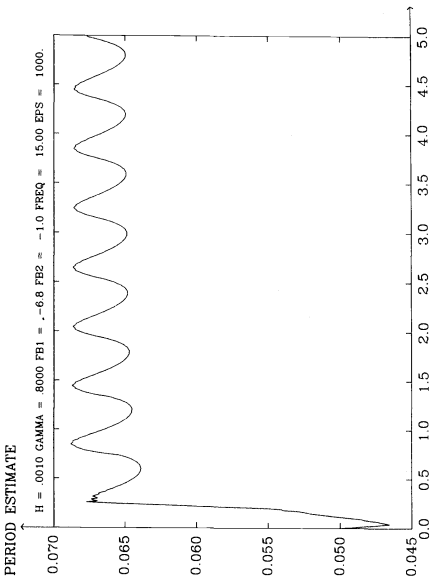


FIG. 20

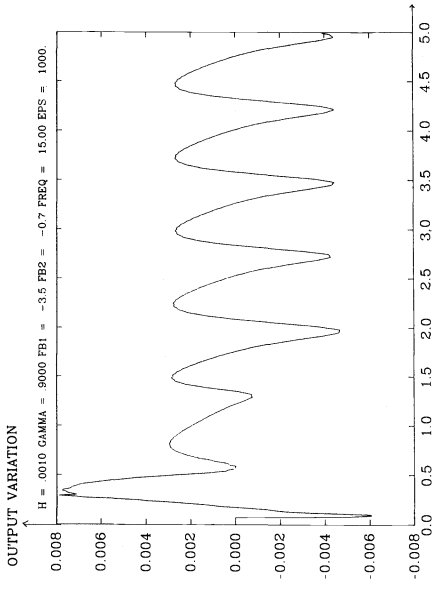


FIG. 17

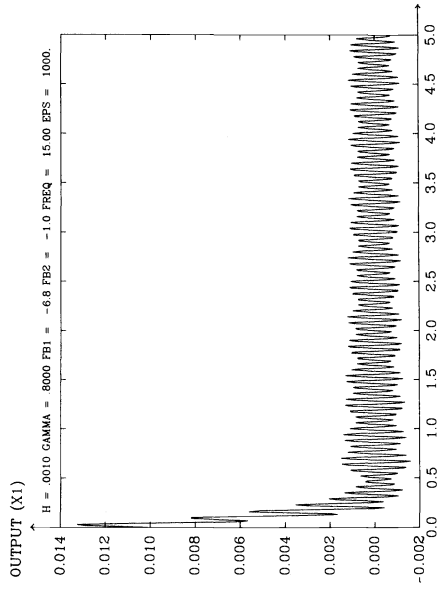


FIG. 19

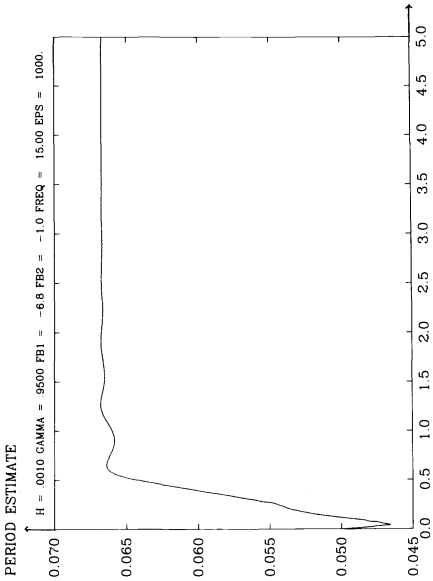


FIG. 22

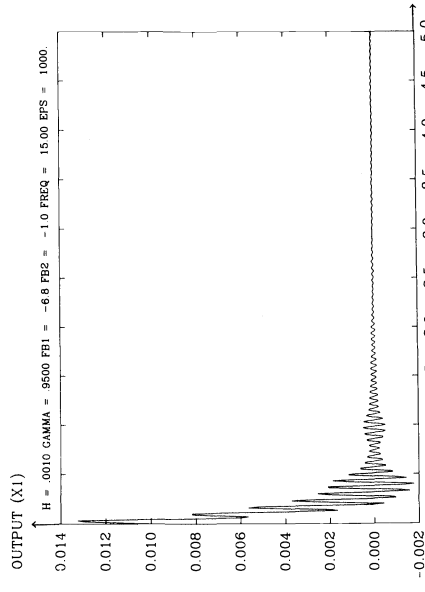


FIG. 21

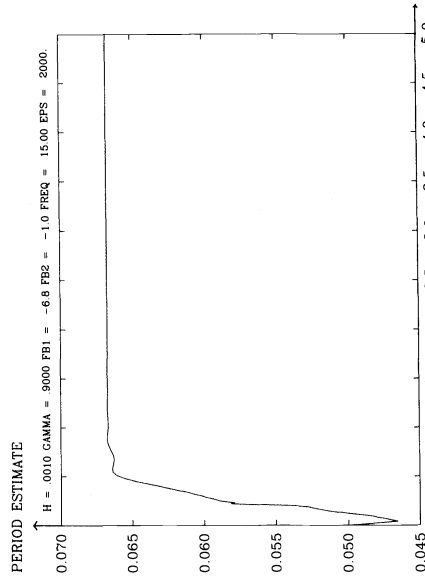


FIG. 24

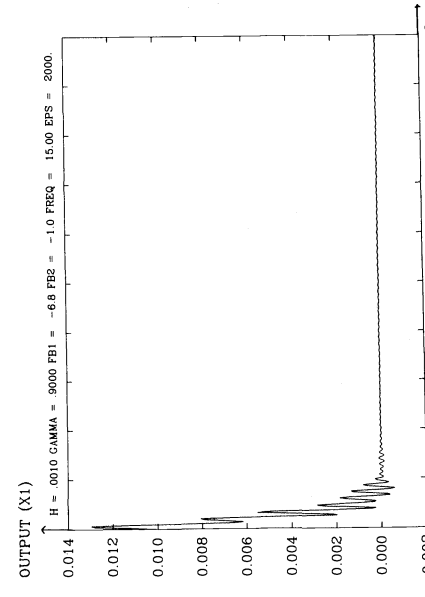


FIG. 23

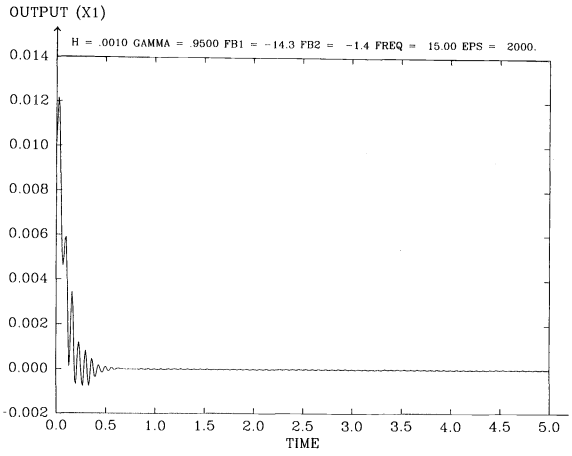


FIG. 25

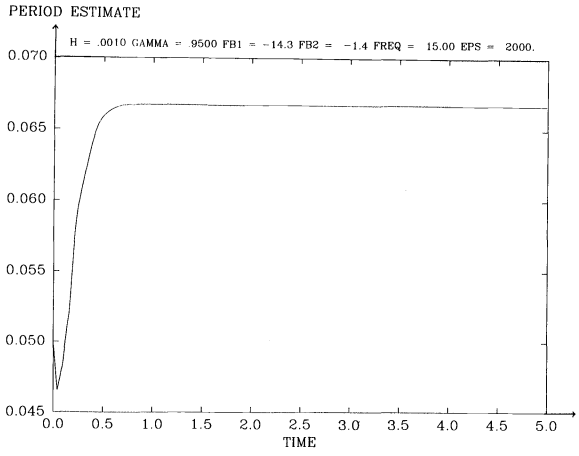


FIG. 26

again quite marked. To obtain the results of Figs. 25 and 26 we "pulled out all the stops" and set GAMMA = .95, $\varepsilon = 2000$ and chose feedback parameters corresponding to a double eigenvalue $\lambda = -20$ for (5.4). Here we finally obtain the sort of sinusoidal disturbance rejection one would hope for.

To summarize our work: we have presented a method, based on a certain optimality criterion, for identification of the period of a periodic disturbance affecting a finite-dimensional linear dynamic disturbance decoupling procedures. We have shown that the linearized dynamics of the plant/compensator/estimator in the neighborhood of the equilibrium trajectory are governed by a certain Volterra type functional equation with periodic coefficients. That functional equation has been analyzed, in brief, and its stability characteristics have been seen to be determined by the location of the characteristic exponents of certain solutions of "Floquet type." Numerical studies have been presented indicating how variation of the design parameters of the system affect the overall performance and indications for needed future research have been given.

6. Appendix: Sketch of the proof of Theorem 5. Let $z(t)$, $W(t, s)$, etc., be as in the statement of Theorem 5. For any $t \geqq 0$ we have

$$\begin{aligned} z(t) &= \int_{-\infty}^{t} W(t, s)z(s)\, ds \\ &= \int_{kT}^{t} W(t, s)z(s)\, ds + \int_{0}^{kT} W(t, s)z(s)\, ds + \int_{-\infty}^{0} W(t, s)\tilde{z}(s)\, ds, \end{aligned} \tag{6.1}$$

where k is the largest integer such that $kT \leqq t$. We define $z_l \in L_m^2[0, T]$ by

$$z_l(\tau) = z((l-1)T + \tau), \quad \tau \in [0, T], \quad l \geqq 1,$$

and we define $\tilde{z}_{-l} \in L_m^2[0, t]$ by

$$\tilde{z}_{-l}(\tau) = \tilde{z}(-(l+1)T + \tau), \quad \tau \in [0, T], \quad l \geqq 0.$$

Then, with

$$t = kT + \tau, \quad s = lT + \sigma, \quad s \in [lT, (l+1)T],$$

(6.1) yields

$$\begin{aligned} z_{k+1}(\tau) &- \int_{0}^{\tau} W(kT + \tau, kT + \sigma)z_{k+1}(\sigma)\, d\sigma \\ &- \sum_{j=1}^{k} \int_{0}^{T} W(kT + \tau, (j-1)T + \sigma)z_j(\sigma)\, d\sigma \\ &= \sum_{l=0}^{\infty} \int_{0}^{T} W(kT + \tau, -(l+1)T + \sigma)\tilde{z}_{-l}(\sigma)\, d\sigma. \end{aligned} \tag{6.2}$$

Using the periodicity relation (3.7) we have

$$W(kT + \tau, (j-1)T + \sigma) = W(\tau, -(k-j)T + \sigma) \equiv W_{k-j}(\tau, \sigma)$$

for any integer $j \leqq k$. Then (6.2) becomes

$$\begin{aligned} z_{k+1}(\tau) &= \int_{0}^{\tau} W_0(\tau, \sigma)z_{k+1}(\sigma)\, d\sigma - \sum_{j=1}^{k} \int_{0}^{T} W_{k+1-j}(\tau, \sigma)z_j(\sigma)\, d\sigma \\ &= \sum_{l=0}^{\infty} \int_{0}^{T} W_{k+1+l}(\tau, \sigma)\tilde{z}_{-l}(\sigma)\, d\sigma. \end{aligned} \tag{6.3}$$

The conditions (3.7), (3.8) show that the last sum converges in $L_m^2[0, T]$. We may write (6.3) as a vector linear recursion equation in $L_m^2[0, T]$

$$(I - P_0)z_{k+1} - \sum_{j=1}^{k} P_{k+1-j}z_j = \sum_{l=0}^{\infty} P_{k+1+l}\tilde{z}_{-l} \tag{6.4}$$

where

$$(P_0 z)(\tau) = \int_{0}^{\tau} W_0(\tau, \sigma)z(\sigma)\, d\sigma,$$

$$(P_k z)(\tau) = \int_{0}^{T} W_k(\tau, \sigma)z(\sigma)\, d\sigma, \qquad k = 1, 2, 3, \cdots.$$

As is well known [3], $I - P_0$ is boundedly invertible, and $P_0, P_1, P_2, \cdots$ are all compact. Then

$$Q_{-k} = -(I - P_0)^{-1}P_k, \qquad k = 1, 2, 3, \cdots \tag{6.5}$$

are all compact and, keeping (3.7) in mind, it may be seen that for some positive number D

$$\|Q_{-k}\| \leqq D e^{-kcT}, \qquad k = 1, 2, 3, \cdots. \tag{6.6}$$

Then (6.4) can be written, with an obvious re-indexing, as

$$z_k + \sum_{l=-k+1}^{-1} Q_l z_{l+k} + \sum_{l=-\infty}^{-k} Q_l \tilde{z}_{l+k}, \qquad k = 1, 2, 3, \cdots. \tag{6.7}$$

Given a sequence $\{y_k | -\infty < k < \infty\} CH$, where H is a Hilbert space, and supposing that

$$\|y_k\| \leqq M^+(\gamma+)^k, \qquad k = 1, 2, 3, \cdots, \tag{6.8}$$

$$\|y_k\| \leqq M^-(\gamma-)^k, \qquad k = 0, -1, -2, -3, \cdots, \tag{6.9}$$

where M^+, M^-, γ^+, γ^- are all positive numbers and $\gamma^+ \geqq \gamma^-$, we define the bilateral "Z-transform" (discrete Laplace transform) of $\{y_k\}$ by

$$\eta(\gamma) = \begin{cases} \sum_{k=1}^{\infty} y_k \lambda^{-k} \equiv \eta^+(\lambda), & |\lambda| > \gamma^+, \\ -\sum_{k=0}^{\infty} y_{-k} \lambda^k \equiv \eta^-(\lambda), & |\lambda| < \gamma^-. \end{cases} \tag{6.10}$$

Clearly, $\eta(\lambda)$ is analytic in neighborhoods of both 0 and ∞. In many cases $\eta^+(\lambda)$ and $\eta^-(\lambda)$ are analytic continuations of each other. For example, if for all integers k

$$y_k = \mu^k y_0, \quad y_0 \in H, \quad \mu \neq 0,$$

then with $\gamma^+ = \gamma^- = |\mu|$, $M^+ = M^- = 1$, all of the above are valid for

$$\eta^+(\lambda) = \eta(\lambda), \quad |\lambda| > |\mu|, \quad \eta^-(\lambda) = \eta(\lambda), \quad |\lambda| < |\mu|, \quad \eta(\lambda) = \frac{1}{\lambda - \mu}.$$

If, correspondingly, $\{Q_k | -\infty < k < \infty\}$, is a sequence of bounded operators on H such that

$$\|Q_k\| \leqq B^+(\rho^+)^{-k}, \qquad k = 1, 2, 3, \cdots, \tag{6.11}$$

$$\|Q_{-k}\| \leqq B^-(\rho^-)^{-k}, \qquad k = 0, -1, -2, \cdots, \tag{6.12}$$

where B^+, B^-, ρ^+, ρ^- are all positive (and $\rho^+ > \gamma^+$, $\rho - < \gamma^-$ in our application), we may define the discrete Fourier transform of $\{Q_k\}$ by

$$Q(\lambda) = \sum_{k=-\infty}^{\infty} Q_k \lambda^k. \tag{6.13}$$

Clearly the series converges and $Q(\lambda)$ is a holomorphic operator valued function for $\rho^- < |\lambda| < \rho^+$. If $Q_k = 0$ for all positive k, then ρ^+ may be taken to be ∞ and $Q(\lambda)$ will be homomorphic for $|\lambda| > \rho^-$, including $\lambda = \infty$.

