SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 1, January 1980

APPLICATIONS OF THE FINITE LAPLACE TRANSFORM TO LINEAR CONTROL PROBLEMS*

RICHARD DATKO†

Abstract. The finite Laplace transform is applied to various control problems involving linear ordinary and linear partial differential equations. Since the finite Laplace transform is an entire function, rather explicit conditions can be given concerning the nature of the controls.

1. Introduction. Consider a differential vector function $f: [0, T] \rightarrow R^n$. Let s be an arbitrary complex number. The finite Laplace transform of f is

$$\begin{aligned} G(s) &= \int_0^T e^{-st}\dot{f}(t)\,dt = f(T)\,e^{-sT} - f(0) + s\int_0^T e^{-st}f(t)\,dt \\ &= f(T)\,e^{-sT} - f(0) + sF(s). \end{aligned} \tag{1.1}$$

The complex valued functions F and G are entire and satisfy the conditions

$$\int_{-\infty}^{\infty} |F(i\omega)|^2\,d\omega < \infty, \qquad \int_{-\infty}^{\infty} |G(i\omega)|^2\,d\omega < \infty \tag{1.2}$$

(see e.g. [4]). Exploiting these two properties and the fact that the initial and terminal states of f occur in the finite Laplace transform of f it is possible to compute controls for a variety of autonomous linear control problems expressed by ordinary or partial differential equations. Briefly what one does in the case of ordinary differential equations is converts the given problem to a finite Laplace transform and isolates the transform of the trajectory, $\bar{x}(s)$, on the left hand side of a certain equation. On the right hand side of the equation the initial and terminal states of the system and the finite Laplace transform of the control function explicitly occur in linear combination multiplied by a holomorphic complex valued linear operator. If the initial and terminal states of the trajectory are given then the constraint that its finite Laplace transform be an entire vector function imposes conditions on the transformed control $\bar{\mu}(s)$, namely that the numerators of certain expressions have zeros of the same order as the poles of the holomorphic operator. These conditions can then be used to find the finite Laplace transform of the control which guides the system from its initial to its terminal state. A variant of the above technique can be used to solve the quadratic regulator problem over finite time intervals. Problems of this type are the content of § 2.

Sections 3 and 4 discuss the case of control problems involving certain types of hyperbolic and parabolic partial differential equations. In these problems the finite Laplace transform converts the original system into a linear elliptic partial differential equation with a forcing term depending on initial, terminal and distributed data. A Green's function for the elliptic nonhomogeneous system is constructed. This function is holomorphic in the complex variable s. The poles of the Green's function determine conditions which the finite Laplace transform of the boundary or distributed data must satisfy. This information, at least in the case of the hyperbolic examples in this paper, permits one to construct finite Laplace transforms of boundary or distributed controls which steer the initial state to the zero state in a finite time.

Sections 3 and 4 of this paper overlap previous work in [9] and [11]. In [11] Russell reduced a control problem for a hyperbolic partial differential equations to construction

* Received by the editors July 10, 1978, and in revised form February 13, 1979.

† Department of Mathematics, Georgetown University, Washington, D.C. 20057.

of a function f which satisfies condition (3.28) of Example 3.2 in this paper. (His conditions are given by Equations (2.24)–(2.27) in [11].) Russell arrived at this condition by observing that an associated homogeneous equation is a Sturm–Liouville eigenvalue problem and hence the Fourier method applies. We obtain the same condition by forcing the finite Laplace transform of the distributed control to be such that it cancels out the poles of a holomorphic family of Green's functions. In the most general setting of system (3.1) given in this paper, the Sturm–Liouville approach will not be applicable since the coefficient $D(x)$ of μ_t in (3.1) prohibits a separation of variables approach.

The one dimensional heat equation considered in Section 4 is a slightly generalized version of the work in [9]. In that paper the finite Laplace transform was used to construct a control which could bring a body from a uniform nonzero temperature to a uniform zero temperature in a finite time. However our viewpoint and that taken in [9] are somewhat different. In [9] Goldwyn et al. seek only bang-bang controls. In this paper we determine conditions which the finite Laplace transform of an admissible control must satisfy and then try to fit entire functions to these conditions. When "bang-bang" controls are sought the problem reduces to the one considered in [9].

The finite Laplace transform does not introduce new properties of linear control systems, but, we believe, it does offer a useful computational tool which can be used to attack control problems expressed by autonomous differential systems. Furthermore it unifies the study of linear autonomous nonhomogeneous differential equations in that it does not discriminate between initial value problems and boundary value problems. That is, a nonhomogeneous linear autonomous ordinary differential or partial differential equation with mixed initial and boundary data is converted by a finite Laplace transform into a problem in which the initial data, the terminal data and the transforms of the forcing terms (or boundary values in the case of partial differential equations) appear explicitly and linearly in the transform. Thus if any two are given they determine nontrivial conditions which the third must satisfy. For example, if the initial data and the forcing terms are known, then the condition that the terminal data be such that a finite Laplace transform exists is equivalent to inverting the ordinary Laplace transform of the original system.

The exposition in this paper is primarily through examples. Its purpose is to demonstrate that converting a linear control system to a finite Laplace transform often leads to conditions which permit direct computation of the controls or at least the transforms of the controls. It should be mentioned that one important class of control problem is not discussed in this paper and that is control of linear functional differential equations. This class certainly falls into the same category of problem as do ordinary and partial differential equations. However so far as the application of the finite Laplace transform is concerned there is one important technical difference. That is, for many problems in linear ordinary and partial differential equations it is legitimate to assume a knowledge of the spectra associated with their differential operators. Unfortunately this is not the case with linear functional differential equations.

Preliminaries. The following are definitions and notational conventions which will be used throughout this paper.

1. $\mathscr{C}$ will denote the complex plane, s will denote a point in $\mathscr{C}$. Re s and Im s will stand for the real and imaginary parts of a complex number.

2. I will be for n-dimensional identity matrix. If A is an $n \times n$ matrix, adj A will denote the transposed cofactor matrix of A and A^* will denote the conjugate transpose of A. The symbol $\sigma(A)$ will denote the set of characteristic values of A. If $\lambda \in \sigma(A)$, $\nu(\lambda)$ will denote the index of λ (see e.g. [5, p. 556]).

3. If $f: [0, T] \to R^n$, $T < \infty$, is an L_1 integrable mapping, then the finite Laplace transform of f, denoted by $\bar{f}$, is given by $\int_0^T e^{-st} f(t)\, dt$. Throughout this paper the finite Laplace transform will be abbreviated to F.L.T., and its dependence on T will not in general be emphasized.

4. The characteristic function of a measurable set E in R will be denoted by $\chi_E(t)$.

We now present the fundamental theorem on which this paper is based (see e.g. [4, pp. 238, 241]).

THEOREM 1.1. *Let $\bar{f}: C \to C$ be an entire function of exponential type, i.e. $|\bar{f}(s)| \leqq a\, e^{b|s|}$ for all $s \in C$ and fixed constants a and b. Then there exist nonnegative constants T and T' and a function $f \in L^2(-\infty, +\infty)$ with $f(t) = 0$ if $t \notin [-T', T]$ and $\bar{f}(s) = \int_{-T}^{T} e^{-st} f(t)\, dt$, if and only if $\int_{-\infty}^{\infty} |\bar{f}(i\omega)|^2\, d\omega < \infty$. Moreover the constants T' and T satisfy the relations*

$$T' = \overline{\lim_{x \to \infty}} \frac{1}{x} \ln |\bar{f}(x)|,$$

$$T = \overline{\lim_{x \to \infty}} \frac{1}{x} \ln |\bar{f}(-x)|.$$

2. Applications of the finite Laplace transform to finite dimensional control problems. Consider the n-dimensional time optimal control problem

$$\dot{x}(t) = Ax(t) + B\mu(t), \qquad x(0) = x_0. \tag{2.1}$$

Here A is an $n \times n$ matrix, B is an $n \times m$ matrix, μ is a measurable m-vector constrained to lie in some compact convex set, $\Omega \subset R^m$, called the control set. The problem is, given x_1 and R^n, select a measurable control μ, with values $\mu(t) \in \Omega$, $t \geqq 0$, such that after some finite time, T, the solution of (2.1), $x(t, x_0, \mu)$, satisfies $x(T, x_0, \mu) = x_1$. Moreover it is desired that T be the smallest possible number for which this can be accomplished. From the general theory of optimal control (see e.g. [6]) we know that if such a μ exists it can be chosen such that, for each t in $[0, T]$, $\mu(t)$ lies on the boundary of Ω.

Suppose there is a measurable $\mu: [0, T] \to \Omega$ such that the solution of (2.1) satisfies $x(0, x_0, \mu) = x_0$ and $x(T, x_0, \mu) = x_1$. Then the F.L.T. of (2.1) for this solution can be written

$$\bar{x}(s) = (sI - A)^{-1}[x_0 - x_1\, e^{-sT} + B\bar{\mu}(s)]. \tag{2.2}$$

Since $\bar{x}(s)$ is an F.L.T. this implies that for each $\lambda \in \sigma(A)$

$$\frac{d^k}{ds^k}[\operatorname{adj}\,(sI - A)(x_0 - x_1\, e^{-sT} + B\bar{\mu}(s))]|_{s=\lambda} = 0, \qquad 0 \leqq k \leqq \nu(\lambda) - 1. \tag{2.3}$$

Conversely, if for a given $T < \infty$ and measurable $\mu: [0, T] \to \Omega$ (2.3) is satisfied for all $\lambda \in \sigma(A)$, then by Theorem 1.1 μ is a control which transfers the solution of (2.1) from x_0 at time $t = 0$ to x_1 at time T. Thus we can state the following theorem.

THEOREM 2.1. *A necessary and sufficient condition for a measurable mapping $\mu: [0, T] \to \Omega$ to transfer the solution of* (2.1) *from x_0 at $t = 0$ to x_1 at $t = T$ is that equation* (2.3) *be satisfied for all $\lambda \in \sigma(A)$.*

The following examples will demonstrate the utility of (2.3) in the computation of optimal controls.

Example 2.1. Consider the system

$$\begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 0 \\ \mu \end{pmatrix}, \qquad \begin{pmatrix} x(0) \\ y(0) \end{pmatrix} = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} \quad \text{and} \quad |\mu(t)| \leqq 1. \tag{2.4}$$

Let

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{2.5}$$

For this system (2.2) has the explicit form

$$\begin{pmatrix} \bar{x}(s) \\ \bar{y}(s) \end{pmatrix} = \frac{\begin{pmatrix} s & 1 \\ -1 & s \end{pmatrix}}{s^2+1} \left[\begin{pmatrix} x_0 \\ y_0 \end{pmatrix} + \begin{pmatrix} 0 \\ \bar{\mu}(s) \end{pmatrix} \right]. \tag{2.6}$$

Since we know the optimal controls are "bang-bang" (see e.g. [6]), we may assume that

$$\mu(t) = \varepsilon_0[\chi_{[0,t_1)}(t) - \chi_{[t_1,t_2)}(t) + \cdots + (-1)^m \chi_{[t_m,T]}(t)]$$

where $\varepsilon_0 = \pm 1$, $0 \leqq t_1 \leqq t_2 \cdots \leqq T$. Thus

$$\bar{\mu}(s) = \varepsilon_0 \left[\int_0^{t_1} e^{-st}\,dt - \int_{t_1}^{t_2} e^{-st}\,dt + \cdots + (-1)^m \int_{t_m}^{T} e^{-st}\,dt \right]. \tag{2.7}$$

The spectrum, $\sigma(A)$, for this problem is $\lambda = \pm i$. Substituting these values into (2.3) we obtain the following independent equations.

$$\begin{aligned} ix_0 + y_0 + \bar{\mu}(i) &= 0. \\ -ix_0 + y_0 + \bar{\mu}(-i) &= 0. \end{aligned} \tag{2.8}$$

Using (2.7) and (2.8) we solve for x_0 and y_0 to obtain

$$\begin{aligned} x_0 &= \varepsilon_0 \left[1 + 2 \sum_{j=1}^{m} (-1)^j \cos t_j + (-1)^{m+1} \cos T \right] \\ y_0 &= \varepsilon_0 \left[2 \sum_{j=1}^{m} (-1)^j \sin t_j + (-1)^{m+1} \sin T \right], \\ \varepsilon_0 &= \pm 1, \qquad 0 \leqq t_1 \leqq t_2 \leqq \cdots \leqq T. \end{aligned} \tag{2.9}$$

Thus given $\begin{pmatrix} x_0 \\ y_0 \end{pmatrix}$ the minimum value of T for which (2.9) is satisfied is the optimal time, the value of ε_0 is $\mu(0)$ and the times $t_1, t_2, \cdots, t_m$ are the switching times (i.e. the times at which $\mu(t)$ changes sign).