The convolution of $\{Q_k\} \equiv Q$ and $\{y_k\} \equiv y$ is defined by

$$f_l \equiv (Q^* y)_l = \sum_{k=-\infty}^{\infty} Q_k y_{k+l}, \tag{6.14}$$

the sum being convergent when (6.8), (6.9), (6.11) and (6.12) apply and $\rho^+ > \gamma^+$, $\rho^- < \gamma^-$, as we suppose. To anyone familiar with transforms of this type the first question occurring concerns the relationship of the transform $\phi(\lambda)$ of $\{f_l\}$ to the

transforms $Q(\lambda)$, $\eta(\lambda)$. The answer is easy but not completely obvious. Let Γ^+ and Γ^- be positively oriented circles centered at $\lambda = 0$ with radii r^+, r^-, $\gamma^+ < r^+ < \rho^+$, $\rho^- < r^- < \gamma^-$. Then, as we show in the more complete discussion [7], if $\Gamma = \Gamma^+ - \Gamma^-$ and λ lies in the exterior of the annular regions between Γ^+ and Γ^-,

$$\phi(\lambda) = \frac{1}{2\pi i} \int_\Gamma \frac{Q(\zeta)\eta(\zeta)\, d\zeta}{\lambda - \zeta}. \tag{6.15}$$

This condition is necessary and sufficient for (6.14) to be true.

When $Q_k = 0$ for k positive and Q_0 has a bounded inverse, (6.14) becomes

$$\sum_{k=-\infty}^{0} Q_k y_{k+l} = f_l \tag{6.16}$$

and, given the initial values

$$y_{-l} = \tilde{y}_{-l}, \qquad l = 0, 1, 2, 3, \cdots, \tag{6.17}$$

and $f_1, f_2, f_3, \cdots$, we can compute $y_1, y_2, y_3, \cdots$. The case of interest to us is the homogeneous case $f_1 = f_2 = f_3 = \cdots = 0$. Here it may be seen that with

$$\tilde{\eta}(\lambda) = \sum_{l=0}^{\infty} \tilde{y}_{-l}\lambda^l, \qquad \eta(\lambda) = \sum_{k=1}^{\infty} y_k \lambda^{-k}, \tag{6.18}$$

and for $|\lambda| > r^-$, λ not a singularity of $Q(\lambda)$

$$\eta(\lambda) = \frac{1}{2\pi i} Q(\lambda)^{-1} \int_{\Gamma^-} \frac{Q(\zeta)\tilde{\eta}(\zeta)\, ds}{\lambda - \zeta}. \tag{6.19}$$

Since $Q(\lambda)$ is analytic at $\lambda = \infty$ and $Q(\infty) = Q_0$, which is nonsingular, if we take r^+, the radius of Γ^+, so large that all singularities of $Q(\lambda)$ are included in the interior of Γ^+, then the individual y_k, $k = 1, 2, 3, \cdots$ may be recovered via

$$y_k = \frac{1}{2\pi i} \int_{\Gamma^+} \eta(\lambda)\lambda^{k-1}\, d\lambda, \qquad k = 1, 2, 3, \cdots. \tag{6.20}$$

With the use of formula (6.16) the proof of Theorem 5 may be completed. In our application $H = L_m^2[0, T]$, $Q_k = 0$ for k positive, $Q_0 = I, Q_{-1}, Q_{-2}, Q_{-3}, \cdots$ are all compact and the series (6.13) converges uniformly for $|\lambda| \geqq r^- + \delta$ for any $\delta > 0$. It is known [1], [5] that the singularities of $Q(\lambda)$ must be isolated in any such region and, for each such singularity λ_k the null space of $Q(\lambda_k)$ must be finite-dimensional. Let Γ_δ^- be the circle centered at zero with radius $r^- + \delta$. We may assume that Γ_δ^- meets no singularities of $Q(\lambda)$. Then, applying (6.16) to the z_k of (6.7) we have

$$\begin{aligned} z_k &= \frac{1}{2\pi i} \int_{\Gamma_\delta} \eta(\lambda)\lambda^{k-1}\, d\lambda + \frac{1}{2\pi i} \int_{\Gamma_\delta^-} \eta(\lambda)\lambda^{k-1}\, d\lambda \\ &= z_{k,F} + z_{k,r^-+\delta}, \qquad k = 1, 2, 3, \cdots, \end{aligned} \tag{6.21}$$

where $\Gamma_\delta = \Gamma^+ - \Gamma_\delta^-$. From (6.19) and (6.21) it is clear that

$$\| z_{k,r^-+\delta} \| \leqq \tilde{M}(r^- + \delta)^k$$

where $\tilde{M}$ is a constant which may be bounded in terms of $\tilde{\eta}$, hence in terms of the $\tilde{z}_{-l}$, $l = 0, 1, 2, \cdots$. On the other hand,

$$z_{k,F} = \sum_{\lambda_2 \in \operatorname{Int} \Gamma_\delta} \lambda_j^k \operatorname{Res} \eta(\lambda_j).$$

In the case of a simple eigenvalue λ_j with one-dimensional null space, which is all we will study here, from the formula (6.19) for η it may be seen that

$$\text{Res } \eta(\lambda_j) = \frac{1}{(\psi_j, Q^*(\lambda_j)\phi_j)} \int_{\Gamma^-} \frac{(\psi_j, Q(\zeta)\tilde{\eta}(\zeta)\, ds}{\lambda_2 - \zeta} \phi_j \tag{6.22}$$

where ϕ_j is a nonzero vector in the one-dimensional null space of $Q(\lambda_j)$ and ψ_j is a corresponding vector in the null space of $Q(\lambda_j)^*$ such that $(\psi_j, \phi_j)_{L_m^2[0,T]} = 0$. We see in any case that $z_{k,F}$ is a sum of the form

$$z_{k,F} = \sum_{\lambda_j \in \text{Int } \Gamma_\delta} \lambda_j^k C_j \phi_j, \qquad k = 1, 2, 3, \cdots. \tag{6.23}$$

The corresponding solutions $z_F(t)$, $z_\beta(t)$ of (3.6) (or (6.1)) are obtained by inverting the transformations which follow (6.1). The term $e^{-\beta T}$ of (3.11) is identified with $r^- + \delta$. It is greater than e^{-CT} which is identified with r^-. Thus $z_\beta(t)$ satisfies (3.11).

Since $z_F(t)$ is a solution of (6.2), the form of that equation shows that $z_F(t)$ must be a continuous function. The form (6.23) then implies that z_F on any interval $[kT, (k+1)T]$ is λ_j times the corresponding value of z_F on $[(k-1)T, kT]$. From this it is clear that

$$\phi_j(t) = \lambda_j \phi_j(0).$$

We identify $e^{\lambda T}$ in (3.12) with λ_j and $P(T)$ with the T-periodic extension of $e^{-\lambda t}\phi_j(t)$ (which satisfies

$$P(T) = e^{-\lambda T}\phi_j(T) = \lambda_j^{-1}\phi_j(T) = \phi_j(0) = e^{-\lambda 0}\phi_j(0) = P(0)).$$

Thus, modulo the usual remarks which must apply to nonsimple poles of $Q(\lambda)^{-1}$, which lead to solutions of the form (3.13), we have completed the proof of Theorem 5. Further details may be found in [7]. The main point of the theorem is that the dominant solutions of (3.6) (or (6.1)) are those associated with the larger singularities of $Q(\lambda)$ and those solutions are of the type (3.12) or (3.13).

Acknowledgment. I would like to express my thanks to Dr. Norman Coleman, Jr. of ARRADCOM, Dover, New Jersey, for background information and helpful conversations which stimulated the development of this work.

REFERENCES

[1] F. V. ATKINSON, *A spectral problem for completely continuous operators*, Acta Math. Acad. Sci. Hungar., 3 (1952), pp. 53-60.

[2] F. R. GANTMACHER, *Theory of Matrices*, Chelsea, New York.

[3] JACK HALE, *Theory of Functional Differential Equations*, Springer-Verlag, New York, 1977.

[4] D. HENRY, *The adjoint of a linear functional differential equation and boundary value problems*, J. Differential Equations, 9 (1971), pp. 55-66.

[5] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966 (see pp. 365 ff).

[6] Y. D. LANDAU, *Adaptive Control: the Model Reference Approach*, Marcel Dekker, New York, 1979.

[7] D. L. RUSSELL, *A Floquet decomposition for Volterra equations with periodic kernel and a transform approach to linear recursion equations*, Math. Res. Ctr., Univ. Wisc., Madison, Techn. Summ. Rept. 2824, June 1985. (Submitted to J. Differential Equations.)

[8] ———, *Mathematics of Finite Dimensional Control Systems; Theory and Design*, Marcel Dekker, New York, 1979.

[9] A. P. SAGE AND J. L. MELSA, *System Identification*, Academic Press, New York, 1971.

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

[11] W. A. WOLOVICH, *Linear Multivariable Systems*, Vol. 11, Applied Mathematical Sciences, Springer-Verlag, New York-Heidelberg-Berlin, 1974.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 6, November 1986

013

CONTROL OF AN ELLIPTIC PROBLEM WITH POINTWISE STATE CONSTRAINTS*

EDUARDO CASAS†

Abstract. This paper deals with a quadratic control problem for elliptic equations with pointwise state constraints. Existence and uniqueness of the solution is proved. Optimality conditions are given and regularity of the optimal solution is investigated.

Key words. quadratic control problems, pointwise state constraints, elliptic equations, optimality conditions, Borel measures, Sobolev spaces, Lagrange multipliers

AMS(MOS) subject classification. 49B22

1. Introduction. This paper is concerned with distributed control problems with restrictions on the control u as well as on the state y. Our main interest is the derivation of optimality conditions and regularity results.

Let $\Omega \subset R^n$ $(1 \leqq n \leqq 3)$ be an open and bounded set whose boundary Γ is Lipschitz continuous (Nečas [17]) and Ω is locally on one side of Γ.

Let us consider the following differential operator

$$Ay = -\sum_{i,j=1}^{n} \partial_{x_j}(a_{ij}\partial_{x_i} y) + a_0 y \tag{1.1}$$

where

$$\begin{aligned} &a_0 \in L^\infty(\Omega), \quad a_0(x) \geqq 0 \quad \text{a.e. } x \in \Omega, \\ &a_{ij} \text{ is Lipschitz on } \bar{\Omega} \qquad (1 \leqq i, j \leqq n), \\ &\sum_{i,j=1}^{n} a_{ij}(x)\xi_i\xi_j \geqq \alpha|\xi|^2, \quad \alpha > 0, \quad \forall x \in \Omega, \quad \forall \xi \in R^n. \end{aligned} \tag{1.2}$$

Our aim is to consider a control problem for the system governed for the following equation

$$\begin{aligned} &Ay = v \quad \text{on } \Omega, \\ &y = 0 \qquad \text{on } \Gamma. \end{aligned} \tag{1.3}$$

It is well known (Nečas [17]) that the problem (1.3) has a unique solution $y(v) \in H_0^1(\Omega)$ for each $v \in L^2(\Omega)$, where $H^m(\Omega)$ and $H_0^m(\Omega)$, m being an integer, are the usual Sobolev spaces on Ω.

We will assume that the problem (1.3) satisfies the following regularity condition: for each $v \in L^2(\Omega)$, $y(v) \in H^2(\Omega)$ and

$$\|y(v)\|_{H^2(\Omega)} \leqq c\|v\|_{L^2(\Omega)} \tag{1.4}$$

where $c \in R$.

The condition (1.4) is satisfied, for example, if either $\Omega \in R^{(1),1}$ (Nečas [17]) or Ω is a polygonally $(n=2)$ or polyhedrally $(n=3)$ convex set open (Grisvard [10]).

* Received by the editors March 20, 1984, and in final revised form July 22, 1985.
† Departamento de Ecuaciones Funcionales, Facultad de Ciencias, 39005-Santander, Spain.

Let K be a nonempty, convex and closed subset of $L^2(\Omega)$ and let J be the functional

$$J(v)=\frac{1}{2}\int_\Omega (y(v)-y_0)^2\,dx+\frac{r}{2}\int_\Omega v^2(x)\,dx \tag{1.5}$$

where y_0 is a fixed element of $L^2(\Omega)$ and $r\geqq 0$.

We consider the following control problem:

$$\text{(P)}\qquad \begin{array}{ll}\text{Minimize} & J(v)\\ \text{Subject to} & v\in K \text{ and } |y(v;x)|\leqq 1 \quad \forall x\in\Omega.\end{array}$$

There are several papers dealing with control problems with state constraints. In these problems, there exists a fundamental difference between integral and pointwise state constraints. Lagrange multipliers in the optimality conditions are integrable functions in the first case and measures in the second case. The choice of the functional spaces is very important for proving the existence of a Lagrange multiplier associated with the constraint in the state, Bonnans and Casas [5], [6].

One of the first papers in distributed optimal control problems with state constraints was written by Mossino [15]. She studies a dual control problem, but she cannot prove the existence of a solution (which would be a Lagrange multiplier) because the choice of the functional spaces is not suited to the state constraint. See also [21].

The existence of this multiplier is interesting because it allows one to prove regularity results of the optimal control and state, which is useful in determining error estimates for approximating schemes. Barbu and Precupanu [3] and Lasiecka [11] derive the existence of a Lagrange multiplier for some control problems with integral state constraints.

Mackenroth in [14] considers a parabolic system controlled by Neumann conditions and subject to pointwise state constraints on the final state. He, like Mossino, studies a dual control problem, but now the functional spaces are well chosen, which allows one to prove the existence of a multiplier as a solution of this dual problem.

In this paper, the first objective is to prove the existence of a Lagrange multiplier which allows one to derive the optimality conditions. This is done by a direct and simple method that uses the well-known results of convex analysis. We do this without introducing any dual control problem. We prove that this multiplier is a Borel measure. This is done in § 2. In § 3, we study some properties of these measures.

In § 4, we give some regularity results for the optimal control and state, which is the second objective of this paper.

Smoothness properties of solutions to distributed control problems for systems governed by parabolic equations subject to control constraints have been proved by Lasiecka and Malanowski [12]. Lasiecka [11] derives some regularity properties of optimal solutions of distributed control problems with state constraints in integral form. The author is not aware of other regularity results for problems with pointwise state constraints governed by partial differential equations other than these proved in this paper and those proved in [5] for a different problem and using different techniques.

2. Existence of solution and optimality conditions. Let $C(\bar\Omega)$ be the space of real and continuous functions on $\bar\Omega$, endowed with the supremum-norm $\|\ \|_\infty$. It is known [1] that $H^2(\Omega)\subset C(\bar\Omega)$ $(1\leqq n\leqq 3)$, the inclusion being continuous.

We will denote by $C_0(\Omega)$ the subspace of $C(\bar\Omega)$ formed by the null functions on Γ. Then it is obvious that $H^2(\Omega)\cap H_0^1(\Omega)$ is included in $C_0(\Omega)$, the inclusion being continuous.

Let us denote the linear map $v\to y(v)$ by T, $T: L^2(\Omega)\to C_0(\Omega)$. An obvious consequence of (1.4) is that T is continuous.

Let B be the following subset:

$$B=\{z\in C_0(\Omega):\ \|z\|_\infty\leqq 1\}$$

and let I_B be the indicator of B

$$I_B(z)=\begin{cases}+\infty & \text{if } z\notin B,\\ 0 & \text{if } z\in B.\end{cases}$$

Then the problem (P) can be stated in the following way:

$$\inf_{v\in K}\{J(v)+(I_B\circ T)(v)\}. \tag{2.1}$$

Now we assume that the following hypothesis holds:

$$\text{Either } K \text{ is bounded in } L^2(\Omega) \text{ or } r>0. \tag{2.2}$$

THEOREM 1. *Suppose that there exists an element $v_0\in K$ such that $Tv_0\in B$. Then, under the hypothesis* (2.2), (P) *has a unique optimal solution.*

The proof is standard (Lions [13]), it is enough to note that $J+I_B\circ T$ is a lower semicontinuous convex functional and, from (2.2), the problem (P) is coercive. Note also that, even if $r=0$, the functional J is strictly convex and so we deduce the uniqueness of the solution.

To derive some optimality conditions, we will assume that the following Slater condition holds:

$$\exists v_0\in K \text{ such that } Tv_0\in \mathring{B}. \tag{2.3}$$

We will denote by $M(\Omega)$ the space of all real and regular Borel measures on Ω endowed with the norm

$$\|\mu\|_{M(\Omega)}=|\mu|(\Omega) \tag{2.4}$$

where $|\mu|$ is the total variation measure of μ.

According to the Riesz representation theorem (Rudin [19]), $M(\Omega)$ is the dual space of $C_0(\Omega)$ and the following equality holds:

$$|\mu|(\Omega)=\sup_{z\in B}\int_\Omega z\,d\mu. \tag{2.5}$$

Now we consider the space $H=H^2(\Omega)\cap H_0^1(\Omega)$ endowed with the norm $\|\ \|_{H^2(\Omega)}$. It is obvious that H is a Hilbert space. On the other hand, since the coefficients a_{ij} $(1\leqq i,j\leqq n)$ are Lipschitz functions, we see that $Az\in L^2(\Omega)$ for all $z\in H^2(\Omega)$ and so it follows from (1.4) that A is an isomorphism from H onto $L^2(\Omega)$.

As usual, we will denote by $\partial\psi$ the subdifferential of a convex function (Rockafellar [18], Ekeland and Temam [9]).

THEOREM 2. *Under the hypothesis* (2.3), *$u\in K$ is a solution of the problem* (P) *if and only if there exist $\mu\in M(\Omega)$, $p\in L^2(\Omega)$ and $y\in H$ such that*

$$Ay=u \quad \text{on } \Omega, \tag{2.6}$$

$$\int_\Omega pAz\,dx=\int_\Omega z(y-y_0)\,dx+\int_\Omega z\,d\mu \quad \forall z\in H, \tag{2.7}$$

$$\int_\Omega y\,d\mu=\sup_{z\in B}\int_\Omega z\,d\mu, \qquad y\in B, \tag{2.8}$$

$$\int_\Omega (p+ru)(v-u)\,dx\geqq 0 \quad \forall v\in K. \tag{2.9}$$

Proof. u is a solution of the problem (P) if and only if u minimizes in $L^2(\Omega)$ the functional

$$\psi(v) = J(v) + (I_B \circ T)(v) + I_K(v) \tag{2.10}$$

where I_K is the indicator of K.

But $\psi(u) = \mathrm{Inf}_{v \in L^2(\Omega)} \psi(v)$ if and only if $0 \in \partial\psi(u)$. By using the hypothesis (2.3) and the standard formulas for subdifferentials of convex functions (Rockafellar [18], Ekeland and Temam [9]) we obtain

$$0 \in J'(u) + T^* \circ \partial I_B(Tu) + \partial I_K(u) \tag{2.11}$$

which is equivalent to $u \in K$ and the existence of $\mu \in \partial I_B(Tu)$ such that

$$\begin{aligned} &\int_\Omega (y(u) - y_0)(y(v) - y(u))\, dx + r \int_\Omega u(v-u)\, dx \\ &\qquad + \int_\Omega T^*\mu(v-u)\, dx \geqq 0 \quad \forall v \in K, \end{aligned} \tag{2.12}$$

$$\mu \in \partial I_B(Tu). \tag{2.13}$$

Take $y = y(u)$, $p_1 = T^*\mu$ and let $p_2 \in H_0^1(\Omega)$ be the solution of the Dirichlet problem

$$\begin{aligned} &A^*p_2 = -\sum_{i,j=1}^{n} \partial_{x_j}(a_{ji}\partial_{x_i}p_2) + a_0 p_2 = y - y_0 \quad \text{on } \Omega, \\ &p_2 = 0 \quad \text{on } \Gamma. \end{aligned} \tag{2.14}$$

Finally, taking $p = p_1 + p_2$, we have

$$\begin{aligned} &\int_\Omega (y - y_0)(y(v) - y(u))\, dx + \int_\Omega T^*\mu(v-u)\, dx \\ &\quad = \int_\Omega A^*p_2(y(v) - y(u))\, dx + \int_\Omega p_1(v-u)\, dx \\ &\quad = \int_\Omega p_2(v-u)\, dx + \int_\Omega p_1(v-u)\, dx = \int_\Omega p(v-u)\, dx. \end{aligned} \tag{2.15}$$

From (2.12) and (2.15) we obtain (2.9), (2.8) follows from (2.13) and (2.6) is obvious. Now we prove (2.7). Let $z \in H$, then

$$\begin{aligned} \int_\Omega pAz\, dx &= \int_\Omega p_2 Az\, dx + \int_\Omega p_1 Az\, dx \\ &= \int_\Omega A^*p_2 z\, dx + \int_\Omega T^*\mu Az\, dx \\ &= \int_\Omega (y - y_0)z\, dx + \int_\Omega (T \circ A)z\, d\mu \\ &= \int_\Omega (y - y_0)z\, dx + \int_\Omega z\, d\mu. \end{aligned} \tag{2.16}$$

Q.E.D.

Remark. From (2.7) it follows that p satisfies the adjoint state equation

$$A^*p = y - y_0 + \mu \quad \text{on } \Omega. \tag{2.17}$$

We will prove in Theorem 4 that p has a null trace on Γ. The approximation by a finite element method of this equation has been studied by the author in [8].

3. Study of the multiplier μ. Next we prove some properties of the Lagrange multiplier μ. First recall that if the solution u of (P) satisfies $\|y(u)\|_\infty < 1$, then we deduce from (2.8) that $\mu = 0$. Now we examine the case $\|y(u)\|_\infty = 1$.

THEOREM 3. *Let u be the solution of* (P) and let $\mu \in M(\Omega)$ *satisfying* (2.8). *Take $y = Tu$ and consider the sets*

$$\Omega_+ = \{x \in \Omega: y(x) = +1\}, \qquad \Omega_- = \{x \in \Omega: y(x) = -1\}$$

and

$$\Omega_0 = \{x \in \Omega: |y(x)| < 1\}.$$

Let $\mu = \mu^+ - \mu^-$ be the Jordan decomposition of μ and $|\mu| = \mu^+ + \mu^-$ the total variation measure of μ. Then μ^+ and μ^- are concentrated in Ω_+ and Ω_-, respectively, that is to say

$$\mu^+(\Omega_- \cup \Omega_0) = \mu^-(\Omega_+ \cup \Omega_0) = |\mu|(\Omega_0) = 0. \tag{3.1}$$

Moreover

$$\int_\Omega (|y(x)| - 1)\, d|\mu|(x) = 0. \tag{3.2}$$

Proof. From (2.5), $\mu = \mu^+ - \mu^-$ and $|\mu| = \mu^+ + \mu^-$ we see that (2.8) is equivalent to

$$\int_\Omega y\, d\mu^+ - \int_\Omega y\, d\mu^- = |\mu|(\Omega) = \int_\Omega 1\, d\mu^+ + \int_\Omega 1\, d\mu^-. \tag{3.3}$$

This can be rewritten as

$$\int_\Omega (y-1)\, d\mu^+ - \int_\Omega (y+1)\, d\mu^- = 0. \tag{3.4}$$

From $-1 \leqq y \leqq +1$ we see that each of the integrals must vanish. The assertion (3.1) follows now from the definition of Ω_+, Ω_- and Ω_0.

Equation (3.2) is an obvious consequence of (3.1). Q.E.D.

Finally, we analyze a very frequent situation. As an immediate consequence of Theorem 3 we prove a fact that the numerical experimentation has clearly shown: the Lagrange multiplier is null at the points where the state constraint is not active and it is a Dirac measure when the constraint is active at a unique point, Casas [7] and Mossino [16].

COROLLARY 1. *Under the same hypothesis as in Theorem* 3 *and, furthermore, if the equality $|y(x)| = 1$ is satisfied at a finite set of points $\{x_j\}_{j=1}^m$, we have*

$$\mu = \sum_{j=1}^{m} \lambda_j \delta_{[x_j]} \tag{3.5}$$

where $\lambda_j \in R$ and $\delta_{[x_j]}$ is the Dirac measure concentrated at x_j. Moreover,

$$\lambda_j \geqq 0 \quad \text{if } y(x_j) = +1, \tag{3.6}$$

and

$$\lambda_j \leqq 0 \quad \text{if } y(x_j) = -1. \tag{3.7}$$

4. Regularity of the optimal solution. Next we investigate the regularity of the optimal solution u as well as the optimal state and Lagrange multiplier p.

If $s \in R$, $s \geqq 1$, and m is an integer, we will denote by $W^{m,s}(\Omega)$ and $W_0^{m,s}(\Omega)$ the usual Sobolev spaces (Adams [1]).

THEOREM 4. *Let $\mu \in M(\Omega)$ and $q \in L^2(\Omega)$ be such that*

$$\int_\Omega qAz\,dx = \int_\Omega z\,d\mu \quad \forall z \in H. \tag{4.1}$$

Then $q \in W_0^{1,s}(\Omega)$ for all $s \in [1, n/(n-1))$ and there exists $c_s > 0$ such that

$$\|q\|_{W_0^{1,s}(\Omega)} \leqq c_s \|\mu\|_{M(\Omega)}. \tag{4.2}$$

Proof. We first consider the one-dimensional case.

A) Case $n = 1$. In this case, it is well known that $W_0^{1,s}(\Omega) \subset C_0(\Omega)$ for all $s \in [1, \infty]$, the inclusion being continuous (Adams [1]); hence $M(\Omega) \subset W^{-1,s}(\Omega)$ for all $s \in (1, \infty)$. In particular, $\mu \in H^{-1}(\Omega)$ and so by density, it follows from (4.1) that

$$\langle q, Az\rangle_{H_0^1(\Omega)H^{-1}(\Omega)} = \int_\Omega z\,d\mu \quad \forall z \in H_0^1(\Omega). \tag{4.3}$$

Since A is an isomorphism from $H_0^1(\Omega)$ onto $H^{-1}(\Omega)$, we deduce the existence of $c_1 > 0$ such that

$$\begin{aligned} \langle q, Az\rangle_{H_0^1(\Omega)H^{-1}(\Omega)} &\leqq \|\mu\|_{M(\Omega)}\|z\|_\infty \\ &\leqq c_0\|\mu\|_{M(\Omega)}\|z\|_{H_0^1(\Omega)} \leqq c_1\|\mu\|_{M(\Omega)}\|Az\|_{H^{-1}(\Omega)} \end{aligned} \tag{4.4}$$

which proves that $q \in (H^{-1}(\Omega))' = H_0^1(\Omega)$ and

$$\|q\|_{H_0^1(\Omega)} \leqq c_1\|\mu\|_{M(\Omega)}. \tag{4.5}$$

On the other hand, since $n = 1$, the operator A can be written in the following way:

$$Az = -a_{11}\frac{d^2z}{dx^2} - \frac{da_{11}}{dx}\frac{dz}{dx} + a_0 z. \tag{4.6}$$

Finally, it is obvious that $A^*q = Aq = \mu$ on Ω, therefore

$$\frac{d^2q}{dx^2} = \frac{1}{a_{11}}\left(-\frac{da_{11}}{dx}\frac{dq}{dx} + a_0 q - \mu\right) \in W^{-1,s}(\Omega) \tag{4.7}$$

for all $s \in (1, +\infty)$.