Equations (2.9) could have been obtained via the usual method. However this would have required integration of (2.4) which has been bypassed using the finite Laplace transform.

Example 2.2. (See e.g. [1, p. 536–540].) Consider the scalar system

$$\dddot{x} + \ddot{x} = \mu, \quad x(0), \quad \dot{x}(0), \quad \ddot{x}(0). \tag{2.10}$$

Assume $|\mu(t)| \leqq 1$ for all $t \geqq 0$. As in Example 2.1, it is desired to drive the initial values to $x(T) = \dot{x}(T) = \ddot{x}(T) = 0$ in some minimum time T. If this is possible the F.L.T. of (2.10) becomes

$$\bar{x}(s) = \frac{1}{s^2(s+1)}[(s^2+s)x(0) - (s+1)\dot{x}(0) + \ddot{x}(0) + \bar{\mu}(s)] \tag{2.11}$$

where $\bar{\mu}(s)$ is the F.L.T. of some measurable μ from $[0, T] \to [-1, 1]$. By Theorems 1.1

or 2.1 this implies that

$$
\begin{aligned}
&x(0)+\dot{x}(0)+\bar{\mu}'(0)=0,\\
(2.12)\qquad &\dot{x}(0)+\ddot{x}(0)+\bar{\mu}(0)=0,\\
&\ddot{x}(0)+\bar{\mu}(-1)=0.
\end{aligned}
$$

Since we may assume $\mu(t)$ is "bang-bang" ([6]) and that there are at most two switch (see e.g. [1]) $\bar{\mu}(s)$ must be of the form

$$
(2.13)\qquad \bar{\mu}(s)=\varepsilon_0\left[\int_0^{t_1} e^{-st}\,dt-\int_{t_1}^{t_2} e^{-st}\,dt+\int_{t_2}^{T} e^{-st}\,dt\right]
$$

where $\varepsilon_0=\pm 1$, $0\leqq t_1\leqq t_2\leqq T$. Thus between (2.12) and (2.13) we are led by some simple calculations to the equations

(2.14)

$$
\begin{aligned}
&\ddot{x}(0)=-\bar{\mu}(-1)=\varepsilon_0[1-2\,e^{t_1}+2\,e^{t_2}-e^{T}],\\
&\dot{x}(0)=\bar{\mu}(-1)-\bar{\mu}(0)=\varepsilon_0[-1+2\,e^{t_1}-2\,e^{t_2}-e^{T}-2t_1+2t_2-T],\\
&x(0)=-\bar{\mu}(-1)+\bar{\mu}(0)-\bar{\mu}'(0)=\varepsilon_0\left[1-2\,e^{t_1}+2\,e^{t_2}-e^{T}+2t_1-2t_2+T+t_1^2-t_2^2+\frac{T^2}{2}\right],\\
&\qquad\qquad 0\leqq t_1\leqq t_2\leqq T,\qquad \varepsilon_0+\pm 1.
\end{aligned}
$$

As in Example 2.1 the smallest value of T for which (2.14) is satisfied is the optimal time of transfer to the origin, the switching times are t_1 and t_2 and ε_0 is the value of $\mu(0)$.

Another type of finite dimensional control problem to which the F.L.T. may be applied is to linear autonomous systems with quadratic cost. Thus let R be an n-dimensional positive semi-definite matrix with real entries, W a real positive n-dimensional real matrix and U a real positive m-dimensional matrix. Consider the problem of minimizing the cost functional

$$
(2.15)\qquad C(\mu)=(Rx(T),x(T))+\int_0^T [(Wx(t),x(t))+(U\mu(t),\mu(t))]\,dt
$$

where $\int_0^T \|\mu(t)\|^2\,dt<\infty$ and $x(t)$ is constrained by the differential equation

$$
(2.16)\qquad \dot{x}(t)=Ax(t)+B\mu(t),\qquad x(0)=x_0.
$$

We assume $T<\infty$.

The solution of the problem (2.15)–(2.16) may be obtained by solving the following $2n$-dimensional system of equations (see e.g. [7]).

$$
\begin{aligned}
&\dot{x}=Ax-BU^{-1}B^{*}q(t),\\
&\dot{q}=-Wx(t)-A^{*}q(t),\\
(2.17)\qquad &\mu(t)=-U^{-1}B^{*}q(t),\\
&x(0)=x_0,\qquad q(T)=Rx(T).
\end{aligned}
$$

The optimal cost $C(\mu)$ is given by

$$
(2.18)\qquad C(\mu)=(q(0),x(0))
$$

where q satisfies (2.17).

The problem may thus be reduced to finding the solution of a two point boundary value problem for a $2n$-dimensional system of linear differential equations. Treating $x(0)$ as known, the F.L.T. of (2.17) is given by

$$\begin{pmatrix}\bar{x}(s)\\ \bar{q}(s)\end{pmatrix}=\begin{pmatrix} sI-A & BU^{-1}B^{*}\\ W & sI+A^{*}\end{pmatrix}^{-1}\left|\begin{pmatrix}x(0)\\ q(0)\end{pmatrix}-\begin{pmatrix}x(T)\,e^{-sT}\\ Rx(T)\,e^{-sT}\end{pmatrix}\right|. \tag{2.19}$$

If we are only interested in determining $q(0)$ and $x(T)$ we see that it is not necessary to integrate (2.17) but only to find the values of $x(T)$ and $q(0)$ for which the right side of (2.19) is a finite Laplace transform. On the basis of Theorem 1.1 the following theorem can be stated.

THEOREM 2.2. *The solution of the problem* (2.15)–(2.16) *is found among the n-vectors $q(0)$ and $x(T)$ which make the right-hand side of* (2.19) *a finite Laplace transform.*

Example 2.3. Consider the problem (see e.g. [1])

$$\begin{pmatrix}\dot{x}\\ \dot{y}\end{pmatrix}=\begin{pmatrix}0 & 1\\ 0 & 0\end{pmatrix}\begin{pmatrix}x\\ y\end{pmatrix}+\begin{pmatrix}0\\ \mu\end{pmatrix},\qquad \begin{pmatrix}x(0)\\ y(0)\end{pmatrix}=\begin{pmatrix}x_0\\ y_0\end{pmatrix}, \tag{2.20}$$

with cost

$$C(\mu)=\int_0^T[(x(t))^2+(y(t))^2+(\mu(t))^2]\,dt. \tag{2.21}$$

For this problem $R=0$, $W=I$, $U=1$. Let

$$q=\begin{pmatrix}\mu\\ \nu\end{pmatrix}. \tag{2.22}$$

For this problem (2.19) becomes

$$\begin{pmatrix}\bar{x}(s)\\ \bar{y}(s)\\ \bar{\mu}(s)\\ \bar{\nu}(s)\end{pmatrix}=\frac{\begin{pmatrix} s^3-s & s & 1 & -s\\ -1 & s^3 & s & -s^2\\ 1-s^2 & -s & s^3-s & 1\\ s & 1-s^2 & -s^2 & s^3\end{pmatrix}}{s^4-s^2+1}\begin{pmatrix}x(0)-x(T)\,e^{-sT}\\ y(0)-y(T)\,e^{-sT}\\ \mu(0)\\ \nu(0)\end{pmatrix}. \tag{2.23}$$

The roots of the denominator on the right hand side of (2.23) are $s_1=e^{(1/6)\pi_i}$, $s_2=e^{(5/6)\pi_i}$, $s_3=e^{(7/6)\pi_i}$ and $s_4=e^{(11/6)\pi_i}$. Equating to zero the first row of the numerator on the right of (2.23) for $s=s_1$ and $s=s_2$ and taking the real and imaginary parts of the resulting equations we obtain four linear equations in the four unknowns $\mu(0)$, $\nu(0)$, $x(T)$ and $y(T)$. The solution of these equations will be the only solutions for which (2.23) is a finite Laplace transform. To see this observe that the last three rows of the matrix in (2.23) are multiples of the first row for some root $s=s_j$ of the denominator. Hence we will obtain four independent relations in $x(T)$, $y(T)$, $\mu(0)$, and $\nu(0)$ if we equate to zero the real and imaginary parts of the numerator of any row on the right side of (2.23) for any two roots s_j and s_k, $j\neq k$, with $\bar{s}_j\neq s_k$. In this case we have chosen s_1 and

s_2 and the first row. When this is done the matrix equation

$$\begin{bmatrix} 1 & \frac{-\sqrt{3}}{2} & e^{-(\sqrt{3}/2)T}\left(\frac{\sqrt{3}}{2}\cos\frac{T}{2}-\frac{1}{2}\sin\frac{T}{2}\right) & e^{-(\sqrt{3}/2)T}\left(-\frac{\sqrt{3}}{2}\cos\frac{T}{2}-\frac{1}{2}\sin\frac{T}{2}\right) \\ 0 & -\frac{1}{2} & e^{-(\sqrt{3}/2)T}\left(\frac{1}{2}\cos\frac{T}{2}-\frac{\sqrt{3}}{2}\sin\frac{T}{2}\right) & e^{-(\sqrt{3}/2)T}\left(-\frac{1}{2}\cos\frac{T}{2}+\frac{\sqrt{3}}{2}\sin\frac{T}{2}\right) \\ 1 & \frac{\sqrt{3}}{2} & e^{(\sqrt{3}/2)T}\left(-\frac{\sqrt{3}}{2}\cos\frac{T}{2}-\frac{1}{2}\sin\frac{T}{2}\right) & e^{(\sqrt{3}/2)T}\left(\frac{\sqrt{3}}{2}\cos\frac{T}{2}-\frac{1}{2}\sin\frac{T}{2}\right) \\ 0 & -\frac{1}{2} & e^{(\sqrt{3}/2)T}\left(-\frac{1}{2}\cos\frac{T}{2}+\frac{\sqrt{3}}{2}\sin\frac{T}{2}\right) & e^{(\sqrt{3}/2)T}\left(-\frac{1}{2}\cos\frac{T}{2}-\frac{\sqrt{3}}{2}\sin\frac{T}{2}\right) \end{bmatrix} \begin{pmatrix} \mu(0) \\ \nu(0) \\ x(T) \\ y(T) \end{pmatrix}$$

$$= \begin{pmatrix} \frac{\sqrt{3}}{2}x(0)-\frac{\sqrt{3}}{2}y(0) \\ -\frac{x(0)}{2}-\frac{y(0)}{2} \\ -\frac{\sqrt{3}}{2}x(0)+\frac{\sqrt{3}}{2}y(0) \\ -\frac{x(0)}{2}-\frac{y(0)}{2} \end{pmatrix} \tag{2.24}$$

is obtained.