So from (4.5) and (4.7) we obtain (4.2). Q.E.D.

Before proving the case $n > 1$, we will prove the following lemma.

LEMMA 1. *Let $f \in W^{-1,t}(\Omega)$ where $t > n$ is a real number, $n = 2$ or 3. Let $z \in H_0^1(\Omega)$ be the solution of the Dirichlet problem*

$$\begin{aligned} Az &= f \quad \text{on } \Omega, \\ z &= 0 \quad \text{on } \Gamma. \end{aligned} \tag{4.8}$$

Then $z \in C_0(\Omega)$ and there exists $c > 0$ such that

$$\|z\|_\infty \leqq c\|f\|_{W^{-1,t}(\Omega)}. \tag{4.9}$$

Proof. Let $f \in W^{-1,t}(\Omega)$ and let $\{f_j\}_{j=0}^n \subset L^t(\Omega)$ be such that (Adams [1, pp. 47–51])

$$f = f_0 + \sum_{j=1}^n \partial_{x_j} f_j \tag{4.10}$$

and

$$\|f\|_{W^{-1,t}(\Omega)} = \sum_{j=0}^{n} \|f_j\|_{L^t(\Omega)}. \tag{4.11}$$

Let $z_0 \in H_0^1(\Omega)$ be the solution of

$$\begin{aligned} Az_0 &= f_0 \quad \text{on } \Omega, \\ z_0 &= 0 \quad \text{on } \Gamma. \end{aligned} \tag{4.12}$$

Then, from (1.4), we deduce that $z_0 \in H^2(\Omega) \cap H_0^1(\Omega)$ and consequently

$$\|z_0\|_\infty \leqq c_1 \|z_0\|_{H^2(\Omega)} \leqq c_2 \|f_0\|_{L^2(\Omega)} \leqq c_3 \|f_0\|_{L^t(\Omega)}. \tag{4.13}$$

Let now $z_1 \in H_0^1(\Omega)$ be the solution of

$$\begin{aligned} Az_1 &= \sum_{j=1}^{n} \partial_{x_j} f_j \quad \text{on } \Omega, \\ z_1 &= 0 \quad \text{on } \Gamma. \end{aligned} \tag{4.14}$$

Since $t > n$, we deduce from (4.14) and the Lax-Milgram theorem (Nečas [17]):

$$\begin{aligned} \|z_1\|_{H_0^1(\Omega)} &\leqq c_4 \left\| \sum_{j=1}^{n} \partial_{x_j} f_j \right\|_{H^{-1}(\Omega)} \\ &\leqq c_5 \sum_{j=1}^{n} \|f_j\|_{L^2(\Omega)} \leqq c_6 \sum_{j=1}^{n} \|f_j\|_{L^t(\Omega)}. \end{aligned} \tag{4.15}$$

Furthermore, by using the theorem (1.4) of Nečas [17, p. 319], we obtain

$$\|z_1\|_\infty \leqq c_7 \left(\sum_{j=1}^{n} \|f_j\|_{L^t(\Omega)} + \|z_1\|_{H_0^1(\Omega)} \right). \tag{4.16}$$

Now, (4.9) follows from (4.13), (4.15) and (4.16).

In order to prove that $z \in C_0(\Omega)$ it is enough to consider a sequence $\{f_n\} \subset L^t(\Omega)$ such that $f_n \to f$ in $W^{-1,t}(\Omega)$. Now from (1.4) it follows that the solution z_n of

$$\begin{aligned} Az_n &= f_n \quad \text{on } \Omega, \\ z_n &= 0 \quad \text{on } \Gamma \end{aligned} \tag{4.17}$$

belongs to $H^2(\Omega) \cap H_0^1(\Omega) \subset C_0(\Omega)$ and using (4.9) we derive

$$\|z - z_n\|_\infty \leqq c_8 \|f - f_n\|_{W^{-1,t}(\Omega)} \to 0 \tag{4.18}$$

which proves that $z \in C_0(\Omega)$. Q.E.D.

B) Case $n > 1$. Take $s \in (1, n/(n-1))$ and let $t > n$ be such that $(1/s) + (1/t) = 1$. Given $\psi \in L^t(\Omega)$, let $z \in H$ be the solution of the Dirichlet problem

$$\begin{aligned} Az &= \psi \quad \text{on } \Omega, \\ z &= 0 \quad \text{on } \Gamma. \end{aligned} \tag{4.19}$$

From (1.4) and (4.9) we derive

$$\begin{aligned} \left| \int_\Omega q\psi \, dx \right| &= \left| \int_\Omega qAz \, dx \right| = \left| \int_\Omega z \, d\mu \right| \\ &\leqq \|\mu\|_{M(\Omega)} \|z\|_\infty \leqq c \|\mu\|_{M(\Omega)} \|\psi\|_{W^{-1,t}(\Omega)}. \end{aligned} \tag{4.20}$$

Since $L^t(\Omega)$ is dense in $W^{-1,t}(\Omega)$ and $W^{-1,t}(\Omega)$ is the dual space of $W_0^{1,s}(\Omega)$ it follows from (4.20) that $q \in W_0^{1,s}(\Omega)$ and q satisfies (4.2).

The case $s=1$ is a consequence of the continuity of the inclusion $W_0^{1,s'}(\Omega)\subset W_0^{1,1}(\Omega)$ for all $s'>1$. Q.E.D.

Stampacchia [20] proved Theorem 4 under the hypothesis: $a_0(x)\geqq\beta>0$ in Ω. Moreover, he supposed that Ω is a H_0^1-admissible subset which is not easy to verify.

THEOREM 5. *Let $\mu\in M(\Omega)$ and $q\in L^2(\Omega)$ be as described in the statement of Theorem 4 and let $E\subset\Omega$ be a closed set such that $|\mu|(\Omega\backslash E)=0$. Then $q\in H^2_{\text{loc}}(\Omega\backslash E)$ and hence is continuous on $\Omega\backslash E$.*

Proof. From (4.1) we derive

$$\int_{\Omega\backslash E} qA\psi\,dx=\int_{\Omega} qA\psi\,dx=\int_{\Omega}\psi\,d\mu=\int_{\Omega\backslash E}\psi\,d\mu=0 \tag{4.21}$$

for all $\psi\in D(\Omega\backslash E)$.

Thus we obtain that $q\in L^2(\Omega)$ is a local solution of $A^*q=0$ on $\Omega\backslash E$, then (Nečas [17]) it follows that $q\in H^2_{\text{loc}}(\Omega\backslash E)$. On the other hand $H^2_{\text{loc}}(\Omega\backslash E)\subset C(\Omega\backslash E)$ which completes the proof. Q.E.D.

As an immediate consequence of this theorem we obtain a regularity result for the Lagrange multiplier p.

COROLLARY 2. *Let $\mu\in M(\Omega)$ and $p\in L^2(\Omega)$ be the Lagrange multipliers obtained in Theorem 2. Then $p\in W_0^{1,s}(\Omega)$ for all $s\in[1, n/(n-1))$. If, furthermore, Ω_0 is the set defined in Theorem 3, we have $p\in H^2_{\text{loc}}(\Omega_0)$.*

Now in order to investigate the regularity of the optimal solution u of (P), we are going to suppose that r is strictly positive in (1.5).

COROLLARY 3. *Let K be one of the following sets:*

$$K_1=\{v\in L^2(\Omega): v(x)\geqq 0 \text{ a.e. in } \Omega\},$$
$$K_2=\{v\in L^2(\Omega): \|v\|_{L^2(\Omega)}\leqq 1\}.$$

Let u be the optimal solution of (P), then $u\in W_0^{1,s}(\Omega)$ for all $s\in[1, n/(n-1))$. Furthermore we have that $u\in H^1_{\text{loc}}(\Omega_0)\cap C(\Omega_0)$ if $K=K_1$ and u belongs to $H^2_{\text{loc}}(\Omega_0)$ if $K=K_2$.

Proof. If $K=K_1$, it follows easily from (2.9)

$$u(x)=-\frac{1}{r}\operatorname{Inf}\{0, p(x)\}\quad \text{a.e. in } \Omega. \tag{4.22}$$

If $K=K_2$, also from (2.9) we derive that

$$\begin{aligned} &u(x)=-\frac{p(x)}{\|p\|_{L^2(\Omega)}} && \text{if } \|p\|_{L^2(\Omega)}>r,\\ &u(x)=-\frac{1}{r}p(x) && \text{if } \|p\|_{L^2(\Omega)}\leqq r. \end{aligned} \tag{4.23}$$

The assertion follows now from (4.22), (4.23) and Corollary 2 (Lions [13]). Q.E.D.

COROLLARY 4. *Let ψ_1, $\psi_2\in H^1(\Omega)$ be such that*

$$K=\{v\in L^2(\Omega): \psi_1(x)\leqq v(x)\leqq\psi_2(x) \text{ a.e. in } \Omega\}$$

Then $u\in W^{1,s}(\Omega)\forall s\in[1, n/(n-1))$ and if, furthermore, the functions ψ_1, ψ_2 are continuous, we have that $u\in H^1_{\text{loc}}(\Omega_0)\cap C(\Omega_0)$.

Proof. From (2.9) it follows

$$u(x)=\max\left\{\psi_1(x), \min\left\{-\frac{1}{r}p(x), \psi_2(x)\right\}\right\}\quad \text{a.e.} \tag{4.24}$$

Now the corollary is a consequence of Corollary 2 (Lions [13]). Q.E.D.

Remark. The formulas (4.22), (4.23) and (4.24) prove that the regularity of u depends essentially on the regularity of p. In general, $p \notin H^1(\Omega)$ and consequently the result $u \in W^{1,s}(\Omega)$ for all $s \in [1, n/(n-1))$ is optimal. Consider, for example, a control problem where the state constraint is active at a unique point $x_0 \in \Omega$, then the Lagrange multiplier μ is a Dirac measure (Corollary 1) and consequently $p \notin H^1(\Omega)$.

When u is given by (4.23) and the coefficients a_{ij} and a_0 of A are C^∞ functions on Ω, then it is easy to prove that p and u are C^∞ functions on Ω_0.

Finally, in order to study the regularity of the optimal state, we assume that Γ is a manifold of class C^∞ and a_0, a_{ij} $(1 \leqq i, j \leqq n)$ are functions of class C^∞ (it is possible to weaken these regularity hypotheses, see [2], [17]).

COROLLARY 5. *Let K be defined as in one of the Corollaries* 3 *or* 4. *Then the optimal state y belongs to the Sobolev space $W^{3,s}(\Omega) \cap H_0^1(\Omega)$ for all $s \in [1, n/(n-1))$ and moreover $y \in H^3_{\text{loc}}(\Omega_0)$. If furthermore $n \leqq 2$, then $y \in C^1(\bar{\Omega})$.*

Proof. Since $u \in W^{1,s}(\Omega)$, it follows from [2] that y belongs to $W^{3,s}(\Omega) \cap H_0^1(\Omega)$. On the other hand, $u \in H^1_{\text{loc}}(\Omega_0)$ implies (see [17]) that $y \in H^3_{\text{loc}}(\Omega_0)$.

Finally, if $n \leqq 2$ and $(n/(n-1)) - s$ is small enough, we have $W^{3,s}(\Omega) \subset C^1(\bar{\Omega})$, from where $y \in C^1(\bar{\Omega})$. Q.E.D.

Acknowledgment. The author wishes to thank the referees for several important suggestions on this paper. In particular, the proof of Theorem 3 presented in this paper was suggested by one of them.

REFERENCES

[1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] S. AGMON, A. DOUGLIS AND L. NIRENBERG, *Estimates near the boundary for solutions for elliptic partial differential equations satisfying general boundary conditions* I., Comm. Pure Appl. Math., 12 (1959), pp. 623-727.

[3] V. BARBU AND TH. PRECUPANU, *Convexity and optimization in Banach spaces*, Sijthoff & Noordhoff-Publishing House of Romanian Academy, 1978.

[4] J. F. BONNANS AND E. CASAS, *Contrôle de systèmes non lineaires comportant des contraintes distribuées sur l'état*, Rapport No. 300, INRIA, 1984.

[5] ———, *On the choice of the function spaces for some state-constrained control problems*, J. Numer. Funct. Anal. Optimiz, to appear.

[6] ———, *Quelques méthodes pour le contrôle optimal de problèmes comportant des contraintes sur l'état*, 4th Workshop on Differential Equations and Control Theory, 9-13 Sept. 1984, Iaši, Romania.

[7] E. CASAS, *Análisis numérico de algunos problemas de optimización estructural*, Ph.D. dissertation, Univ. Santiago de Compostela (Spain), 1982.

[8] ———, *L^2 Estimates for the finite element method for the Dirichlet problem with singular data*, Numer. Math, to appear.

[9] I. EKELAND AND R. TEMAM, *Analyse convexe et problèmes variationnels*, Dunod, Gauthier-Villars, Paris, 1984.

[10] P. GRISVARD, *Behavior of the solutions of an elliptic boundary value problem in a polygonal or polyhedral domain*, in Numerical Solution of Partial Differential Equations III (SYNSPADE 1975), B. Hubbard, ed., Academic Press, New York, 1976, pp. 207-274.

[11] I. LASIECKA, *State constrained control problems for parabolic systems: regularity of optimal solutions*, Appl. Math. Optim., 6 (1980), pp. 1-29.

[12] I. LASIECKA AND K. MALANOWSKI, *On the regularity of solutions to convex optimal control problems with control constraints for parabolic systems*, Control Cybernet., 6 (1977), pp. 57-74.

[13] J. L. LIONS, *Optimal control of systems governed by partial differential equations*, Springer-Verlag, Berlin, 1971.

[14] U. MACKENROTH, *Convex parabolic boundary control problems with pointwise state constraints*, J. Math. Anal. Appl., 87 (1982), pp. 256-277.

[15] J. MOSSINO, *An application of duality to distributed optimal control problems with constraints on the control and the state*, J. Math. Anal. Appl., 50 (1975), pp. 223-242.