Remark 2.1. The last example shows that the quadratic cost problem may be solved by a method conceptually simpler than the usual Riccati method (see e.g. [1]) since it bypasses direct integration of (2.17) and yet yields the terminal value $x(T)$ and the initial value $q(0)$. In general the above procedure may be used to solve the initial value problem for an autonomous n-dimensional matrix Riccati equation of the form

$$\dot{W} = WEW + DW + WF + C, \qquad W(0) = R \tag{2.25}$$

(see e.g. [2]). This is gone using the F.L.T. as follows. We consider the $2n$-dimensional matrix system

$$\begin{aligned} \dot{X} &= -FX(t) - EQ(t), \\ \dot{Q} &= CX(t) + DQ(t), \qquad X(0) = I, \qquad Q(0) = R. \end{aligned} \tag{2.26}$$

On the interval $[0, b)$ for which $X^{-1}(t)$ exists we set

$$W(t) = X^{-1}(t)Q(t) \tag{2.27}$$

and observe that it satisfies (2.25). This is not new; what is new is that we may obtain $X(t)$ and $Q(t)$ by using the requirement that the F.L.T. be an entire function. Thus the F.L.T. of (2.26) is

$$\begin{pmatrix} \bar{X}(s) \\ \bar{Q}(s) \end{pmatrix} = \begin{pmatrix} sI + F & E \\ -C & sI - D \end{pmatrix}^{-1} \begin{pmatrix} I - X(t)\,e^{-st} \\ R - Q(t)\,e^{-st} \end{pmatrix}. \tag{2.28}$$

Hence the matrices $Q(t)$ and $X(t)$ for which the right-hand side of (2.28) is a F.L.T. over $[0, t]$ will solve the Riccati equation (2.26).

3. Applications to hyperbolic partial differential equations. Consider an autonomous hyperbolic partial differential equation with mixed boundary and initial data given by the equations

$$\begin{aligned} &u_{tt}+C(x)u_{xx}+D(x)u_t+E(x)u_x+F(x)=R(x)f(t),\\ &u(x,0)=\phi(x), \qquad u_t(x,0)=\psi(x),\\ &u(0,t)=a_0(t), \qquad u(1,t)=a_1(t). \end{aligned} \tag{3.1}$$

Here C, D, E, F and R are continuous for x in $[0, 1]$, $C(x)\leqq C_0<0$ for some constant C_0 and $t\geqq 0$. The functions ϕ and ψ are assumed to be measurable and bounded and the functions a_0, a_1 and f are integrable over finite intervals in $[0,\infty)$.

Setting

$$\begin{aligned} &U(x,s)=\int_0^T e^{-st}u(x,t)\,dt,\\ &A_0(s)=\int_0^T e^{-st}a_0(t)\,dt, \qquad A_1(s)=\int_0^T e^{-st}a_1(t)\,dt \end{aligned} \tag{3.2}$$

and

$$F(s)=\int_0^T e^{-st}f(t)\,dt \tag{3.3}$$

we can write the F.L.T. of (3.1) in the form

$$\begin{aligned} &C(x)\frac{d^2U}{dx^2}(x,s)+E(x)\frac{dU}{dx}(x,s)+(s^2+sD(x)+F(x))U(x,s)\\ &\quad=s(\phi(x)-u(x,T)\,e^{-sT})+D(x)(\phi(x)-u(x,T)\,e^{-st})+F(s)Q(x)\\ &\qquad U(0,s)=A_0(s), \qquad U_1(s)=A_1(s). \end{aligned} \tag{3.4}$$

We fix s and assume $U_1(x,s)$ and $U_2(x,s)$ are linearly independent solutions of the homogeneous equation

$$C(x)\frac{d^2u}{dx^2}(x,s)+E(x)\frac{dU}{dx}(x,s)+(s^2+sD(x)+F(x))U(x,s)=0 \tag{3.5}$$

such that

$$U_1(0,s)=U_2(1,s)=0. \tag{3.6}$$

(This last condition will be assumed to hold except for at most a countable number of s, as it indeed does when $D(x)\equiv 0$. See e.g. [8, the chapter on boundary value problems].)

In terms of (3.5) and (3.6) we can write the solution of (3.4) in the form

$$\begin{aligned} U(x,s)=&\frac{A_1(s)U_1(x,s)U_2'(1,s)}{\Delta_0(1,s)}-\frac{A_0(s)U_2(x,s)U_1'(0,s)}{\Delta_0(0,s)}\\ &+\int_0^1 G(x,\sigma,s)[(s\phi(\sigma)-u(\sigma,T)\,e^{-sT})+D(\sigma)(\phi(\sigma)-u(\sigma,T)\,e^{-sT})\\ &\qquad+(\psi(\sigma)-u_t(\sigma,T)\,e^{-sT})]\,d\sigma+\int_0^1 G(x,\sigma,s)R(\sigma)F(s)\,d\sigma. \end{aligned} \tag{3.7}$$

In (3.7)

$$U'(x, s) = \frac{d}{dx}(U(x, s)), \tag{3.8}$$
$$\Delta_0(\sigma, s) = -U_2(0, s)U_1'(0, s) \exp\left[-\int_0^\sigma E(\tau)/C(\tau)\, d\tau\right]$$

and

$$G(x, \sigma, s) = \frac{U_2(x, s)U_1(\sigma, s)}{\Delta_0(\sigma, s)}, \qquad 0 \leqq \sigma \leqq x, \tag{3.9}$$
$$G(x, \sigma, s) = \frac{U_1(x, s)U_2(\sigma, s)}{\Delta_0(\sigma, s)}, \qquad x \leqq \sigma \leqq 1.$$

Notice from the form of (3.7) and (3.8) that $U(x, s)$ has a pole of order k at a point s if and only if $\Delta_0(0, s)$ has a zero of order k at the same point, i.e. $U_2(0, s)U_1'(0, s) = 0$. Thus for the right-hand side of (3.7) to be a F.L.T. it is necessary that when we write

$$U(x, s) = \frac{1}{\Delta_0(0, s)} Q(x, s), \tag{3.10}$$

$Q(x, s)$ have a zero of order k whenever $\Delta_0(0, s)$ has a zero of order k. This condition allows us to find $u(x, T)$ and $u_t(x, T)$ for $T > 0$. Conversely given $u(x, T)$ and $u_t(x, T)$, $0 \leqq x \leqq 1$, $T > 0$, it permits us to find conditions on $A_0(s)$, $A_1(s)$ and $F(s)$ such that $a_0(t)$, $a_1(t)$ and $f(t)$ may act as boundary and distributed controls taking ϕ and ψ to $u(\cdot, T)$ and $u_t(\cdot, T)$ in time T. This is the content of Theorems 3.1 and 3.2.

THEOREM 3.1. *Given integrable mappings ϕ and ψ on* $[0, 1]$ *a necessary and sufficient condition that there exist a pair of integrable functions a_0 and a_1 on the interval* $[0, T]$ *which act as boundary controls transferring ϕ, ψ to $u(\cdot, T)$, $u_t(\cdot, T)$ subject to the dynamics* (3.1) *is that the finite Laplace transforms A_0 and A_1 of a_0 and a_1 be such that $Q(x, s)$ in* (3.10) *have a zero of order k whenever $\Delta_0(0, s)$ has a zero of order k.*

THEOREM 3.2. *Given integrable mappings ϕ and ψ on* $[0, 1]$ *a necessary and sufficient condition that there exist an integrable function f on* $[0, T]$ *which acts as a distributed control transferring ϕ, ψ to $u(\cdot, T)$ and $u_t(\cdot, T)$ subject to the dynamics* (3.1) *is that the finite LaPlace transform F of f be such that $Q(x, s)$ in* (3.10) *have a zero of order k whenever $\Delta_0(0, s)$ has a zero of order k.*

Remark 3.1. If instead of the boundary conditions in (3.1) we substitute the conditions

$$\alpha\mu(0, t) + \beta\mu_x(0, t) = q_0(t), \quad \alpha^2 + \beta^2 \neq 0, \tag{3.11}$$
$$\nu\mu(1, t) + \delta\mu_x(1, t) = q_1(t), \quad \gamma^2 + \delta^2 \neq 0$$

α, β, ν, and δ constants, we obtain a theorem similar to Theorem 3.1. This is a consequence of the fact that the F.L.T. of (3.11) has the form

$$\alpha A_0(s) + \beta U'(0, s) = \bar{q}_0(s), \quad \alpha^2 + \beta^2 \neq 0, \tag{3.12}$$
$$\nu A_1(s) + \delta U'(1, s) = \bar{q}_1(s), \quad \gamma^2 + \delta^2 \neq 0.$$

where in (3.12) we let $U'(x_0, s) = (dU/dx)(x, s)|_{x = x_0}$. Differentiation of (3.7) and substitution into (3.12) leads to a linear equation in which we can replace the quantities $\bar{q}_i$, $i = 0, 1$ with the quantities A_i, $i = 0, 1$. Thus the two problems are basically equivalent.

Example 3.1. Consider the control problem

$$
\begin{aligned}
&u_{xx} = u_{tt}, \qquad t \geqq 0, \quad 0 \leqq x \leqq 1,\\
&u(x,0) = \phi(x), \qquad u_t(x,0) = \psi(x),\\
&u(x,T) = u_t(x,T) = 0, \qquad T < \infty \text{ but unspecified.}
\end{aligned}
\tag{3.13}
$$

The object is to select $u(0,t) = a_0(t)$ and $u(1,t) = a_1(t)$ such that (3.13) is satisfied. For the special case of (3.13) equation (3.7) becomes

$$
\begin{aligned}
U(x,s) = \frac{Q(x,s)}{\sinh s} &= \frac{1}{\sinh s}[A_0(s)\sinh(s(1-x)) + A_1(s)\sinh(sx)]\\
&+\int_0^x \frac{\sinh(s(1-x))\sinh(s\sigma)}{s\sinh s}[s\phi(\sigma)+\psi(\sigma)]\,d\sigma\\
&+\int_x^1 \frac{\sinh(sx)\sinh(s(1-\sigma))}{s\sinh s}[s\phi(\sigma)+\psi(\sigma)]\,d\sigma.
\end{aligned}
\tag{3.14}
$$

It is easily seen that no matter what A_0 and A_1 are $s=0$ is not a pole of (3.14). However $\sinh s$ has zeros of order one at the points $s = \pm n\pi i$, $n = 1, 2, \cdots$. Thus by a simple calculation the finite Laplace transform of a_0 and a_1 must satisfy the requirements

$$
(-1)^{n+1}A_0(n\pi i) + A_1(n\pi i) + \frac{(-1)^{n+1}}{n\pi}\int_0^1 \sin n\pi\sigma(n\pi i\phi(\sigma)+\psi(\sigma))\,d\sigma = 0
\tag{3.15}
$$

at these points.

If we set

$$
\begin{aligned}
\phi_n &= \int_0^1 (\sin(n\pi\sigma))\phi(\sigma)\,d\sigma\\
\psi_n &= \int_0^1 (\sin(n\pi\sigma))\psi(\sigma)\,d\sigma
\end{aligned}
\tag{3.16}
$$

$n = \pm 1, \cdots$, we see that (3.15) is equivalent to

$$
(-1)^{n+1}A_0(n\pi i) + A_1(n\pi i) + (-1)^{n+1}\left(i\phi_n + \frac{\psi_n}{n\pi}\right) = 0.
\tag{3.17}
$$

Whatever requirements are placed on the controls their finite Laplace transforms must satisfy (3.17). If, for example, we are merely interested in driving ϕ and ψ to zero we could select

$$
\begin{aligned}
A_0(s) &= \sum_{n=1}^{\infty} \frac{1-e^{-2s}}{s(s^2+n^2\pi^2)} n\pi\psi_n\\
&\text{and}\\
A_1(s) &= \sum_{n=1}^{\infty} (-1)^{n+1}\frac{1-e^{-2s}}{s^2+n^2\pi^2} n\pi\phi_n.
\end{aligned}
\tag{3.18}
$$

These would satisfy (3.17) and are respectively the finite Laplace transforms of the functions

$$
\begin{aligned}
a_0(t) &= \sum_{n=1}^{\infty} \frac{(1-\cos(n\pi t))}{n\pi}\psi_n, \qquad 0 \leqq t \leqq 2,\\
a_1(t) &= \sum_{n=1}^{\infty} (-1)^{n+1}(\sin(n\pi t))\phi_n, \qquad 0 \leqq t \leqq 2.
\end{aligned}
\tag{3.19}
$$

Example 3.2. This example is a special case of a problem considered by Russell [11]. The problem is to control to zero the system

$$\begin{aligned}&u_{tt}=u_{xx}+r(x)u+v(x)f(t), \qquad 0\leqq x\leqq 1, \qquad t\geqq 0,\\ &u(x,0)=\phi(x), \qquad u_t(x,0)=\psi(x),\\ &u(0,t)=u(1,t)=0.\end{aligned} \tag{3.20}$$

The function v is assumed to satisfy certain conditions which will be given below and f: $[0, T]\rightarrow R$, T as yet unspecified, is L_2 integrable over $[0, T]$.

Using (3.7) we can write the transformed solution of (3.20) in the form

$$U(x,s)=\int_0^1 G(x,\sigma,s)[s\phi(\sigma)+\psi(\sigma)]\,d\sigma+\int_0^1 G(x,\sigma,s)v(\sigma)F(s)\,d\sigma. \tag{3.21}$$

$G(x, \sigma, s)$ in (3.21) has the structure

$$\frac{U_2(x,s)U_1(\sigma,s)}{-U_2(0,s)U_1'(0,s)}=G(x,\sigma,s), \qquad 0\leqq\sigma\leqq x,$$

$$\frac{U_1(x,s)U_1(\sigma,s)}{U_2(0,s)U_1'(0,s)}=G(x,\sigma,s), \qquad x\leqq\sigma\leqq 1.$$

$U_1(x, s)$ and $U_2(x, s)$ satisfy

$$U_1(0,s)=U_2(1,s)=0$$

and are, for all but a countable number of s, independent solutions of the differential equation

$$\frac{d^2U}{dx^2}(x,s)+(r(x)-s^2)U(x,s)=0. \tag{3.22}$$

As Russell has pointed out in Section 2 of [11] there exists a strictly increasing sequence of nonnegative numbers $\{\lambda_k\}$, $k=0, 1, \cdots$, such that when $s=\pm\lambda_k i$

$$U_1(x,s)=a(s)U_2(x,s), \qquad a(s)\neq 0,$$

i.e. $G(x, \sigma, s)$ has a pole of order one at these values of s. Moreover the $\{\lambda_k\}$ satisfy the following two conditions, there exists $D>0$ such that

$$\lim_{k\to\infty}\frac{k}{\lambda_k}=D, \qquad \underline{\lim}_{k}\,(\lambda_{k+1}-\lambda_k)=\frac{1}{D}. \tag{3.23}$$

We write

$$\begin{aligned}&F(x,\sigma,s)=U_2(x,s)U_1(\sigma,s), \qquad 0\leqq\sigma\leqq x,\\ &F(x,\sigma,s)=U_1(x,s)U_2(\sigma,s), \qquad \sigma\leqq x\leqq 1.\end{aligned} \tag{3.24}$$

We see that (3.21) is an entire function in s for all $0\leqq x\leqq 1$ if and only if at the points $s=\pm\lambda_k i$

$$\begin{aligned}&\int_0^1 F(x,\sigma,s)[s\phi(\sigma)+\psi(\sigma)]\,d\sigma+\int_0^1 F(x,\sigma,s)v(\sigma)F(s)\,d\sigma=0\\ &=\int_0^1 U_1(\sigma,s)[s\phi(\sigma)+\psi(\sigma)]\,d\sigma+\int_0^1 U_1(\sigma,s)v(\sigma)F(s)\,ds.\end{aligned} \tag{3.25}$$

In [11] Russell assumes $v(x)$ satisfies the conditions

$$\int_0^1 U_1(\sigma, \lambda_k i)v(\sigma)\,d\sigma \neq 0 \quad \text{and} \quad \varliminf_k k\left|\int_0^1 U_1(\sigma, \lambda_k i)v(\sigma)\,d\sigma\right| > 0. \tag{3.26}$$

Under the assumptions (3.26) we seek a function f: $[0, T] \to R$ which is in $L_2[0, T]$ and whose F.L.T. F satisfies (3.25) at $s = \pm\lambda_k i$, i.e.

$$F(s) = \frac{\int_0^1 U_1(\sigma, s)[s\phi(\sigma) + \psi(\sigma)]\,d\sigma}{\int_0^1 U_1(\sigma, s)v(\sigma)\,d\sigma} \tag{3.27}$$

for $s = \pm\lambda_k i$.

Russell [11] shows that this is possible for $T \geqq 2$ (see Theorems 2 and 3 in [11]). Our object is not to duplicate Russell's results, but to indicate how one might proceed to construct a F.L.T., F, satisfying (3.27). First observe that when $s = \pm\lambda_k i$

$$U_1(x, \lambda_k i) = U_1(x, -\lambda_k i)$$

is a real function of x, since for these values of s (3.22) satisfies the usual Sturm-Liouville boundary value problem

$$\frac{d^2U}{dx^2} + (r(x) + \lambda_k^2)U = 0, \qquad U(0) = U(1) = 0.$$

Thus when $s = \pm\lambda_k i$, $F(s)$ must satisfy

$$F(\pm\lambda_k i) = \frac{\int_0^1 U_1(\sigma, \pm\lambda_k i)[\pm\lambda_k i\phi(\sigma) + \psi(\sigma)]\,d\sigma}{\int_0^1 U_1(\sigma, \pm\lambda_k i)\nu(\sigma)\,d\sigma} = \pm i\lambda_k q_k + r_k \tag{3.28}$$

where q_k and r_k are real numbers. Russell [11], using properties of nonharmonic Fourier series, has shown that there exists an f in $L_2[0, T]$ whose F.L.T. satisfies (3.28) for $T \geqq 2$, i.e.

$$F(s) = \int_0^T e^{-st}f(t), \qquad T \geqq 2. \tag{3.29}$$

His method is constructive in that it depends on finding a biorthogonal set in $L_2[0, T]$ (see e.g. [11]).

The procedure given below avoids the use of biorthogonal sets, but presents another difficulty. This is the inversion of a complicated Laplace transform.

The construction of F satisfying (3.28). For each $\lambda_k \neq 0$ we let

$$G_k(s) = \frac{\lambda_k^2}{\pi k} \frac{1 - e^{-2k\pi s/\lambda_k}}{s^2 + \lambda_k^2}(sq_k + r_k), \quad \text{and} \quad G_0(s) = \frac{1 - e^{-s}}{s} r_0 \quad \text{if } 0 \in \{\lambda_k i\}. \tag{3.30}$$

Notice that G_k is a F.L.T. such that

$$G_k(\pm\lambda_k i) = \pm\lambda_k q_k i + r_k.$$

Furthermore

$$\overline{\lim_{t\to\infty}}\frac{1}{t}\ln|G_k(t)|=0$$

and if $\lambda_k\neq 0$

$$\overline{\lim_{t\to\infty}}\frac{1}{t}\ln|G_k(-t)|=\frac{2n\pi}{\lambda_k}.$$

Thus by Theorem 1.1 G_k is the finite Laplace transform of a function g_k whose support is over $[0, 2n\pi/\lambda_k]$. Using work of Redheffer [10] we can show there exists a countable set $S=\{\omega_k i\}$, $\omega_k\geqq 0$, such that $\{\lambda_k i\}\subset S$ and

$$H(s)=s\prod_1^\infty\left(1+\frac{s^2}{\omega_k^2}\right) \tag{3.31}$$

is an entire function which satisfies

$$H(s)=\int_0^T e^{-st}h(t)\,dt, \tag{3.32}$$

$$|H(i\omega)|\leqq M \quad \text{for } \omega\in(-\infty,\infty) \tag{3.33}$$

where $M<\infty$, $T<\infty$ and $h\in L_2[0, T]$. (See also [11, p. 550–551] and use the Laplace transform in place of the Fourier transform.)

For each $\lambda_k i$ we now construct the functions

$$\begin{aligned} F_k(s)&=\frac{H(s)}{s^2+\lambda_k^2}\,\frac{2i\lambda_k}{H'(\lambda_k i)},\\ F_0(s)&=\frac{H(s)}{s}. \end{aligned} \tag{3.34}$$

Since

$$\frac{H'(\lambda_k i)}{\lambda_k i}=\frac{H'(-\lambda_k i)}{-\lambda_k i},$$

$F_k(s)$, as defined by (2.33), satisfies

$$F_k(\lambda_j i)=\delta_{ij}=F_k(-\lambda_j i) \tag{3.35}$$

for each pair of integers k and j. Moreover because of (3.33) it is easy to verify that $|F_k(s)|\leqq M_k<\infty$ for $s=i\omega$ and ω real, and that there exists $f_k\in L_2[0, T]$ such that

$$F_k(s)=\int_0^T e^{-st}f_k(t)\,dt. \tag{3.36}$$

We now define

$$F(s)=\sum_{k=0}^\infty F_k(s)G_k(s). \tag{3.37}$$

By our construction for each k

$$F_k(s)G_k(s) \tag{3.38}$$

is an entire function of s, is the Laplace transform of the convolution of f_k and g_k defined above and hence the support of this product for each k must be at most

$$\left[0,\ T+\frac{2k\pi}{\lambda_k}\right] \quad \text{if } \lambda_k \neq 0$$

and

$$[0,\ T+1] \quad \text{if } \lambda_0 = 0.$$

Thus by assumption (3.24) the maximum support of $f_k g_k$ cannot be larger than

$$[0,\ T+l_0] \tag{3.39}$$

where

$$l_0 = \max\left[1,\ \sup_k \frac{2\pi k}{\lambda_k}\right].$$

This brings us to the question. Is $F(s)$ defined by (2.37) the F.L.T. of some function f over $[0, T+l_0]$? We do not know the answer to this if $(q_k, r_k) \neq (0, 0)$ for an infinite number of k. However if $(q_k, r_k) = (0, 0)$ for $k \geqq k_0$, then the answer is yes. Since in that case $G_k(s) \equiv 0$ for $k \geqq k_0$ and hence

$$f(t) = \sum_{k=0}^{k_0} \int_0^t f_k(t-\sigma) g_k(\sigma)\, d\sigma. \tag{3.40}$$

The last example in this section concerns the boundary control to zero of the two dimensional wave equation in a square region. The mechanics of this example are much like those of the one dimensional wave equation of Example 3.1. However in this example there is some geometry involved which permits us to make statements concerning the boundary control to zero of the two dimensional wave equations for arbitrary simply connected regions of the plane. To be specific, suppose we are given a compact region D in R^2 and we wish to control the two dimensional wave equation to zero in this region. We circumscribe about D some square, S. We may assume, if need be, that the boundaries of S and D are disjoint. Let

$$\begin{aligned} &u_{xx} + u_{yy} = u_{tt}, \\ &\phi(x, y) = u(x, y, 0), \qquad (x, y) \in D, \\ &\psi(x, y) = u_t(x, y, 0), \qquad (x, y) \in D, \end{aligned} \tag{3.41}$$

represent the initial data. We wish to select $u(x, y, t)$, $t > 0$ and $(x, y) \in \partial(D)$ (boundary of D) such that after some as yet unspecified time, T, $u(x, y, T) = u_t(x, y, T) = 0$ for all $(x, y) \in D$. We accomplish this by extending the initial data on D to S. Thus on S we have the boundary control problem

$$\begin{aligned} &u_{xx} + u_{yy} = u_{tt}, \\ &u(x, y, 0) = \hat{\phi}, \qquad \hat{\phi}|_D = \phi, \\ &u_t(x, y, 0) = \hat{\psi}, \qquad \hat{\psi}|_D = \psi. \end{aligned} \tag{3.42}$$

We seek a boundary control

$$u(x, y, t), \qquad (x, y) \in \partial(S) \text{ (boundary of } S), \qquad t > 0$$

such that

$$u(x, y, T) = u_t(x, y, T) = 0, \quad \text{for some } T > 0 \text{ and all } (x, y) \in S.$$

But if we accomplish this we have also solved the original problem. For $u(x, y, t)$ restricted to $(x, y) \in \partial(D)$ will act as a boundary control on $\partial(D)$ which drives the original system to zero in time T. However in general this procedure cannot be time optimal, since it restricts the class of boundary controls acting on $\partial(D)$. The main point is that boundary control on a square is sufficient for boundary control on compact regions in R^2.