[16] J. MOSSINO, *Approximation numérique de problèmes de contrôle optimal avec contraintes sur le contrôle et sur l'état*, Calcolo, 13 (1976), pp. 21-62.
[17] J. NEČAS, *Les méthodes directes en théorie des équations elliptiques*, Masson, Paris, 1967.
[18] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[19] W. RUDIN, *Real and Complex Analysis*, McGraw-Hill, New York, 1966.
[20] G. STAMPACCHIA, *Le problème de Dirichlet pour les équations elliptiques de second ordre à coefficients discontinus*, Ann. Ins. Fourier Grenoble, 15 (1965), pp. 189-258.
[21] L. W. WHITE, *Control of a hyperbolic problem with pointwise stress constraints*, J. Optim. Theory Appl., 41 (1983), pp. 359-369.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 6, November 1986

014

EVEN MORE STATES REACHABLE BY BOUNDARY CONTROL FOR THE HEAT EQUATION*

E. J. P. GEORG SCHMIDT†

Abstract. In this note we show how both known and new sufficient conditions for reachability in boundary control for the heat equation can easily be derived from null controllability. A characterisation of reachable states remains elusive.

Key words. distributed control, heat equation, reachability

Let Ω be a bounded domain in R^n whose boundary $\partial\Omega$ is a C^∞ manifold. Let Δ denote the Laplacian operator on R^n, $\partial/\partial\nu$ denote differentiation in the direction of the outward pointing normal ν to $\partial\Omega$, a be a nonnegative constant, and $\mathcal{B}$"=" $a(\partial/\partial\nu)+1$. We consider the following initial boundary value problem:

$$\begin{aligned} &\frac{\partial u}{\partial t}(x,t)=\Delta u(x,t) \quad \text{for } x\in\Omega, \qquad t\in(0,T], \\ &\mathcal{B}u(x,t)=f(x,t) \quad \text{for } x\in\partial\Omega, \qquad t\in(0,T], \\ &u(x,0)=u_0(x) \quad \text{for } x\in\Omega. \end{aligned} \tag{1}$$

It can be shown that, given u_0 in $\mathcal{H}=L_2(\Omega)$ and f in $\mathcal{U}=L_\infty(\partial\Omega\times(0,T))$ (1) has a unique weak solution $u(x,t)$ in $L_2(\Omega\times(0,T))\cap C^\infty(\Omega\times(0,T))$ in the following sense (see [4] or [8]):

$$\begin{aligned} \int_\Omega\int_0^T u(x,t)\left[\frac{\partial\phi}{\partial t}(x,t)+\Delta\phi(x,t)\right]dx\,dt+\int_\Omega u_0(x)\phi(x,0)\,dx \\ +\int_{\partial\Omega}\int_0^T f(x,t)\phi^\partial(x,t)\,dS_x\,dt=0, \end{aligned} \tag{2}$$

where dS_x denotes an element of surface area of $\partial\Omega$, ϕ belongs to the space of test functions

$$\mathcal{T}=\{\phi\in C^\infty(\bar{\Omega}\times[0,T]):\ \phi(\cdot,T)\equiv 0,\ \mathcal{B}\phi(\cdot,t)\equiv 0\}$$

and

$$\phi^\partial(x,t)=\begin{cases} a^{-1}\phi(x,t) & \text{for } x\in\partial\Omega,\ t>0 \quad \text{if } a>0, \\ -\dfrac{\partial\phi}{\partial\nu}(x,t) & \text{for } x\in\partial\Omega,\ t>0 \quad \text{if } a=0. \end{cases} \tag{3}$$

Moreover $u(\cdot,T)$ lies in $\mathcal{H}$, so that one can define the set of states reachable from u_0 in time T by

$$\mathcal{R}_T(u_0,\mathcal{U})=\{u(\cdot,T):\ \text{there exists } f\in\mathcal{U} \text{ with } u(x,t) \text{ the corresponding solution of (1)}\}.$$

* Received by the editors June 12, 1985 and in revised form October 7 1985. This work was supported by the Natural Sciences and Engineering Research Council of Canada under grant A7271.

† Department of Mathematics and Statistics, McGill University, Montreal, Quebec, Canada H3A 2K6.

It is well known that $\mathcal{R}_T(u_0, \mathcal{U})$ is dense in $\mathcal{H}$ (see, for example [5]), that $0 \in \mathcal{R}_T(u_0, \mathcal{U})$ (a fundamental and deep result known as "null controllability" and proved in [6] or [9]) and that $\mathcal{R}_T(u_0, \mathcal{U})$ is independent of u_0 and T (see [2] or [10]). From now on we denote $\mathcal{R}_T(u_0, \mathcal{U})$ by $\mathcal{R}$; it forms a nonclosed subspace of $\mathcal{H}$. We refer to the functions in that subspace as "reachable." No intrinsic characterisation of reachable functions has yet been given. However there exist various sufficient conditions for reachability, which allow one to identify some interesting classes of reachable functions (see, for example, [6], [8], [7], [11], [12]). In this note a new sufficient condition is given which subsumes most of the known conditions.

We remark that the space $\mathcal{U}$ of controls can be chosen in other ways without affecting the validity of all the previously stated facts. If, for example, $a \neq 0$ one can choose $\mathcal{U} = L_2(\partial\Omega \times (0, T))$ while, if $a = 0$ (the case of Dirichlet boundary conditions) $\mathcal{U}$ can be chosen as in [11].

Given w in $\mathcal{H}$ with Δw in $\mathcal{H}$ (in the distribution sense) we say that $\mathcal{B}w = f$ (with f in $L_\infty(\partial\Omega)$) if

$$\int_\Omega \Delta w(x)\phi(x)\,dx = \int_\Omega w(x)\Delta\phi(x)\,dx + \int_{\partial\Omega} f(x)\phi^\partial(x)\,dS_x, \tag{4}$$

for any ϕ in $C^\infty(\bar{\Omega})$ satisfying $\mathcal{B}\phi = 0$.

Let $\|\cdot\|_{\mathcal{H}}$ and $\|\cdot\|_{\mathcal{U}}$ denote the norms on $\mathcal{H}$ and $\mathcal{U}$, respectively. Since $L_\infty(\partial\Omega)$ can be regarded as a subspace of $\mathcal{U}$ (consisting of functions independent of t), $\|\cdot\|_{\mathcal{U}}$ can be used on $L_\infty(\partial\Omega)$, and in fact coincides with the usual norm.

We can now state our main theorem.

THEOREM. *Let v in $\mathcal{H}$ satisfy*

(i) $\Delta^n v \in \mathcal{H}$, *for all positive integers n and for some $p > 0$ the series*

$$\sum_{n=0}^{\infty} \frac{(-p)^n}{n!} \Delta^n v(\cdot)$$

converges in $\mathcal{H}$.

(ii) $\mathcal{B}(\Delta^n v) \in L_\infty(\partial\Omega)$, *for all positive integers n, and furthermore the series*

$$\sum_{n=0}^{\infty} \frac{(-p)^n}{n!} \mathcal{B}(\Delta^n v)$$

converges in $L_\infty(\partial\Omega)$. Then v is reachable.

Proof. Pick T in $(0, p)$ and note that the series

$$\sum_{n=0}^{\infty} \frac{(t-T)^n}{n!} \Delta^n v \quad \text{and} \quad \sum_{n=0}^{\infty} \frac{(t-T)^n}{n!} \mathcal{B}(\Delta^n v)$$

converge uniformly and "absolutely" (with the appropriate norm replacing absolute values) to yield functions $u(x, t)$ in $L_2(\Omega \times (0, T))$ and $f(x, t)$ in $\mathcal{U}$ respectively. Note that $u(x, T) = v(x)$. Let $u_N(x, t)$ and $f_N(x, t)$ denote the partial sums of the two series (with n running from 0 to N); then $(\partial/\partial t)u_N(x, t) = \Delta u_{N-1}(x, t)$. Now integration by parts and (4) readily yield, for test functions ϕ in $\mathcal{T}$,

$$\int_\Omega \int_0^T \left[u_N(x, t) \frac{\partial\phi}{\partial t}(x, t) + u_{N-1}(x, t)\Delta\phi(x, t) \right] dx\,dt + \int_\Omega u_N(x, 0)\phi(x, 0)\,dx$$
$$+ \int_{\partial\Omega} \int_0^T f_{N-1}(x, t)\phi^\partial(x, t)\,dS_x\,dt = 0.$$

Letting N tend to infinity one obtains (2), so that one can conclude that $v(x) = u(x, T)$ belongs to $\mathcal{R}(u(x, 0), \mathcal{U}) = \mathcal{R}$.

This theorem has a number of easy corollaries.

COROLLARY 1. *The following classes of functions are included in* $\mathcal{R}$:

(a) *Polynomials in the space variables* (*see also* [7] *and* [11]).

(b) *Functions* v *which satisfy* $\Delta v = \lambda v$ *with* $\mathcal{B}v$ *in* $L_\infty(\partial\Omega)$, *a class of functions which includes "holdable targets"* (*see* [8]), *the eigenfunctions of* Δ *corresponding to homogeneous boundary conditions, the restrictions to* Ω *of such eigenfunctions on bigger domains and functions of the form* $\prod_{i=1}^{n} \psi_i(x_i)$ *with* $\psi_i(x_i) = \cos(a_i x_i)$, $\sin(a_i x_i)$ *or* $\exp(a_i x_i)$ (*see* [7, Cor. 6]).

(c) *The restrictions to* Ω *of functions* v *in* $L_2(R^n)$ *whose Fourier transforms* $\hat{v}$ *satisfy* $\exp(p\xi^2)\hat{v}(\xi) \in L_2(R^n)$ *for some* $p > 0$.

The proofs of (a) and (b) are trivial; (c) requires Plancherel's theorem and the observation that, when p is replaced by a slightly smaller positive constant, the series occurring in hypotheses (i) and (ii) of the theorem, written in terms of Fourier transforms, converge uniformly in x.

The next corollary partially generalises a result to be found in [6]. To formulate this one needs to introduce the selfadjoint operator $\mathcal{L}$ defined as Δ acting on the appropriate domain of functions ϕ in $L_2(\Omega)$ satisfying $\mathcal{B}\phi = 0$. Then (see [1]) $\mathcal{L}$ has a complete orthonormal system of eigenfunctions $\{\phi_k\}_{k=1}^{\infty}$ corresponding to negative eigenvalues $\{-\lambda_k\}_{k=1}^{\infty}$. Given v in $L_2(\Omega)$ let v_k denote the coefficient of v with respect to ϕ_k.

COROLLARY 2. *Let* v *in* $L_2(\Omega)$ *satisfy* $\sum_{k=1}^{\infty} \exp(2p\lambda_k) v_k^2 < \infty$ *for some* $p > 0$; *then* v *is reachable.*

To prove this we note that $\Delta^n v = \sum_{k=1}^{\infty} (-\lambda_k)^n v_k \phi_k$ belongs to the domain of $\mathcal{L}$ for all n and hence $\mathcal{B}(\Delta^n v) = 0$, so that hypothesis (ii) of our theorem is satisfied for any choice of p. Moreover (i) is satisfied since one easily justifies

$$\sum_{n=1}^{\infty} \frac{(-p)^n}{n!} \Delta^n v = \sum_{k=1}^{\infty} \exp(p\lambda_k) v_k \phi_k,$$

which is convergent in $L_2(\Omega)$ by hypothesis.

Without describing the details we remark that Theorems 5 and 5^0 of [11] can also be obtained immediately from our theorem (adapted to handle the space of controls used in the cited paper).

It is also possible to give a generalisation of a result in [7] which is not contained in the previous theorem.

PROPOSITION. *Let* w *in* $L_2(\Omega)$ *be a solution of*

$$(\Delta - \lambda) w = v \quad \text{and} \quad \mathcal{B}w = g,$$

where v *is in* $\mathcal{R}$ *and* g *is in* $L_\infty(\partial\Omega)$. *Then* w *is reachable.*

When $v = 0$ this is just Corollary 1(b); by linearity it is then enough to prove the result for $g = 0$. This involves minor changes in the proof for the case $\lambda = 0$ (given in [7, Thm. 3]).

An easy induction proof on the degree of the polynomial then yields

COROLLARY 3. *The functions* $p(x)v(x)$, *where* p *is a polynomial and* v *is as described in Corollary* 1(b), *are reachable.*

REFERENCES

[1] S. AGMON, *Lectures on Elliptic Boundary Value Problems*, Van Nostrand, New York, 1965.

[2] H. O. FATTORINI, *Reachable states in boundary control of the heat equation are independent of time*, Proc. Royal Soc. Edinburgh 81, (1976), pp. 71-77.

[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, *The time optimal problem for boundary control of the heat equation*, Calculus of Variations and Control Theory, D. L. Russell, ed., Academic Press, New York, 1976, pp. 305-320.

[5] R. C. MACCAMY, V. J. MIZEL AND T. I. SEIDMAN, *Approximate boundary controllability for the heat equation*, J. Math. Anal. Appl. 23, (1968), pp. 699-703.

[6] D. L. RUSSELL, *A unified boundary controllability theory for hyperbolic and parabolic partial differential equations*, Studies in Appl. Math. LII (1973), pp. 189-211.

[7] E. SACHS AND E. J. P. G. SCHMIDT, *On reachable states in boundary control for the heat equation and an associated moment problem*, Appl. Math. Optim. 7, (1981), pp. 225-232.

[8] G. SCHMIDT, *Boundary control for the heat equation with steady state targets*, this Journal, 18 (1980), pp. 145-154.

[9] T. I. SEIDMAN, *A well-posed problem for the heat equation*, Bull. AMS, 80 (1974), pp. 901-902.

[10] ———, *Time-invariance of the reachable set for linear control problems*, J. Math. Anal. Appl. 72 (1979), pp. 17-20.

[11] N. WECK, *More states reachable by boundary control of the heat equation*, this Journal, 22 (1984), pp. 699-710.

[12] ———, *On exact controllability for parabolic equations*, in Optimal Control of Partial Differential Equations, K.-H. Hoffmann and W. Krabs, eds., Birkhauser, Boston, 1984, pp. 243-261.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 24, No. 6, November 1986

015

ON HERMITIAN SOLUTIONS OF THE SYMMETRIC ALGEBRAIC RICCATI EQUATION*

I. GOHBERG†, P. LANCASTER‡ AND L. RODMAN§

Abstract. The structure of the set of hermitian solutions of the matrix quadratic equation $XDX - XA - A^*X - C = 0$ is studied under the conditions that $C = C^*$, D is positive semidefinite and (A, D) is stabilizable. New features (e.g., nonexistence of the minimal solution) appear in contrast with the known case when (A, D) is controllable.

Key words. algebraic Riccati equation, hermitian solutions, extremal solutions, stabilizability, invariant subspaces

AMS(MOS) subject classifications. 93B25, 49E30, 15A24

1. Introduction. Our primary concern is the analysis of hermitian solutions X of the Riccati equation of the form

$$XDX - XA - A^*X - C = 0. \tag{1}$$

This equation will be considered under the assumptions that A, C, D are $n \times n$ matrices with complex entries with C hermitian, D positive semidefinite, and the pair (A, D) stabilizable.

The now classical problem of quadratic optimization of a time-invariant finite-dimensional linear problem can be reduced to the solution of (1). Namely, the control $u(t)$ of the time invariant system

$$\dot{x} = Ax + Bu \tag{2}$$

which minimizes the cost functional

$$\int_0^\infty (x^*Cx + u^*Ru)\, dt \tag{3}$$

with constant positive definite matrix R and positive semidefinite C, is given by the formula

$$u(t) = -R^{-1}B^*X_+x(t),$$

where X_+ is the maximal hermitian solution (which is assumed to exist) of the matrix equation

$$XBR^{-1}B^*X - XA - A^*X - C = 0. \tag{4}$$

The maximality of X_+ means that $X_+ - X$ is positive semidefinite for any hermitian solution X of (4).

It is important to know when the solutions of (2) are stable, and it is well known that this happens if and only if all the eigenvalues of A are in the open left halfplane. More generally, we ask when the stability of (2) can be achieved by feedback. This

* Received by the editors October 30, 1984, and in revised form September 20, 1985. This research was partially supported by the Natural Sciences and Engineering Research Council of Canada.

† School of Mathematical Sciences, Tel-Aviv University, Tel-Aviv, Ramat Aviv, Israel.

‡ Department of Mathematics and Statistics, University of Calgary, Alberta, Canada T2N 1N4.

§ School of Mathematical Sciences, Tel-Aviv University, Tel-Aviv, Ramat Aviv, Israel. The research of this author was partially supported by the Fund for Basic Research administrated by the Israel Academy of Sciences and Humanities.

means that after a transformation $u = Fx + v$ with some matrix F (where v is the new control) the equation

$$\dot{x} = (A + BF)x + Bv$$

is stable. In other words, the eigenvalues of $A + BF$ are in the open left halfplane. If such an F exists, the pair (A, B) is called *stabilizable.*

It is possible to check that (A, B) is stabilizable if and only if $(A, BR^{-1}B^*)$ is stabilizable. So a natural additional restriction on the coefficients of (1) is that (A, D) be stabilizable. Indeed, this is a useful sufficient condition for the problem of minimizing the cost functional to be well posed.

In particular, (A, D) is stabilizable if this pair is *controllable*, which means that its controllability subspace $C_{A,D} = \sum_{j=1}^{n} \operatorname{Im}(A^{j-1}D)$ coincides with $\mathbb{C}^n$. For a theory of equation (1) under the controllability of (A, D) see [6], [15] and expositions in [2], [3], [12] (many references concerning the development of this theory are found in [12]). Our purpose in this paper is to extend this algebraic theory to the case when only stabilizability of (A, D) is required. In the course of this extension new features appear: First, in contrast with the controllability case, the minimal hermitian solution of (1) does not always exist. However, the reader's attention is drawn to the sufficient condition (called *regularity*) under which a minimal solution exists given in Corollary 4.4, and to two other characterizations of equations for which minimal solutions exist (Corollary 5.2).

Second, in the geometric description of the hermitian solutions there is a distinction between two cases. In one case (the *regular* case) this description goes in the same way as when (A, D) is controllable (see [6], [15]). In the nonregular case the situation is more complicated and only partial results are obtained.

The paper consists of 6 sections. In the second section we study the existence question and properties of the maximal hermitian solution and also give an example of nonexistence of the minimal hermitian solution. Existence of hermitian solutions of (1) is studied in § 3 in the geometric terms of invariant subspaces associated with

$$\begin{bmatrix} -A & D \\ C & A^* \end{bmatrix}.$$

The regular and nonregular cases are studied in §§ 4 and 5, respectively. In § 6 we discuss the Riccati equation with real coefficients and the properties of real symmetric solutions.

2. Existence and properties of extremal solutions. Consider the algebraic Riccati equation

$$XDX - XA - A^*X - C = 0, \tag{1}$$

where A, D and C are $n \times n$ (complex) matrices, D is positive semidefinite, C is hermitian and the pair (A, D) is stabilizable, and hermitian solutions X of (1) are required. These conditions on A, D, C will be maintained throughout the paper. Our first result asserts the existence of the maximal hermitian solution. We say that a hermitian solution X_+ of (1) is *maximal* if $X_+ \geqq X$ for any other hermitian solution of (1). Here and elsewhere $X \geqq Y$ for hermitian matrices X and Y means that $X - Y$ is positive semidefinite. Clearly, a maximal hermitian solution is unique if it exists.

THEOREM 2.1. *If* (1) *has a hermitian solution, then it has the maximal hermitian solution* X_+. *The matrix* X_+ *has the property that all eigenvalues of* $A - DX_+$ *are in the closed left halfplane.*

The case when $C \geqq 0$ arises naturally in quadratic optimal control problems (see the Introduction). For this case we have the following.

THEOREM 2.2. *If $C \geqq 0$, then* (1) *has a hermitian solution, and its maximal hermitian solution X_+* (*which exists by Theorem* 2.1) *is positive semidefinite.*

For completeness, we remark that hypotheses on the pair (A^*, C) arise naturally in the quadratic optimal control problem and give stronger conclusions. It is shown in [19], for example, that if, in addition to the hypotheses of Theorem 2.2, (A^*, C) is stabilizable then the matrix $A - DX_+$ is stable (i.e. has all its eigenvalues in the open left half-plane), and if (A^*, C) is controllable, then X_+ is positive definite.

We shall prove Theorem 2.1 together with the following result concerning existence of hermitian solutions if the hermitian matrix C is increased.

THEOREM 2.3. *Assume* (1) *has a hermitian solution. Then for every hermitian matrix C' satisfying $C' \geqq C$ the equation*

$$XDX - XA - A^*X - C' = 0 \tag{2}$$

has a hermitian solution as well. Moreover, the maximal hermitian solutions X_+ and X'_+ of (1) *and* (2), *respectively* (*which exist by Theorem* 2.1), *satisfy the inequality $X'_+ \geqq X_+$.*

Theorem 2.2 is a simple corollary of Theorem 2.3. Indeed, equation $XDX - XA - A^*X = 0$ obviously has the hermitian solution $X = 0$. Now use Theorem 2.3 to obtain Theorem 2.2.

In case (A, D) is controllable the result of Theorem 2.1 is known and goes back to [13] (see also [2]).

Theorem 2.2 seems to be new although closely related results with stronger hypotheses on the pair (A, D) have been known for some time (see [4], [5], [1], [9]). Similarly, Theorem 2.3 seems to be new, although a related result (without the existence statement) appears in the work of Willems [16]. Also, using other methods, and during the preparation of this paper, A.C.M. Ran has proved Theorem 2.3 in the case that (A, D) is controllable.

Theorems 2.3 and 2.1 will be proved together. In the proof we use an approach introduced (in a special case) in [5] and further developed in [18], [2].

It is well known that under the stronger condition of controllability of (A, D) equation (1) has the *minimal* hermitian solution X_- which is defined by the property that $X \geqq X_-$ for any hermitian solution X of (1), provided (1) has hermitian solutions at all (see e.g., [2]). This fact follows easily from Theorem 2.1. Indeed, if (A, D) is controllable, so is $(-A, D)$. In particular, $(-A, D)$ is stabilizable. Apply Theorem 2.1 to the equation

$$XDX + XA + A^*X - C = 0, \tag{3}$$

and observe that X satisfies (3) if and only if $-X$ satisfies (1). The following example shows that, in contrast with the controllable case, the minimal hermitian solution need not exist under the conditions of Theorem 2.1:

Example 2.1. Let

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

A calculation shows that the hermitian solutions of (1) are

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad \begin{bmatrix} -1 & b \\ \bar{b} & -|b|^2/2 \end{bmatrix}, \quad b \in \mathbb{C}.$$

Clearly,

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

is the maximal solution, but there is no minimal solution.

Proof of Theorems 2.1 *and* 2.3. Let

$$\mathrm{C} = \mathrm{Im}\,[D, AD, \cdots, A^{n-1}D]$$

be the controllable subspace of (A, D). Clearly, C is A-invariant and contains Im D. Hence with respect to the orthogonal decomposition $\mathbb{C}^n = C \oplus C^{\perp}$ we have

$$A = \begin{bmatrix} A_1 & A_{12} \\ 0 & A_2 \end{bmatrix}, \qquad D = \begin{bmatrix} D_1 & 0 \\ 0 & 0 \end{bmatrix}, \tag{4}$$

where D_1 is positive semidefinite and (A_1, D_1) controllable. Further, the stabilizability of (A, D) implies that A_2 is stable. As (A_1, D_1) is controllable, there is a hermitian matrix Y such that $A_1 - D_1 Y$ is stable. Indeed,

$$Y = \left(\int_0^1 e_0^{-A_1 t} D_1 e^{-A_1^* t} \, dt \right)^{-1}$$

will do (see [8] and also [14, Thm. III.2.1]).

Now for the hermitian matrix

$$X_0 = \begin{bmatrix} Y & 0 \\ 0 & 0 \end{bmatrix}$$

we have that $A - DX_0$ is stable.

Starting[1] with X_0, we shall define a sequence of hermitian matrices $\{X_\nu\}_{\nu=0}^{\infty}$ satisfying the equalities

$$X_{\nu+1}(A - DX_\nu) + (A - DX_\nu)^* X_{\nu+1} = -X_\nu D X_\nu - C', \qquad \nu = 0, 1, \cdots, \tag{5}$$

where $C' \geqq C$ is a fixed hermitian matrix. The sequence will also have the property that $A - DX_\nu$ is stable for all ν. Assuming inductively that we have defined $X_\nu = X_\nu^*$ already with $A - DX_\nu$ stable, (5) has a unique solution $X_{\nu+1}$. Taking adjoints in (5) and using the uniqueness of the solution, we obtain $X_{\nu+1} = X_{\nu+1}^*$. We are to show that $A - DX_{\nu+1}$ is stable. To this end note the following identity which holds for any hermitian matrices Y and $\hat{Y}$:

$$\begin{aligned} Y(A - DY) + (A - DY)^* Y + YDY \\ = Y(A - D\hat{Y}) + (A - D\hat{Y})^* Y + \hat{Y}D\hat{Y} - (Y - \hat{Y})D(Y - \hat{Y}). \end{aligned} \tag{6}$$

By assumption, there exists a hermitian solution X of (1). Letting $Y = X$ and $\hat{Y} = X_\nu$ in (6), we get

$$X(A - DX_\nu) + (A - DX_\nu)^* X + X_\nu D X_\nu - (X - X_\nu)D(X - X_\nu) = -C. \tag{7}$$

Subtract (7) from (5):

$$(X_{\nu+1} - X)(A - DX_\nu) + (A - DX_\nu)^*(X_{\nu+1} - X) = -(X - X_\nu)D(X - X_\nu) - (C' - C).$$

[1] In this part of the proof the line of argument follows that of the second proof of Theorem 2.1 in [2].

As $A-DX_\nu$ is stable it follows from Proposition 8.13.1 of [7], for example, that

$$X_{\nu+1}-X=\int_0^\infty e^{(A-DX_\nu)t}[(X-X_\nu)D(X-X_\nu)+C'-C]\,e^{(A-DX_\nu)^*t}\,dt.$$

In particular, $X_{\nu+1}\geqq X$.

Next, use (6) again with $Y=X_{\nu+1}$, $Y=X_\nu$ and apply (5) to get

$$\begin{aligned}(8)\qquad &X_{\nu+1}(A-DX_{\nu+1})+(A-DX_{\nu+1})^*X_{\nu+1}+X_{\nu+1}DX_{\nu+1}\\ &\quad=-C'-(X_{\nu+1}-X_\nu)D(X_{\nu+1}-X_\nu).\end{aligned}$$

Subtracting the equality (7) with X_ν replaced by $X_{\nu+1}$, we obtain

$$\begin{aligned}(9)\qquad &(X_{\nu+1}-X)(A-DX_{\nu+1})+(A-DX_{\nu+1})^*(X_{\nu+1}-X)\\ &\quad=-(X_{\nu+1}-X_\nu)D(X_{\nu+1}-X_\nu)-(X_{\nu+1}-X)D(X_{\nu+1}-X)-C'+C.\end{aligned}$$

Assume now $(A-DX_{\nu+1})x=\lambda x$ for some λ with $\operatorname{Re}\lambda\geqq 0$ and vector $x\neq 0$. Then

$$(10)\qquad (\bar\lambda+\lambda)x^*(X_{\nu+1}-X)x=x^*Wx,$$

where $W\leqq 0$ is the right-hand side of (9). As $X_{\nu+1}-X\geqq 0$, the equality (10) implies $x^*Wx=0$ which, using the definition of W, in turn implies

$$x^*(X_{\nu+1}-X_\nu)D(X_{\nu+1}-X_\nu)x=0.$$

But $D\geqq 0$, so $D(X_{\nu+1}-X_\nu)x=0$. Now

$$(A-DX_\nu)x=(A-DX_{\nu+1})x=\lambda x,$$

a contradiction with the stability of $A-DX_\nu$. Hence $A-DX_{\nu+1}$ is stable as well.