Example 3.3. In this example all functions are assumed to have the necessary integrability conditions.

Let

$$S=\{(x, y): 0 \leqq x \leqq 1, 0 \leqq y \leqq 1\} \tag{3.43}$$

and on S consider

$$\begin{aligned} &u_{tt} = u_{xx} + u_{yy}, \\ &u(x, y, 0) = \phi(x, y), \\ &u_t(x, y, 0) = \psi(x, y), \\ &u(0, y, t) = u(1, y, t) = 0. \end{aligned} \tag{3.44}$$

Let the boundary controls be

$$\begin{aligned} &a_0(x, t) = u(x, 0, t), \qquad t>0, \quad 0 \leqq x \leqq 1 \\ &a_1(x, t) = u(x, 1, t), \qquad t>0, \quad 0 \leqq x \leqq 1. \end{aligned} \tag{3.45}$$

We shall attempt to find conditions on the a_i, $i = 0, 1$ such that after some time $T > 0$

$$u(x, y, T) = u_t(x, y, T) = 0, \qquad (x, y) \in S. \tag{3.46}$$

The F.L.T. of (3.44) is

$$s^2U(x, y, s) - s\phi(x, y) - \psi(x, y) = U_{xx}(x, y, s) + U_{yy}(x, y, s). \tag{3.47}$$

$$\begin{aligned} &A_0(x, s) = \int_0^T u(x, 0, t)\, e^{-st}\, dt = U(x, 0, s), \\ &A_1(x, s) = \int_0^T u(x, 1, t)\, e^{-st} = U(x, 1, s), \\ &U(0, y, s) = U(1, y, s) = 0. \end{aligned} \tag{3.48}$$

We shall assume that A_0 and A_1 in (3.48) are representable in the forms

$$\begin{aligned} A_0(x, s) &= \sum_{n=1}^{\infty} A_0^n(s) \sin(n\pi x), \\ A_1(x, s) &= \sum_{n=1}^{\infty} A_1^n(s) \sin(n\pi x). \end{aligned} \tag{3.49}$$

For convenience we shall let

$$\int_0^1 (\sin n\pi\tau)(s\phi(\tau, \sigma) + \psi(\tau, \sigma))\, d\tau = q_n(\sigma), \qquad n = 1, \cdots. \tag{3.50}$$

Using the notation (3.49) and (3.50) and the method of separation of variables the solution of (3.47)–(3.48) can be written

$$
\begin{aligned}
U(x, y, s) = {} & \sum_{n=1}^{\infty} \frac{A_0^n(s) \sin(n\pi x) \sinh \sqrt{s^2+n^2\pi^2}(1-y)}{\sinh \sqrt{s^2+n^2\pi^2}} \\
& + \sum_{n=1}^{\infty} \frac{A_1^n(s) \sin(n\pi x) \sinh \sqrt{s^2+n^2\pi^2}\, y}{\sinh \sqrt{s^2+n^2\pi^2}} \\
& + \sum_{n=1}^{\infty} \sin n\pi x \int_0^y \frac{\sinh \sqrt{s^2+n^2\pi^2}(1-y) \sinh \sqrt{s^2+n^2\pi^2}\, \sigma}{\sqrt{s^2+n^2\pi^2} \sin \sqrt{s^2+n^2\pi^2}} q_n(\sigma)\, d\sigma \\
& + \sum_{n=1}^{\infty} \sin(n\pi x) \int_y^1 \frac{\sinh \sqrt{s^2+n^2\pi^2}\, y \sinh \sqrt{s^2+n^2\pi^2}(1-\sigma)}{\sqrt{s^2+n^2\pi^2} \sinh \sqrt{s^2+n^2\pi^2}} q_n(\sigma)\, d\sigma.
\end{aligned}
\tag{3.51}
$$

A necessary condition for (3.51) to be a F.L.T. is that for each integer n the corresponding entry in (3.51) be a finite Laplace transform. Since the zeros of $\sinh \sqrt{s^2 n^2 \pi^2}$ occur at the points

$$s = \pm i\sqrt{m^2+n^2}\, \pi, \qquad m = 1, 2, \cdots, \tag{3.52}$$

this is possible only when

$$
\begin{aligned}
& (-1)^{m+1} A_0^n(\sqrt{n^2+m^2}\, \pi i) + A_1^n(\sqrt{n^2+m^2}\, \pi i) \\
& \qquad + \frac{(-1)^{m+1}}{m\pi} \int_0^1 (\sin(m\pi\sigma)) q_n(\sigma)\, d\sigma = 0,
\end{aligned}
\tag{3.53}
$$

$m = 1, 2, \cdots$. A similar expression holds when we replace $\sqrt{(n^2+m^2)}\pi i$ by $-\sqrt{(n^2+m^2)}\pi i$. Taking note of (3.50) and setting

$$\alpha_{mn} = \int_0^1 \int_0^1 (\sin(m\pi\sigma))(\sin(n\pi\tau))\phi(\tau, \sigma)\, d\tau\, d\sigma \tag{3.54}$$

and

$$\beta_{mn} = \int_0^1 \int_0^1 \sin(m\pi\sigma) \sin(n\pi\tau)\psi(\tau, \sigma)\, d\tau\, d\sigma \tag{3.55}$$

we can rewrite the conditions (3.53) in the form

$$
\begin{aligned}
& (-1)^{m+1} A_0^n(\pm\sqrt{n^2+m^2}\, \pi i) + A_1^n(\pm\sqrt{n^2+m^2}\, \pi i) \\
& \qquad + \frac{(-1)^{m+1}}{m}(\pm i)\sqrt{n^2+m^2}\, \alpha_{mn} + \frac{(-1)^{m+1}}{m} \beta_{mn} = 0.
\end{aligned}
\tag{3.56}
$$

A choice for $A_0^n(s)$ will be of the form

$$A_0^n(s) = \sum_{m=1}^{\infty} A_0^{mn}(s) \tag{3.57}$$

where if $\sqrt{n^2+m^2}$ is an integer

$$A_0^{mn}(s) = \frac{s^2 e^{-2s}(\sinh 2\sqrt{s^2+n^2\pi^2}) b_{mn}}{\sqrt{s^2+n^2\pi^2}(s^2+(n^2+m^2)\pi^2)}. \tag{3.58}$$

The coefficient b_{mn} is chosen so that

$$A_0^{mn}(\pm i\sqrt{n^2+m^2}\, \pi) = -\frac{(m^2+n^2)\pi}{m} b_{mn} = \frac{-\beta_{mn}}{m}. \tag{3.59}$$

Thus

$$b_{mn}=\frac{\beta_{mn}}{(m^2+n^2)\pi},$$

if $\sqrt{n^2+m^2}$ is an integer.

If $\sqrt{m^2+n^2}$ is not an integer it must be irrational and there is a natural number l such that

$$l-1<\sqrt{m^2+n^2}<l.$$

We define

$$\frac{k}{2}=\frac{l}{\sqrt{m^2+n^2}} \tag{3.60}$$

and observe that the inequality

$$k_m-\frac{2}{\sqrt{m^2+n^2}}<2<k_m \tag{3.61}$$

is satisfied. We then define

$$A_0^{mn}(s)=\frac{s^2\,e^{-k_ms}\sinh{(k_m-2)}\sqrt{s^2+n^2\pi^2}(\sinh 2\sqrt{s^2+n^2\pi^2})b_{mn}}{(s^2+n^2\pi^2)(s^2+(n^2+m^2)\pi^2)}. \tag{3.62}$$

It is not difficult to verify that (3.62) and (3.58) define finite Laplace transforms over intervals of the form $[0, k_m]$, $2<k_m\leqq 2+\sqrt{2}$. In the case of (3.62) we observe that because $\sqrt{m^2+n^2}$ is irrational $(k_m)m$ is also. Thus $\sin{(k_m\pi)}\neq 0$ and the equation

$$A_0(\pm\sqrt{n^2+m}\;\pi i)=\frac{(m^2+n^2)}{m^3\pi}(\sin m\pi k_m)b_{mn}=-\frac{\beta_{mn}}{m}$$

has a solution for b_{mn} which is

$$b_{mn}=-\frac{m^2\pi}{m^2+n^2}\,\frac{\beta_{mn}}{\sin{(m\pi k_m)}}. \tag{3.63}$$

Similarly we define

$$A_1^n(s)=\sum_{m=1}^{\infty}A_1^{mn}(s) \tag{3.64}$$

where if $\sqrt{m^2+n^2}$ is an integer

$$A_1^{mn}(s)=\frac{s\,e^{-2s}\sinh 2\sqrt{s^2+n^2\pi^2}}{\sqrt{s^2+n^2\pi^2}(s^2+(n^2+m^2)\pi^2)}a_{mn}$$

with

$$a_{mn}=(-1)^m\pi_m\alpha_{mn}. \tag{3.65}$$

If $\sqrt{m^2+n^2}$ is not an integer we define

$$A_1^{mn}(s)=\frac{s\,e^{-k_ms}\sinh{(k_m-2)}\sqrt{s^2+n^2\pi^2}(\sinh 2\sqrt{s^2+n^2\pi^2})a_{mn}}{(s^2+n^2\pi^2)(s^2+(n^2+m^2)\pi^2)} \tag{3.66}$$

where

$$a_{mn} = (-1)^{m+1} \frac{m^2\pi^2}{\sin(k_m \Pi_m)} \alpha_{mn}. \tag{3.67}$$

Notice that when $\sqrt{m^2+n^2}$ is not an integer the inequality (3.61) implies that equations (3.63) and (3.67) tend to the quantities $\pm(m/\sqrt{m^2+n^2})\beta_{mn}$ and $\pm m\pi\alpha_{mn}$ respectively. Also observe that

$$A_j^{mn}(\pm i\sqrt{l^2+n^2}) = 0, \qquad j = 0, 1,$$

if $l \neq m$.

As a final remark in this section it should be pointed out that the convergence of $A_0^n(s)$ and $A_1^n(s)$ in equations (3.57) and (3.64) has not been discussed. Since this paper is meant to demonstrate a technique we shall not concern ourselves with this question. Suffice it to say that if $\phi(x, y)$ and $\psi(x, y)$ have finite Fourier expansions i.e. $\{\alpha_{mn}\}$ and $\{\beta_{mn}\}$ contain only a finite number of nonzero terms, then the above constructions will always yield controls which drive (3.41) to the zero state in some time $T \leqq 2 + \sqrt{2}$ (i.e. the maximum possible k_m given by inequality (3.61)).

4. A parabolic problem. Consider the heat equation in one dimension

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial t^2}, \qquad t \geqq 0, \quad 0 \leqq x \leqq 1. \tag{4.1}$$

Assume

$$\begin{aligned} u(x, 0) &= \phi(x), \\ u(0, t) &= a_0(t), \\ u(1, t) &= a_1(t), \end{aligned} \tag{4.2}$$

where ϕ, a_0 and a_1 are integrable over finite intervals.

If $u(x, T) = \psi(x)$ for some $T > 0$ and

$$\int_0^T e^{-st} a_i(t)\, dt = A_i(s), \qquad i = 0, 1,$$

then using the techniques of Example 3.1, the F.L.T. of (4.1)–(4.2) is given by

$$U(\bar{x}, s) = \frac{A_0(s) \sinh \sqrt{s}(1-x) + A_1(s) \sinh \sqrt{s}\, x}{\sinh \sqrt{s}} + \int_0^1 F(x, \sigma, s)[\phi(\sigma) - \psi(\sigma)\, e^{-sT}]\, d\sigma \tag{4.3}$$

where

$$F(x, \sigma, s) = \frac{1}{\sqrt{s} \sinh \sqrt{s}} \sinh \sqrt{s}\, \sigma \sinh \sqrt{s}(1-x) \quad \text{if } 0 \leqq \sigma \leqq x$$

and

$$F(x, \sigma, s) = \frac{1}{s \sinh \sqrt{s}} \sinh \sqrt{s}\, x \sinh \sqrt{s}(1-\sigma) \quad \text{if } \sigma \leqq x \leqq 1. \tag{4.4}$$

Suppose it is desired to drive an initial temperature ϕ to the zero temperature in some time $T>0$. Since the poles of (4.3) are of order one and occur at the points $s=-n^2\pi^2$, $n=1,2,\cdots$, the numerator in (4.3) must satisfy the equations

$$(4.5)\qquad (-1)^{n+1}A_0(-n^2\pi^2)+A_1(-n^2\pi^2)+\frac{(-1)^{n+1}}{n\pi}\int_0^1 (\sin(n\pi\sigma))\phi(\sigma)\,d\sigma=0.$$

Thus it is desired to find entire functions $A_0(s)$ and $A_1(s)$ which are finite Laplace transforms over $[0, T]$ and which also satisfy (4.5). This is not always a practical problem as the following special case shows (see e.g. [3]). Let

$$(4.6)\qquad \phi(x)=\phi_0=\text{constant}\neq 0.$$

For this value of ϕ the equations (4.5) reduce to

$$(4.7)\qquad A_0(-n^2\pi^2)=A_1(-n^2\pi^2),\qquad n \text{ even},$$

and

$$(4.8)\qquad A_0(-n^2\pi^2)+A_1(-n^2\pi^2)=\frac{-2\phi_0}{n^2\pi^2},\qquad n \text{ odd}.$$

Assume

$$(4.9)\qquad A_0(s)=A_1(s),\quad \text{i.e.}\quad a_0(t)=a_1(t),\quad \text{and}\quad |a_0(t)|\leqq 1\quad \text{on } [0, T].$$

If we also assume $a_0(t)$ is piecewise constant with a finite number of switches on $[0, T]$ then it is easily seen that (4.8) can never be satisfied. For if

$$a_0(t)=\sum_{j=1}^{N}\alpha_j\chi_{[t_{j-1},t_j)}(t),\qquad 0=t_0<\cdots<t_n=T,$$

then

$$(4.10)\qquad A_0(s)=\frac{1}{s}\sum_{j=1}^{N}\alpha_j(e^{-st_{j-1}}-e^{-st_j}).$$

Clearly $A_0(s)$ given by (4.10) can never satisfy (4.8).

However if we permit an infinite number of switches then it is possible to bring the temperature to zero in any finite time T. For let

$$a_0(t)=a_1(t)=\sum_{j=1}^{\infty}\alpha_j\chi_{[t_{j-1},t_j)}(t),\qquad 0=t_0<t_1<\cdots<t_n\to T.$$

Then

$$(4.11)\qquad A_0(s)=A_1(s)=\frac{1}{s}\sum_{j=1}^{\infty}\alpha_j(e^{-st_{j-1}}-e^{-st_j})$$

and at the points $-n^2\pi^2$ we have

$$(4.12)\qquad A_0(-n^2\pi^2)=A_1(-n^2\pi^2)=\frac{1}{n^2\pi^2}\sum_{j=1}^{\infty}\alpha_j(e^{n^2\pi^2t_{j-1}}-e^{n^2\pi^2t_j}).$$

Thus (4.8) reduces to the moment problem (see e.g. [3]) of selecting $\{\alpha_j\}$, $|\alpha_j|\leqq 1$ for $j=1,2,\cdots$, and $0=t_0<t_1<\cdots<t_n\to T$ such that

$$(4.13)\qquad \sum_{j=1}^{\infty}\alpha_j(e^{n^2\pi^2t_{j-1}}-e^{n^2\pi^2t_j})=\phi_0.$$

This is a solvable problem (see e.g. [3]) which can be solved for any $T>0$.

REFERENCES

[1] M. ATHANS AND P. L. FALB, *Optimal Control, An Introduction to the Theory and Its Applications*, McGraw-Hill, New York, 1966.

[2] S. BARNETT, *Matrices in Control Theory*, Van Nostrand Reinhold, London, 1971.

[3] A. G. BUTKOVSKIY, *Distributed Control Systems*, American Elsevier, New York, 1969.

[4] G. DOETSCH, *Handbuch der LaPlace Transformation*, Vol. III, Birkhäuser Verlag, Basel, 1956.

[5] N. DUNFORD AND F. T. SCHWARTZ, *Linear Operators*, vol. I, Wiley, New York, 1971.

[6] J. P. LASALLE, *The time optimal control problem*, Contributions to Differential Equations, Vol. V, Princeton Univ. Press, Princeton, NJ, 1960.

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

[8] N. N. PETROVSKY, *Lectures on Partial Differential Equations*, Interscience, New York, 1957.

[9] R. M. GOLDWYN, K. P. SRIRAM AND M. GRAHAM, *Time optimal control of a linear diffusion process*, this Journal, 5 (1967), pp. 295–308.

[10] R. M. REDHEFFER, *Remarks on incompleteness of* $\{e^{i\lambda_n x}\}$, *nonaveraging sets, and entire functions*, Proc. Amer. Math. Soc., 2 (1951), pp. 365–369.

[11] D. L. RUSSELL, *Nonharmonic Fourier series in the control theory of distributed parameter systems*, J. Math. Anal. Appl., 18 (1967), pp. 542–560.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 1, January 1980

0363-0129/80/1801-0002$01.00/0

DEFINITIONS OF ORDER AND JUNCTION CONDITIONS IN SINGULAR OPTIMAL CONTROL PROBLEMS*

R. M. LEWIS†

Abstract. The generalized Legendre–Clebsch higher order tests for optimality of singular arcs in optimal control problems depend upon the orders of the arcs involved. To date three distinct definitions of order have been given but many authors do not distinguish among them. The features of each definition are discussed with special reference to the applicability of the higher order tests and of the conditions at junctions between singular and nonsingular arcs; only in terms of one of the definitions are the junction conditions generally valid. An illustrative example is presented.

Introduction. In optimal control problems an extremal arc or subarc is called singular if it trivially satisfies the Pontryagin minimum principle, that is, a first order control variation on the arc or subarc produces no change in cost, to first order (a statement of the minimum principle and a precise definition of extremality are given in the following section). Higher order conditions are needed to check the optimality of such arcs and two different types of condition have evolved; those based upon higher order control variations (see, for example [1]) and those in which the higher order changes in cost due to the first order control variations are studied (for example, the Gabasov–Jacobson condition [4]).

We are concerned here with the former type and in particular among these, the generalized Legendre–Clebsch necessary conditions. Associated with these is the notion of the order of a singular arc, of which various definitions have been given (some authors define a quantity called degree which is just 2× order). We point out here that these definitions are in need of interpretation and show by means of examples how differing interpretations can yield different values for the order of some singular arcs. This is not in itself a cause for concern; however, by failing to state precisely which interpretation they are considering and, worse, by using different ones alternately, a number of authors have created some confusion about this issue.

The purpose here is to clear this up. In § 1 the class of problems is defined and the phenomenon of singularity is briefly discussed. We then give a naive definition of order: two interpretations of this yield, respectively, the notions of intrinsic and local order. Most of the definitions in the literature correspond to one or the other of these but we do find a third distinct one, combining the features of the other two yet being more suitable with regard to applying the higher order optimality tests.

Conditions at the junctions between singular and nonsingular arcs are discussed in § 3 where it is shown that the theorems of McDanell and Powers are valid only if stated in terms of intrinsic order. Moreover, there exist problems to which none of their theorems are applicable. The section is ended with an example around which much of the work hinges. We conclude by considering the implications of these findings.

1. Problem formulation. We consider the optimal control problem of the form: find the scalar control function $u(\cdot)\in L^1[t_0, t_1]$ which minimizes the cost functional

$$J(u(\cdot))=G(x(t_1))+\int_{t_0}^{t_1} L_0(t, x(t))+L_1(t, x(t))u(t)\,dt \tag{1.1}$$

* Received by the editors November 14, 1977, and in final revised form March 6, 1979.

† School of Mathematics, University of Bath, Bath, England. Now at Department of Electrical Engineering, University of Newcastle, New South Wales, 2308, Australia.

subject to the system equation

$$\dot{x}(t) = f_0(t, x(t)) + f_1(t, x(t))u(t) \quad \text{for almost every } t \in [t_0, t_1] \tag{1.2}$$

and the constraints

$$|u(t)| \leqq K \quad \text{for almost every } t \in [t_0, t_1], \tag{1.3}$$

$$x(t_0) = x_0, \qquad g(x(t_1)) = 0. \tag{1.4}$$

Here $x(\cdot)$ is an absolutely continuous n-vector function of t and $x(t)$ is called the state of the system at time t. The functions G, L_0, L_1, f_0, f_1 and g are assumed to be analytic with $g: \mathbb{R}^n \to \mathbb{R}^m$. $x_0 \in \mathbb{R}^n$ is given and t_0 and t_1 are specified initial and final times. We further assume that a nontrivial set of solutions to (1.2), (1.3) and (1.4) exists and that J has a minimum over this set.

The above problem is linear in the control. We shall deal with nonlinear problems, in which $L_0(t, x) + L_1(t, x)u$ and $f_0(t, x) + f_1(t, x)u$ are replaced by analytic functions $L(t, x, u)$ and $f(t, x, u)$ respectively, separately in § 4. Many authors obtain results for nonlinear problems by considering locally equivalent linearizations (see [1], [2]).

Restricting attention to scalar control problems considerably simplifies notation and results, many of which (the junction conditions for example) are not available in the case of vector controls. Admitting variable end times does not substantially affect what follows.

As usual the Hamiltonian for the problem is defined by:

$$H(t, x, \lambda, u) = \lambda^T f_0(t, x) + L_0(t, x) + [\lambda^T f_1(t, x) + L_1(t, x)]u \tag{1.5}$$

where $\lambda \in \mathbb{R}^n$. The well-known minimum principle provides that a necessary condition for the control-state pair $(u^*(\cdot), x^*(\cdot))$ to be optimal is the existence of an absolutely continuous function $\lambda^*(\cdot)$ (the adjoint) satisfying

$$\dot{\lambda}^*(t) = -H_x^T(t, x^*(t), \lambda^*(t), u^*(t)) \quad \text{a.e. in } [t_0, t_1], \tag{1.6a}$$

$$\lambda^*(t_1) = \nu_0 G_x^T(x^*(t_1)) + \nu^T g_x^T(x^*(t_1)) \tag{1.6b}$$

where ν_0 is a nonnegative scalar and $\nu \in \mathbb{R}^m$. Further,

$$H(t, x^*(t), \lambda^*(t), u^*(t)) \leqq H(t, x^*(t), \lambda^*(t), v) \quad \text{for all } |v| \leqq K \text{ and for almost every } t \in [t_0, t_1]. \tag{1.7}$$

Here $H_x^T(t, x^*(t), \lambda^*(t), u^*(t))$ denotes the partial derivative of H with respect to x, evaluated at $(t, x^*(t), \lambda^*(t), u^*(t))$. T denotes transpose (the derivative is a row vector). The right hand sides of (1.6b) and (1.8) (below) have a similar interpretation.

For any triple $(x(\cdot), \lambda(\cdot), u(\cdot))$ satisfying (1.2), (1.3) and (1.6a) set

$$\phi(t) = \lambda^T f_1(t, x(t)) + L_1(t, x(t)) = (\partial/\partial u)H(t, x(t), \lambda(t), u(t)). \tag{1.8}$$

Expression (1.7) yields two distinct possibilities for optimal controls $u^*(\cdot)$ on subintervals $(t_a, t_b) \subset [t_0, t_1]$, either

$$\phi^*(t) \neq 0, \qquad u^*(t) = -K \operatorname{sgn}(\phi^*(t)), \qquad t \in (t_a, t_b), \tag{1.9}$$

or

$$\phi^*(t) = 0, \qquad t \in (t_a, t_b). \tag{1.10}$$

Any triple $(x^*(\cdot), \lambda^*(\cdot), u^*(\cdot))$ satisfying (1.2), (1.3), (1.4), (1.6) and (1.7) is called an *extremal* for the problem of interest (ϕ^* above denotes evaluation of ϕ along an extremal). An arc of an extremal corresponding to a subinterval (t_a, t_b) is called *nonsingular* if (1.9) holds, otherwise, if (1.10) holds, it is called *singular*. Along nonsingular arcs the minimum principle is strictly satisfied, that is, there exists v with $|v| \leqq K$ giving strict inequality in (1.7). On singular arcs however, the first order necessary condition (minimum principle) is trivially satisfied and we are led to seek higher order tests for optimality.