Next it will be shown that the sequence $\{X_\nu\}_{\nu=0}^\infty$ is nonincreasing. Consider the equality (8) with ν replaced by $\nu-1$, and subtract from it the equality (5) to get

$$(X_\nu-X_{\nu+1})(A-DX_\nu)+(A-DX_\nu)^*(X_\nu-X_{\nu+1})=-(X_\nu-X_{\nu-1})D(X_\nu-X_{\nu-1}).$$

As $A-DX_\nu$ is stable,

$$X_\nu-X_{\nu+1}=\int_0^\infty e^{(A-DX_\nu)t}(X_\nu-X_{\nu-1})D(X_\nu-X_{\nu-1})\,e^{(A-DX_\nu)^*t}\,dt\geqq 0.$$

So $\{X_\nu\}_{\nu=0}^\infty$ is a nonincreasing sequence of hermitian matrices bounded below by X. Hence the limit $X'_+=\lim_{\nu\to\infty}X_\nu$ exists. Passing to the limit in (5) when $\nu\to\infty$ shows that X'_+ is a hermitian solution of (2). Since $A-DX_\nu$ is stable for all $\nu=0,1,\cdots$, the matrix $A-DX'_+$ has all its eigenvalues in the closed left halfplane. Also, $X'_+\geqq X$ for every hermitian solution of (1). In particular, taking $C'=C$ we see that the hermitian solution X'_+ of (1) is maximal. It follows now that $X'_+\geqq X_+$, where X_+ is the maximal hermitian solution of (1). □

3. Existence of hermitian solutions in geometric terms. We continue our discussion of (2.1) with the hypothesis $D\geqq 0$, $C=C^*$ and (A, D) stabilizable. Write

$$(1)\qquad A=\begin{bmatrix}A_1 & A_{12}\\ 0 & A_2\end{bmatrix},\quad D=\begin{bmatrix}D_1 & 0\\ 0 & 0\end{bmatrix},\quad C=\begin{bmatrix}C_1 & C_{12}\\ C_{12}^* & C_2\end{bmatrix}$$

with respect to the decomposition $\mathbb{C}^n=\mathcal{C}\oplus\mathcal{C}^\perp$, where $\mathcal{C}$ is the controllable subspace of (A, D), as in (2.4). So (A_1, D_1) is controllable, and A_2 is stable. Letting

$$X=\begin{bmatrix}X_1 & X_{12}\\ X_{12}^* & X_2\end{bmatrix}$$

be partitioned conformally with (1), it is found that (2.1) is equivalent to the three equations:

$$X_1D_1X_1 - X_1A_1 - A_1^*X_1 - C_1 = 0, \tag{2}$$

$$(A_1^* - X_1D_1)X_{12} + X_{12}A_2 = -(C_{12} + X_1A_{12}), \tag{3}$$

$$A_2^*X_2 + X_2A_2 = X_{12}^*D_1X_{12} - X_{12}^*A_{12} - A_{12}^*X_{12} - C_2. \tag{4}$$

(Molinari [9] and Wimmer [17] have also taken advantage of this observation.)

The existence of a hermitian solution of (2) is equivalent to any one of the following properties:

(a) There exists an m-dimensional, H-neutral, M_1-invariant subspace of $C \times C$, where $m = \dim C$,

$$H = i\begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix} \quad \text{and} \quad M_1 = \begin{bmatrix} -A_1 & D_1 \\ C_1 & A_1^* \end{bmatrix}$$

and I is the identity map on C. (See [6]. This generalizes the description using eigenvectors of M_1 introduced in [10].) (Recall that a subspace $L \subset C \times C$ is called H-neutral if $(Hx, y) = 0$ for every $x, y \in L$, where $(\cdot, \cdot)$ stands for the standard scalar product in $C \times C$.)

(b) The partial multiplicities of M_1 (i.e. the sizes of Jordan blocks in the Jordan normal form of M_1) which correspond to the pure imaginary eigenvalues of M_1 (if any) are all even (see [6]).

(c) The rational matrix function

$$Z(\lambda) = I + D_0^*(\lambda I + iA_1^*)^{-1}C_1(\lambda I - A_1)^{-1}D_0,$$

where D_0 is any $m \times m$ matrix with $D_1 = D_0D_0^*$, is positive semidefinite for every real λ which is not a pole of $Z(\lambda)$ (see [16] or [3]).

It turns out that these properties determine also the existence of hermitian solutions of (2.1):

THEOREM 3.1. *Equation* (2.1) *has a hermitian solution if and only if there is an m-dimensional, H-neutral invariant subspace of M_1, or, equivalently, if one of the conditions* (b) *and* (c) *holds.*

Proof. Assume that (2) has a hermitian solution. Taking the maximal hermitian solution of (2) for X_1 (which exists by Theorem 2.1), we find that, because all eigenvalues of $A_1 - D_1X_1$ are in the closed left halfplane, equation (3) is uniquely solvable for X_{12}. Then (4) is uniquely solvable for X_2 and this solution is hermitian. Thus a hermitian solution X of (2.1) is obtained.

Conversely, if (2.1) has a hermitian solution X, then (2) holds with X_1 hermitian. □

We shall distinguish between two cases. If

$$\sigma(M_1) \cap \overline{\sigma(A_2)} = \varnothing, \tag{5}$$

where A_2 is defined by (1), the equation (2.1) will be called *regular.* Otherwise, it will be called *nonregular.* This distinction is dictated by the fact that in the regular case the hermitian solutions of (2.1) are completely described in terms of certain M_1-invariant subspaces, while in the nonregular case such a description is only partial. These two cases are discussed separately in the next two sections.

4. The regular case. In this section we shall assume that the equation (2.1) is regular, i.e. M_1 and A_2^* have no common eigenvalues. It turns out that there is a

one-to-one correspondence between the set of hermitian solutions of (1) and the set of M_1-invariant subspaces corresponding to the spectrum of M_1 in the open right halfplane, as follows. We denote by L_+ the maximal M-invariant subspace with $\sigma(M_1|_{L_+})$ lying in the open right halfplane.

THEOREM 4.1. *Assume* (2.1) *has hermitian solutions and is regular. For every* M_1*-invariant subspace* $L \subset L_+$ *there exists a unique hermitian solution* $X = \varphi(L)$ *of* (2.1) *satisfying*

$$\text{Im}\begin{bmatrix} I \\ PXP \end{bmatrix} \cap L_+ = L, \tag{1}$$

where P *is the orthogonal projector on* C. *Conversely, for every hermitian solution* X *of* (2.1) *there is a unique* M_1*-invariant subspace* $L \subset L_+$ *such that* (1) *holds, i.e.* $X = \varphi(L)$.

Proof. Write (3.1) in the equivalent form of (3.2), (3.3), (3.4). As shown in [6] (see also [3]) there is a one-to-one correspondence between the set of hermitian solutions X_1 of (2) and the set of all M_1-invariant subspaces L which are contained in L_+ given by the formula

$$\text{Im}\begin{bmatrix} I \\ X_1 \end{bmatrix} \cap L_+ = L. \tag{2}$$

Further note that the restriction of M_1 to $\text{Im}\,[{}^{I}_{X_1}]$ is similar to $-(A_1 - D_1X_1)$. Therefore the regularity condition (3.5) ensures that equation (3.3) can be uniquely solved for X_{12} for any hermitian solution X_1 of (3.2). Also (3.4) can be uniquely solved for X_2 in view of the stability of A_2. □

The one-to-one correspondence φ in Theorem 4.1 preserves topology, analyticity and a partial order as shown in the next theorem. In connection with topology note that the set of subspaces in $\mathbb{C}^{2m}$ (where $m = \dim C$) is considered as a metric space with the gap metric

$$\theta(L_1, L_2) = \|P_{L_1} - P_{L_2}\|, \qquad L_1, L_2 \subset \mathbb{C}^{2m},$$

where P_{L_i} is the orthogonal projector on L_i, $i = 1, 2$, and the norm is understood in the Hilbert space sense.

THEOREM 4.2. (i) *The correspondence* φ *introduced in Theorem* 4.1 *is a homeomorphism between the set of hermitian solutions of* (2.1) *and the set of all* M_1*-invariant subspaces which are contained in* L_+.

(ii) *Let* L_1 *and* L_2 *be* M_1*-invariant subspaces contained in* L_+, *and let* $X_1 = \varphi(L_1)$, $X_2 = \varphi(L_2)$ *be the corresponding hermitian solutions of* (1). *Then* $L_1 \supset L_2$ *if and only if* $X_1 \geqq X_2$.

(iii) *If* $P(t)$ *is a* $2m \times 2m$ *matrix function which is analytic on an open connected set* $\Omega \subset \mathbb{R}^k$ *and whose values are orthogonal projectors such that for every* $t \in \Omega$, $\text{Im}\,P(t)$ *is an* M_1*-invariant subspace contained in* L_+, *then* $X(t) \overset{\text{def}}{=} \varphi(\text{Im}\,P(t))$ *is an analytic* $n \times n$ *matrix function whose values are hermitian solutions of* (2.1). *Conversely, if* $X(t)$, $t \in \Omega$ *is an analytic matrix function whose values are hermitian solutions of* (2.1), *then the orthogonal projector* $P(t)$ *on* $\varphi^{-1}(X(t))$ *is an analytic* $2m \times 2m$ *matrix function on* Ω *as well.*

Proof. (i) As shown in [13] and Theorem 4.2 of [9], the one-to-one correspondence between the set of hermitian solutions X_1 of (3.2) and the set of all M_1-invariant subspaces $L \subset L_+$ given by (2) is actually a homeomorphism. Now (i) follows from the proof of Theorem 4.1, taking into account the fact that both X_{12} and X_2 depend continuously on the hermitian solution X_1 of (2).

(iii) This part follows from the corresponding analyticity property for hermitian solutions of (3.2) (see [11]) taking into account the fact that the solutions X_{12} and X_2 of equations (3.3) and (3.4), respectively, depend analytically on the data of these equations.

(ii) It is known (see [1], [11]) that the correspondence between hermitian solutions $X_1 = \psi(L)$ of (3.2) and the subspace

$$L = \operatorname{Im} \begin{bmatrix} I \\ X_1 \end{bmatrix} \cap L_+$$

is partial order preserving in the sense that $L_1 \supset L_2$ holds if and only if $\psi(L_1) - \psi(L_2)$ is positive semidefinite. In view of this property, the proof of statement (ii) will follow from the following result which concerns the possibility of extending (or dilating) solutions of (3.2) to solutions of (2.1) in such a way as to retain their partial ordering.

LEMMA 4.3. *Let $D_1, A_1, C_1, A_2, C_2, A_{12}, C_{12}$ be complex matrices of sizes $m \times m$, $m \times m$, $m \times m$, $p \times p$, $p \times p$, $m \times p$, $m \times p$, respectively, where D_1, C_1 and C_2 are hermitian. Assume that $\sigma(-A_2) \cap \overline{\sigma(A_2)} = \varnothing$, and that*

$$\sigma(-(A_1 - D_1 X_1)) \cap \overline{\sigma(A_2)} = \varnothing$$

for every hermitian solution X_1 of the equation

$$X_1 D_1 X_1 - X_1 A_1 - A_1^* X_1 - C_1 = 0. \tag{3}$$

Then for every hermitian solution X_1 of (3) there are unique matrices X_{12} and X_2 such that

$$X = \begin{bmatrix} X_1 & X_{12} \\ X_{12}^* & X_2 \end{bmatrix}$$

satisfies the equation

$$X \begin{bmatrix} D_1 & 0 \\ 0 & 0 \end{bmatrix} X - X \begin{bmatrix} A_1 & A_{12} \\ 0 & A_2 \end{bmatrix} - \begin{bmatrix} A_1^* & 0 \\ A_{12}^* & A_2^* \end{bmatrix} X - \begin{bmatrix} C_1 & C_{12} \\ C_{12}^* & C_2 \end{bmatrix} = 0, \tag{4}$$

and if $X_1 \geqq Y_1$ are hermitian solutions of (3) then the corresponding solutions X and Y of (4) satisfy $X \geqq Y$.

Proof. Equation (4) is equivalent to three equations: Equation (3) together with

$$(A_1 - D_1 X_1)^* X_{12} + X_{12} A_2 = -(C_{12} + X_1 A_{12}),$$

$$A_2^* X_2 + X_2 A_2^* = X_{12}^* D_1 X_{12} - X_{12}^* A_{12} - A_{12}^* X_{12} - C_2.$$

Hence the existence and uniqueness of X_{12} and X_2 follow from the assumptions on the spectra of A_2 and $A_1 - D_1 X_1$.

To prove the second statement of the lemma, we can (and will) assume without loss of generality that $C_1 = 0$, $C_2 = 0$, $C_{12} = 0$ and that A_2 is in upper triangular form. Indeed, for a fixed hermitian solution $\tilde{X}$ of (4), a hermitian matrix X is a solution of (4) if and only if

$$(X - \tilde{X}) D (X - \tilde{X}) - (X - \tilde{X})(A - D\tilde{X}) - (A^* - D\tilde{X})(X - \tilde{X}) = 0, \tag{5}$$

where

$$A = \begin{bmatrix} A_1 & A_{12} \\ 0 & A_2 \end{bmatrix}, \qquad D = \begin{bmatrix} D_1 & 0 \\ 0 & 0 \end{bmatrix},$$

i.e. $X - \tilde{X}$ satisfies an equation analogous to (4) with $C_1 = 0$, $C_2 = 0$, $C_{12} = 0$ and A replaced by $A - D\tilde{X}$. One checks easily that the assumptions on the spectra of A_2 and

of $A_1 - D_1X_1$ remain valid for the equation (5) as well. Further, for any nonsingular $p \times p$ matrix S, X satisfies (4) if and only if

$$\hat{X} = \begin{bmatrix} I & 0 \\ 0 & S^* \end{bmatrix} X \begin{bmatrix} I & 0 \\ 0 & S \end{bmatrix}$$

satisfy the equation

$$\hat{X}\begin{bmatrix} D_1 & 0 \\ 0 & 0 \end{bmatrix}\hat{X} - \hat{X}\begin{bmatrix} A_1 & A_{12}S \\ 0 & S^{-1}A_2S \end{bmatrix} - \begin{bmatrix} A_1^* & 0 \\ S^*A_{12}^* & S^*A_2^*S^{-1*} \end{bmatrix}\hat{X} - \begin{bmatrix} C_1 & C_{12}S \\ S^*C_{12}^* & S^*C_{12}S \end{bmatrix} = 0,$$

and for a suitable S the matrix $S^{-1}A_2S$ is upper triangular.