Remark (i) Along nonsingular arcs (1.9) completely determines $u^*(\cdot)$ whilst on singular arcs the control is usually completely specified by conditions implicit in (1.10), namely $(d^i/dt^i)\phi^*(t) = 0$, $i = 0, 1, 2, \cdots, t \in (t_a, t_b)$. Singularity does not imply indeterminacy of the control but that the first order control variations used to derive the minimum principle produce no first order variations in cost, when applied at $t \in (t_a, t_b)$.

Remark (ii) Singularity is strictly a property of extremals $(x^*(\cdot), \lambda^*(\cdot), u^*(\cdot))$ and not of state-control pairs $(x^*(\cdot), u^*(\cdot))$ since for some such pair there may be more than one adjoint function $\lambda^*(\cdot)$ making the triple extremal. This is a consequence of the nonuniqueness of ν in (1.6b).

Example. Maximize $x_1(1)$ (minimize $-x_1(1)$) subject to

$$\dot{x}_1(t) = x_2(t) + u(t), \qquad \dot{x}_2(t) = x_2^{-1}(t) - u(t),$$

$$x_1(0) = x_2(0) = x_2(1) = 1, \qquad |u(t)| \leqq 1.$$

A candidate for optimality is $(x_1(t), x_2(t), u(t)) = (1 + 2t, 1, 1)$ $0 \leqq t \leqq 1$. (The functions defining the problem are analytic in a neighborhood of the trajectory.) We have

$$\dot{\lambda}_1(t) = 0, \qquad \lambda_1(1) = \nu_0 \leqq 0,$$

$$\dot{\lambda}_2(t) = -\lambda_1(t) + \lambda_2(t) x_2^{-2}(t) = -\lambda_1(t) + \lambda_2(t), \qquad \lambda_2(1) = \nu,$$

that is

$$\lambda_1(t) = \nu_0 \quad \text{and} \quad \lambda_2(t) = (\nu - \nu_0) \exp(t - 1) + \nu_0;$$

hence

$$\phi(t) = \lambda_1(t) - \lambda_2(t) = (\nu_0 - \nu) \exp(t - 1).$$

Therefore $\phi(t) = 0$ if $\nu = \nu_0$ whilst if $\nu > \nu_0$, $\phi(t) < 0$ and $u(t) = -\text{sgn}\, \phi(t) = 1$.

Both singular and nonsingular extremals corresponding to $(x_1(\cdot), x_2(\cdot), u(\cdot))$ are possible.

2. The order of singular extremal arcs. We begin this section by defining the order of a singular extremal arc. The value obtained in a particular problem is seen to depend upon the interpretation given to the definition and this leads us to two different notions of order. Four examples from the literature are then investigated to determine which notions their authors had in mind and in so doing we discover a third independent one.

DEFINITION 2.1. The *order* of a singular extremal arc on (t_a, t_b) is that integer q such that $(d^{2q}/dt^{2q})[H_u]$ is the lowest order total derivative of H_u in which u appears explicitly. $(H_u = (\partial/\partial u)H = L_1 + \lambda^T f_1)$.

As $H_u = H_u(t, x, \lambda)$, total derivatives of H_u are defined only when $x(\cdot)$ and $\lambda(\cdot)$ are specified as functions of t. Then, strictly, $H_u = H_u(t)$ and detection of the explicit appearance of u is impossible. It is therefore necessary to interpret the definition.

Interpretation 2.2. To determine the order of a singular extremal arc, form the derivatives of H_u as follows:

$$\begin{aligned}(d/dt)H_u &= H_{ut} + H_{ux}^T \dot{x} + H_{u\lambda}^T \dot{\lambda} \\ &= H_{ut} + H_{ux}^T [f_0(t, x) + f_1(t, x)u] - H_{u\lambda}^T H_x^T,\end{aligned}$$

i.e. substitute the functional forms given by the right hand sides of (1.2) and (1.6a) for $\dot{x}$ and λ respectively. These forms hold along all extremal arcs; hence the above expression for $(d/dt)H_u$ is valid along all extremal arcs and is explicitly a function of (t, x, λ, u), $M(t, x, \lambda, u)$, say. It is easy to show directly that $M = M(t, x, \lambda)$ and hence $(\partial/\partial u)(d/dt)H_u = 0$.

Continuing, form

$$(d^2/dt^2)H_u = M_t + M_x^T[f_0(t, x) + f_1(t, x)u] - M_\lambda^T H_x^T = N(t, x, \lambda, u).$$

This expression is also valid along all extremal arcs. If N depends explicitly upon u, i.e. if for some $(t, x, \lambda) \in \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n$ $(\partial/\partial u)N(t, x, \lambda, u) \neq 0$, then the order of the singular arc is $q = 2/2 = 1$; otherwise the above process is continued until a total derivative of H_u is found which is dependent upon u. If no such derivative exists, set $q = \infty$.

Note (a) It is implicit in Definition 2.1–Interpretation 2.2 that the first appearance of u is in an even order derivative of H_u. This is proved by Robbins [2] whose definition of order is the same as Definition 2.1–Interpretation 2.2 (see below).

Note (b) Though we set out to determine the order of a *particular* singular extremal arc, we note from Interpretation 2.2 that the number q arrived at there is a property of all the extremal arcs in a given problem. This motivates:

DEFINITION 2.3. The *intrinsic order* of an optimal control problem in which the control appears linearly is the least integer q such that $(d^{2q}/dt^{2q})H_u$ depends explicitly upon u, with the Interpretation 2.2.

Note (i) The intrinsic order of a problem linear in the control is always greater than or equal to one.

Note (ii) For problems linear in the state x as well as in the control, $q = \infty$.

Note (iii) u appears linearly in (d^{2q}/dt^{2q}) H_u, that is $(d^{2q}/dt^{2q})H_u = A(t, x, \lambda) + B(t, x, \lambda)u$ where B is not the null function. The importance of this is discussed in § 3.

A necessary condition for optimality of singular arcs is:

THEOREM 2.4. *Suppose* $(x^*(\cdot), \lambda^*(\cdot), u^*(\cdot))$ *is a normal extremal for a problem of intrinsic order q, with a singular arc on the interval* (t_a, t_b), *and that* $|u^*(t)| < K$ *for* $t \in (t_a, t_b)$. *Then for the extremal to be optimal it is necessary that*

$$(2.5) \qquad (-1)^q\{(\partial/\partial u)(d^{2q}/dt^{2q})H_u\}^* \geqq 0$$

for all points $t \in (t_a, t_b)$ *at which* u^* *is analytic.*

* denotes evaluation along the extremal, i.e. the left hand side of (2.5) is $(-1)^q B(t, x^*(t), \lambda^*(t))$, where B is as in Note (iii) above.

Normality guarantees that the terminal constraints can be satisfied by varied trajectories and is equivalent to uniqueness of the $\lambda^*(\cdot)$ making $(x^*(\cdot), \lambda^*(\cdot), u^*(\cdot))$ extremal [1]. Nonnormal extremals are included in a modified version of Theorem 2.4 at the end of this section.

A proof of Theorem 2.4 can be found in [2]. Condition (2.5) is known as the generalized Legendre–Clebsch (GLC) condition. By the strengthened GLC condition we mean that strict inequality holds. Of course it is possible to have $B(t, x^*(t), \lambda^*(t)) = 0$ for $t \in (t_c, t_d) \subset (t_a, t_b)$ even though B is not the null function. Then on (t_c, t_d) the GLC condition is trivially satisfied and does not provide a test for optimality of this subarc. To obtain a test the following is needed:

DEFINITION 2.6. The *local order* of an extremal on the interval (t_c, t_d) is the least integer p such that

$$\{(\partial/\partial u)(d^{2p}/dt^{2p})H_u\}^* \neq 0 \quad \text{for all } t \in (t_c, t_d).$$

Interpretation. The derivatives up to order $2q$ (q = intrinsic order) are formed as before, yielding

$$(d^{2q}/dt^{2q})H_u = A(t, x, \lambda) + B(t, x, \lambda)u$$

where B is not the null function. If however $B(t, x^*(t), \lambda^*(t)) = 0$ on (t_c, t_d) we then form

$$\begin{aligned}(d^{2q+1}/dt^{2q+1})H_u &= A_t + B_t u + [A_x^T + B_x^T u][f_0(t, x) + f_1(t, x)u]\\ &\quad + [A_\lambda^T + B_\lambda^T u][-H_x^T] + B\dot{u}\\ &= P(t, x, \lambda, u, \dot{u}).\end{aligned}$$

It can be shown that $(\partial/\partial u)(d^{2+1}/dt^{2q+1})H_u = (\partial/\partial u)P(t, x, \lambda, u, \dot{u})$ is zero along $(x^*(t), \lambda^*(t), u^*(t))$ for $t \in (t_c, t_d)$ because $B = 0$ there [7]. Continuing, we have

$$(d^{2q+2}/dt^{2q+2})H_u = P_t + P_x^T[f_0 + f_1 u] - P_\lambda^T H_x^T + \left(\frac{\partial}{\partial u}P\right)\dot{u} + B\ddot{u} = Q(t, x, \lambda, u, \dot{u}, \ddot{u}).$$

We now evaluate $(\partial/\partial u)(d^{2q+2}/dt^{2q+2})H_u = (\partial/\partial u)Q(t, x, \lambda, u, \dot{u}, \ddot{u})$ along $(x^*(t), \lambda^*(t), u^*(t))$ for $t \in (t_c, t_d)$. If it is nonzero there, then the local order of the singular extremal $(x^*(t), \lambda^*(t), u^*(t))$ is $p = (2q+2)/2 = q+1$. Otherwise, continue the above process until a total derivative of order $2q+2r$, $r > 1$, is found such that $(\partial/\partial u)(d^{2q+2r}/dt^{2q+2r})H_u$ is not zero along the extremal. If no such derivative exists, set $p = \infty$.

Note (i) To obtain $(d^{2q+2r}/dt^{2q+2r})H_u$ we appear to require $u^*(t)$ to be $2r$ times differentiable on (t_c, t_d). As the value of r is not known a priori we therefore assume that $u^*(\cdot)$ is piecewise analytic. This implies that the interval $[t_0, t_1]$ divides into at most a finite number of subintervals on which the local order of the extremal is different. That the local order can change along the extremal is shown by the example in § 3.

Note (ii) Note that the coefficients of $\dot{u}$, $\ddot{u}$ etc. are zero along the extremal concerned. This may enable us to extend Definition 2.6 to extremals corresponding to nonpiecewise analytic controls.

Note (iii) It is evident that if the local order p is greater than the intrinsic order q, then u no longer appears linearly in $(d^{2p}/dt^{2p})H_u$. Indeed $(d^{2p}/dt^{2p})H_u$ is generally a polynomial of degree $2(p-q)+1$ in u. (The function Q defined above is cubic in u.)

Note (iv) Proof that the first nonzero term $\{(\partial/\partial u)(d^k/dt^k)H_u\}^*$ occurs for k even is given in [7].

In terms of local order, the necessary conditions for optimality are the same.