For a fixed j, $1 \leqq j \leqq p-1$, write

$$A_2 = \begin{bmatrix} A_{22} & A_{23} \\ 0 & A_{33} \end{bmatrix}, \qquad A_{12} = [A_{122}, \quad A_{123}],$$

where A_{22}, A_{23}, A_{33}, A_{122} and A_{123} are matrices of sizes $j \times j$, $j \times (p-j)$, $(p-j) \times (p-j)$, $m \times j$ and $m \times (p-j)$, respectively. It is easy to see that, since $\sigma(-A_2) \cap \overline{\sigma(A_2)} = \varnothing$, the assumptions of the lemma hold with D_1, A_1, A_2, A_{12} replaced by

$$\begin{bmatrix} D_1 & O_{m\times j} \\ O_{j\times m} & O_{j\times j} \end{bmatrix}, \quad \begin{bmatrix} A_1 & A_{122} \\ 0 & A_{22} \end{bmatrix}, \quad A_{33}, \quad \begin{bmatrix} A_{123} \\ A_{23} \end{bmatrix},$$

respectively (here $O_{r\times s}$ stands for the $r \times s$ zero matrix). Hence, using induction on the size p of the matrix A_2, we can restrict ourselves to the case $p = 1$, i.e. $A_2 = a_2$ is a scalar. In this case (4) is equivalent to the three equations

$$X_1D_1X_1 - X_1A_1 - A_1^*X_1 = 0, \tag{6}$$

$$(A_1 - D_1X_1)^*X_{12} + a_2X_{12} = -X_1A_{12}, \tag{7}$$

$$(a_2 + \bar{a}_2)X_2 = X_{12}^*D_1X_{12} - X_{12}^*A_{12} - A_{12}^*X_{12}. \tag{8}$$

Equation (7) gives (recall that, from the hypotheses of the lemma, $a_2I + (A_1 - D_1X_1)^*$ is nonsingular)

$$X_{12} = -[a_2I + (A_1 - D_1X_1)^*]^{-1}X_1A_{12}.$$

But in view of (6) the equality

$$[\lambda I + (A_1 - D_1X_1)^*]^{-1}X_1 = X_1(\lambda I - A_1)^{-1}$$

holds for every λ not belonging to $\overline{\sigma(-A_1 + D_1X_1)} \cup \sigma(A_1)$ and therefore

$$[a_2I + (A_1 - D_1X_1)^*]^{-1}X_1 = X_1(a_2I - A_1)^{-1},$$

where the right-hand side is meaningful in the sense that $X_1(a_2I - A_1)^{-1} = \lim_{\lambda \to a_2}[X_1(\lambda I - A_1)^{-1}]$ even when $a_2 \in \sigma(A_1)$. Hence

$$X_{12} = -X_1(a_2I - A_1)^{-1}A_{12}.$$

Now (8) becomes

$$\begin{aligned}(a_2 + \bar{a}_2)X_2 = {} & A_{12}^*(\bar{a}_2I - A^*)X_1D_1X_1(a_2I - A_1)^{-1}A_{12} + A_{12}^*(\bar{a}_2I - A_1^*)^{-1}X_1A_{12} \\ & + A_{12}^*X_1(a_2I - A_1)^{-1}A_{12},\end{aligned}$$

and assuming $a_2 \notin \sigma(A_1)$ and inserting $X_1A_1 + A_1^*X_1$ in place of $X_1D_1X_1$, we obtain eventually

$$(a_2 + \bar{a}_2)X_2 = A_{12}^*(a_2I - A_1^*)^{-1}(a_2 + \bar{a}_2)X_1(a_2I - A_1)^{-1}A_{12}.$$

As $a_2+\bar{a}_2\neq 0$, we have

$$X_2=A_{12}^*(\bar{a}_2I-A_1^*)^{-1}X_1(a_2I-A_1)^{-1}A_{12}$$

and

$$X=\begin{bmatrix} X_1 & -X_1(a_2I-A_1)^{-1}A_{12} \\ -A_{12}^*(\bar{a}_2I-A_1^*)^{-1}X_1 & A_{12}^*(\bar{a}_2I-A_1^*)^{-1}X_1(a_2I-A_1)^{-1}A_{12}\end{bmatrix}$$
$$=K^*\begin{bmatrix} X_1 & X_1 \\ X_1 & X_1\end{bmatrix}K,$$

where $K=\operatorname{diag}(I,-(a_2I-A_1)^{-1}A_{12})$.

Analogously, for a hermitian solution Y_1 of (6) and the corresponding solution

$$Y=\begin{bmatrix} Y_1 & Y_{12} \\ Y_{12}^* & Y_2\end{bmatrix}$$

of (6), (7) and (8) we have

$$Y=K^*\begin{bmatrix} Y_1 & Y_1 \\ Y_1 & Y_1\end{bmatrix}K.$$

So if $X_1\geqq Y_1$, then

$$\begin{bmatrix} X_1 & X_1 \\ X_1 & X_1\end{bmatrix}\geqq\begin{bmatrix} Y_1 & Y_1 \\ Y_1 & Y_1\end{bmatrix}$$

and consequently $X\geqq Y$. The assumption $a_2\notin\sigma(A_1)$ is immaterial, because one can take a complex sequence $\{b_m\}_{m=1}^{\infty}$ tending to a_2 with $b_m\notin\sigma(A_1)$, apply the already proved result for each b_m (in place of a_2), and pass to the limit when $m\to\infty$. □

As a corollary to Theorem 4.2 we obtain the existence of minimal solutions:

COROLLARY 4.4. *Under the regularity assumption* $\sigma(M_1)\cap\overline{\sigma(A_2)}=\varnothing$, *there exists a minimal hermitian solution of* (2.1).

Indeed, the hermitian solution $X=\varphi(\{0\})$ (in the notation of Theorem 4.1) is minimal in view of Theorem 4.2.

5. The nonregular case. We consider now the nonregular case:

$$\sigma(M_1)\cap\overline{\sigma(A_2)}\neq\varnothing.$$

Let $\tilde{L}_+$ be the maximal M_1-invariant subspace such that $\sigma(M_1|_{\tilde{L}_+})$ lies in the open right halfplane and $\sigma(M_1|_{\tilde{L}_+})\cap\sigma(-A_2)=\varnothing$. Such an $\tilde{L}_+$ always exists; if $\sigma(M_1|_L)\cap\sigma(-A_2)\neq\varnothing$ for every nonzero M-invariant subspace L with $\sigma(M_1|_L)$ lying in the open right halfplane, we put $\tilde{L}_+=\{0\}$. We have now a partial description of hermitian solutions of (2.1) in terms of M_1-invariant subspaces contained in $\tilde{L}_+$:

THEOREM 5.1. *Assume that* (2.1) *has a hermitian solution. For every* M_1*-invariant subspace* $L\subset\tilde{L}_+$ *there exists a unique hermitian solution* $X=\tilde{\varphi}(L)$ *of* (2.1) *satisfying*

$$\operatorname{Im}\begin{bmatrix} I \\ PMP\end{bmatrix}\cap\tilde{L}_+=L, \tag{1}$$

where P *is the orthogonal projector on the controllable subspace* C *of* (A, D). *The solution* X *satisfies*

$$\sigma(-A|_C+PDPXP)\cap\overline{\sigma(A_2)}=\varnothing. \tag{2}$$

Conversely, for every hermitian solution X of (2.1) *with* $\sigma(-A|_C + PCPXP) \cap \overline{\sigma(A_2)} = \varnothing$ *there exists a unique* M_1*-invariant subspace* $L \subset \tilde{L}_+$ *such that* (1) *holds. The correspondence* $\tilde{\varphi}$ *between* M_1*-invariant subspaces contained in* $\tilde{L}_+$ *and hermitian solutions satisfying* (2) *is analytic and a homeomorphism and preserves the partial order.*

The continuity, analyticity and the partial order preservation of $\tilde{\varphi}$ are understood as in Theorem 4.2.

Theorem 5.1 is proved in the same way as Theorem 4.2.

It can happen, in general, that there exist hermitian solutions X of (2.1) for which (2) does not hold (see Example 2.1). In this case, for at least one hermitian solution X_1 of (3.2) the linear equation (3.3) has many solutions X_{12}, and therefore the set of hermitian solutions of (2.1) is not compact. In particular, there is no minimal hermitian solution. Conversely, if the set of hermitian solutions of (2.1) is compact, then equation (3.3) has no solution X_{12} for any hermitian solution X_1 of (3.2) satisfying $\sigma(-A_1 + D_1X_1) \cap \overline{\sigma(A_2)} = \varnothing$. Hence all hermitian solutions X of (2.1) satisfy (2). In particular, Theorem 5.1 ensures the existence of a minimal hermitian solution (which is obtained from $L = (0)$). We have proved the following:

COROLLARY 5.2. *The following statements are equivalent for the equation* (2.1) *with* $D \geqq 0$, $C^* = C$ *and* (A, D) *stabilizable* (*in either the regular or nonregular case*):

(i) *the set of hermitian solutions of* (2.1) *is compact*;

(ii) *there exists a minimal hermitian solution of* (2.1);

(iii) *every hermitian solution X of* (2.1) *satisfies*

$$\sigma(-A|_C - PDPXP) \cap \overline{\sigma(A_2)} = \varnothing,$$

where P is the orthogonal projector on the controllable subspace C of (A, D), and $A_2 = (I - P)A(I - P)$.

6. The real case. Consider the equation

$$XDX - XA - A^*X - C = 0 \tag{1}$$

with $D \geqq 0$, $C = C^*$, (A, D) stabilizable, and assume in addition that A, C, D are real matrices. Naturally, in this case real symmetric solutions X of (1) are sought. The results of § 2, together with their proofs, remain valid if "hermitian" is replaced by "real symmetric" throughout § 2. Note that the maximal hermitian solution of (1) is necessarily real. Note also that if there is a hermitian solution of (1), then there is also a real symmetric solution.

To characterize real symmetric solutions of (1) in terms of invariant subspaces, introduce the M_1-invariant subspace $L_{++} \subset \mathbb{C}^{2m}$ which corresponds to the eigenvalues of M_1 lying in the quadrant $\{\lambda \in \mathbb{C} \mid \operatorname{Re} \lambda > 0, \operatorname{Im} \lambda \geqq 0\}$.

THEOREM 6.1. *Assume* (1) *has a real symmetric solutions and is regular. For every* M_1*-invariant subspace* $L \subset L_{++}$ *there exists a unique real symmetric solution* $X = \tau(L)$ *of* (1) *satisfying*

$$\operatorname{Im} \begin{bmatrix} I \\ PXP \end{bmatrix} \cap L_{++} = L, \tag{2}$$

where P is the orthogonal projector on the controllable subspace of (A, D). Conversely, for every real symmetric solution X of (1) *there exists a unique M-invariant subspace* $L \subset L_{++}$ *such that* (2) *holds, i.e.* $X = \tau(L)$. *The correspondence φ is homeomorphic, analytic, and preserves partial order.*

The proof of Theorem 6.1 is obtained in the same way as the proof of Theorems 4.1 and 4.2 using Theorem II.4.14 of [3].

In the nonregular case a result holds for real symmetric solutions of (1) which is analogous to Theorem 5.1. We leave its statement to the reader. Corollary 5.2 is valid in the real symmetric case as well. Finally, we remark that if a minimal hermitian solution of (1) exists, it is necessarily real.

REFERENCES

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

[2] W. A. COPPEL, *Matrix quadratic equations*, Bull. Austral. Math. Soc., 10 (1974), pp. 377-401.

[3] I. GOHBERG, P. LANCASTER AND L. RODMAN, *Matrices and Indefinite Scalar Products*, Birkhauser Verlag, Basel, 1983.

(4) R. E. KALMAN, *Contributions to the theory of optimal control*, Bol. Soc. Mat. Mexicana, 5 (1960), pp. 102-119.

[5] D. L. KLEINMAN, *On an iterative technique for Riccati equation computation*, IEEE Trans. Automat. Control, 13 (1968), pp. 114-115.

[6] P. LANCASTER AND L. RODMAN, *Existence and uniqueness theorems for algebraic Riccati equations*, Internat. J. Control, 32 (1980), pp. 285-309.

[7] P. LANCASTER AND M. TISMENETSKY, *Theory of Matrices*, 2nd Edition, Academic Press, New York, 1985.

[8] D. L. LUKES, *Equilibrium feedback control in linear games with quadratic costs*, SIAM J. Control, 9 (1971), pp. 234-252.

[9] B. P. MOLINARI, *The stabilizing solution of the algebraic Riccati equation*, SIAM J. Control, 11 (1973), pp. 262-271.

[10] J. E. POTTER, *Matrix quadratic solutions*, SIAM J. Appl. Math., 14 (1966), pp. 496-501.

[11] A. C. M. RAN AND L. RODMAN, *Stability of invariant maximal semidefinite subspaces* II. *Applications: selfadjoint rational matrix functions, algebraic Riccati equations*, Linear Algebra Appl., to appear.

[12] ———, *The algebraic matrix Riccati equation*, in Topics in Operator Theory, Systems and Networks, the Rehovot Workshop, H. Dym and I Gohberg, eds., Operator Theory: Advances and Applications, Vol. 12, Birkhauser Verlag, Basel, 1984, pp. 351-381.

[13] W. T. REID, *Riccati matrix differential equations and non-oscillation criteria for associated linear differential systems*, Pacific J. Math., 13 (1963), pp. 665-685.

[14] D. L. RUSSEL, *Mathematics of Finite-Dimensional Control Systems*, Marcel Dekker, New York, 1979.

[15] M. A. SHAYMAN, *Geometry of the algebraic Riccati equation* II, this Journal, 21 (1983), pp. 395-409.

[16] J. C. WILLEMS, *Least squares stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Automat. Control, 16 (1971), pp. 621-634.

[17] H. K. WIMMER, *The algebraic Riccati equation: conditions for existence and uniqueness of solutions*, Linear Algebra Appl., 58 (1984), pp. 441-452.

[18] W. M. WONHAM, *On a matrix Riccati equation of stochastic control*, SIAM J. Control, 6 (1968), pp. 681-697. Erratum, ibid., 7 (1969), p. 365.

[19] ———, *Linear Multivariable Control*, Springer-Verlag, New York-Heidelberg-Berlin, 1974.