THEOREM 2.7. *If $(x^*(\cdot), \lambda^*(\cdot), u^*(\cdot))$ is a normal extremal with a singular arc of local order p on (t_c, t_d) and $|u^*(t)| < K$ for $t \in (t_c, t_d)$ then a necessary condition for the extremal to be optimal is*

$$(2.8)\qquad (-1)^p\{(\partial/\partial u)(d^{2p}/dt^{2p})H_u\}^* \geqq 0$$

for all $t \in (t_c, t_d)$ at which $u^*(\cdot)$ is analytic.

Using local order, for a normal extremal the higher order tests for optimality are never trivially satisfied unless $p = \infty$; no further definitions of order nor tests for optimality of GLC type are possible. Of course satisfaction of the strengthened GLC does not guarantee optimality, as an example due to Jacobson and Bell [4, pp. 94–96], shows.

We conclude, provisionally, that there are two useful notions of order, the intrinsic order of a problem and the local order of a particular extremal subarc and that these are not the same. The literature on singular control problems abounds in different statements of a definition of order. The stress here is on "statements" since many authors do not give interpretations of their definitions and in fact use the intrinsic order on some and local order on other occasions. Some of these definitions are now examined, in order of publication.

Note. Some authors quoted below refer to optimal arcs instead of extremals. Their definitions actually apply to pairs (x^*, u^*) satisfying (1.6) and (1.7), for which the extremal (x^*, λ^*, u^*) is unique.

Among the earliest derivations of the generalized Legendre–Clebsch conditions is that of Kopp and Moyer [3] who treat general, nonlinear problems and would therefore be expected to use local order (see § 4). Actually, they do not explicitly define the order of a singular arc but state the GLC as:

$$(-1)^k(\partial/\partial u)[(d^{2k}/dt^{2k})(\partial H/\partial u)] \geqq 0 \qquad [3, (24)].$$

The left hand side is evaluated as above, along the singular arc of interest. k is to be found as follows [3, p. 1443]: "If the inequality is met marginally (equality) for the first necessary condition, in which case the test is inconclusive on the nature of the extremal arc, the second test is applied and so on." The first test refers to [3, (24)] with $k = 1$, the second with $k = 2$ etc. This confirms that local order is being used as it is the same procedure as used to determine local order. The authors also state the strengthened GLC condition as necessary for optimality [3, (A12)]; this is true only when k is local order.

Robbins, by contrast, gives a very detailed definition of degree (= 2× order). In [2], linear problems with vector valued controls of dimension n_c are considered. r control variables are assumed singular, that is, correspond to components of H_u which are zero. The definition of degree is then [2, p. 365] (in [2] λ is written as p):

> The condition $H_u \equiv 0$ is independent of u (as already noted) and hence gives a relation among x, p and t. By use of the equations $\dot{x} = H_p$ and $\dot{p} = -H_x$, the other conditions given in (17) can successively be reduced to similar relations among these variables, until sooner or later (in general) a relation will be encountered which explicitly involves u. Let Q_m denote the $r \times r$ matrix whose elements are
>
> $$Q_{mij} = \frac{\partial}{\partial u}\left[\left(\frac{d^m}{dt^m}\right)\frac{\partial H}{\partial u}\right] \qquad (18)$$
>
> and let M denote the smallest value of m for which Q_m has at least one nonzero element. In general, M is a function of x, p and t but to simplify the discussion we shall assume that M is constant in the neighborhood of the extremal arc of interest and make a similar assumption for the rank of Q_M. These assumptions exclude certain atypical cases in which the extremal arc coincides with a line or surface in the x, p or t space where M is greater, or the rank of Q_M is less, than at neighboring points. (I am indebted . . . to my attention.) These atypical cases will be discussed in section 8. In all other cases, M is the first value of m for which the elements of Q_m do not all vanish identically in the region of interest.

Defined in this way, degree equals twice intrinsic order. Robbins "atypical cases" are those arcs along which local order is greater than intrinsic order. For the application of the GLC condition to these cases, he specifies a procedure equivalent to the use of the local order of the arc [2, p. 272].

The precision of Robbins definition is lacking in many subsequent to it. Typical of these (and important since it appears in the first text on singular control) is (vector valued controls are considered: $[t_1, t_2]$ is a subinterval of $[t_e, t_f]$, the interval of interest; the word optimal should be replaced by extremal):

"Let u_k be an optimal singular element of the control vector u on the interval $[t_1, t_2]$ which appears linearly in the Hamiltonian. Let the $2q$th time derivative of H_{u_k} be the lowest order total derivative in which u_k appears explicitly with a coefficient which is not identically zero on $[t_1, t_2]$. Then the integer q is called the order of the singular arc. The control variable u_k is referred to as a singular control." [4, p. 4 definition (1.1).].

This is difficult to interpret satisfactorily because of the simultaneous demands "appears explicitly" and "not identically zero on $[t_1, t_2]$". The first phrase indicates that intrinsic order is meant, that is, explicit dependence of $(d^{2q}/dt^{2q})H_{u_k}$ on u_k, considered as a function of (t, x, λ, u). The second phrase confuses this by requiring that along the arc of interest the coefficient be not identically zero, as a function of t along the arc: unless one is to understand that the "atypical cases" of Robbins are excluded from the definition it is not sensible. On the contrary, the authors do not seem to intend intrinsic order for they write [4, p. 63]:

"In the following derivation of the GLC a sequence of special control variations will be constructed which in turn will generate a sequence of necessary conditions. Should the first condition of this sequence be trivially satisfied, then the second condition is tested and so on until new information is obtained." This sequence of testing implies consideration of local order.

It must be remarked again that from the point of view of applying the GLC tests it does not matter which order is used, except that in terms of local order, the GLC condition holds strongly except on a null set of points. However, this robustness of the GLC condition is not shared by other necessary conditions (see § 3).

Recently, Krener has given a hybrid definition of degree (= twice order) which obviates the need for normality in the GLC condition, [1, p. 278 et seq.], (a nonlinear problem is considered):

DEFINITION 2.9. Suppose $u^*(\cdot)$ and $x^*(\cdot)$ are a singular extremal control and trajectory on $[t_a, t_b]$. The pair is singular of degree m on this interval if m is the smallest integer for which there exists $\lambda(\cdot)$ satisfying the adjoint differential equation

$$\lambda(t) = -H_x^T(t, x^*(t), \lambda(t), u^*(t)),$$

the necessary conditions $H(t, x^*(t), \lambda(t), u^*(t)) = 0$, $(d^k/dt^k)H_u(t, x^*(t), \lambda(t), u^*(t)) = 0$, $k = 0, \cdots, \infty$, and $(\partial/\partial u)(d^m/dt^m)H_u(t, x^*(t), \lambda(t), u^*(t))$ is not identically zero on $[t_a, t_b]$.

The adjoint $\lambda(\cdot)$ in this definition need not be the same as the $\lambda^*(\cdot)$ forming the extremal in that it need not satisfy the boundary conditions (1.6b).

Definition 2.9 has something of the flavor of the intrinsic order definition whilst having local statement and $q \leqq m/2 \leqq p$ where q is intrinsic and p, local order. In example 5.1 of [1], $m/2 = q < p$ while in the example in the following section $q < p = m/2$, so the above definition is not equivalent to either the local or the intrinsic one. It should not be difficult to combine the features of these two examples to produce a problem for which $q < m/2 < p$.

The GLC condition now takes the form:

THEOREM 2.10. *Assume that $u^*(\cdot)$ and $x^*(\cdot)$ are extremal and singular of degree m on $[t_a, t_b]$ and that $|u^*(t) < K$ for $t \in [t_a, t_b]$. Then m is even and if $u^*(\cdot)$ is optimal there exists a $\lambda(\cdot)$ such that $(x^*(\cdot), \lambda(\cdot), u^*(\cdot))$ is extremal and*

$$(-1)^{m/2}(\partial/\partial u)(d^m/dt^m)H_u(t, x^*(t), \lambda(t), u^*(t)) \leqq 0.$$

For normal problems, the $\lambda(\cdot)$ above must be $\lambda^*(\cdot)$ but in nonnormal cases the GLC condition need not hold for $\lambda^*(\cdot)$.

The major disadvantage of Definition 2.9 is that one may need to compute a large number of multipliers $\lambda(\cdot)$ to determine m, unless either (i) the multiplier satisfying the conditions of Definition 2.9 is unique, in which case $m = p$ or (ii) a multiplier exists for which

$$(\partial/\partial u)(d^{2q}/dt^{2q})H_u(t, x^*(t), \lambda(t), u^*(t)) \neq 0, \quad \text{whence } m = q.$$

Against this there is the advantage of including nonnormal extremals in Theorem 2.10.

3. Junction conditions. The generalized Legendre–Clebsch condition provides a test for optimality along the singular subarcs of an extremal. At the junction between singular and nonsingular subarcs, further tests are required. In [5], McDanell and Powers give the first general junction conditions for problems linear in the control.

THEOREM 3.1. *Let t_c be a point at which singular and nonsingular subarcs of an optimal control $u(\cdot)$ are joined and let q be the order of the singular arc. Suppose the strengthened GLC condition is satisfied at t_c, i.e., $(-1)^q(\partial/\partial u)(d^{2q}/dt^{2q})H_u > 0$ and assume that the control is piecewise analytic in a neighborhood of t_c. Let $u^{(r)}$ $(r \geqq 0)$ be the lowest order derivative of u which is discontinuous at t_c. Then $q + r$ is an odd integer.*

This is Theorem 1 of [5]. The control does not actually have to be optimal but has to satisfy the minimum principle and the GLC condition.

The discussion in § 2 leads one to the question, for which definitions of order is this theorem true? The definition of order given in [5], (definition 3), is virtually identical to the confusing one offered in [4] except that it is restricted to linear problems with scalar control. Local order is therefore indicated. However, the proof of the theorem given in [5] depends upon $(d^{2q}/dt^{2q})H_u$ being linear in u which is generally true only if q is intrinsic order. Failure of the theorem when local order is used is shown in the example below.

Accepting that intrinsic order is employed, the theorem is inapplicable when the GLC condition holds trivially at t_c. Then "To treat this case note from Definition 3 that for a qth order singular arc the GLC expression $(\partial/\partial u)(d^{2q}/dt^{2q})H_u$ (i.e. β) cannot be identically zero on the singular arc. Therefore in view of our analyticity assumptions a derivative of some order must be nonzero at the junction point t_c even if $\beta(t_c) = 0$. This then leads to the following theorem, $\cdots$" [5, pp. 166–167]. This is incorrect, or at least incompatible with q being intrinsic order, for the example below shows that it is in fact possible for $(\partial/\partial u)(d^{2q}/dt^{2q})H_u$ to be identically zero on the singular arc. We see then that there is a class of problems whose junction point behavior is not specified by the theorems in [5].

Example. Consider the problem of minimizing

$$J(u) = \int_{t_0}^{t_1} (x_1 - \tfrac{1}{2})^2 \, dt \tag{3.1}$$

subject to

$$\begin{aligned} \dot{x}_1 &= x_2 u, & x_1(t_0) &= \xi_1 \neq \tfrac{1}{2}, \\ \dot{x}_2 &= u - x_1, & x_2(t_0) &= \xi_2 \neq 0; \end{aligned} \tag{3.2}$$

$$|u| \leqq 1. \tag{3.3}$$

t_0, t_1, ξ_1 and ξ_2 are fixed but remain unspecified for the present. The Hamiltonian, the multiplier equations and the switching function are given by

$$H = \lambda_1 x_2 u + \lambda_2(u - x_1) + (x_1 - \tfrac{1}{2})^2, \tag{3.4}$$

$$\begin{aligned} \dot{\lambda} &= \lambda_2 - 2(x_1 - \tfrac{1}{2}), & \lambda_1(t_1) &= 0,\\ \dot{\lambda}_2 &= -\lambda_1 u, & \lambda_2(t_1) &= 0; \end{aligned} \tag{3.5}$$

$$\phi = H_u = (\lambda_1 x_2 + \lambda_2). \tag{3.6}$$

Arcs where $\phi \neq 0$ are nonsingular and extremal controls for such arcs are given by $u^* = -\mathrm{sgn}\,(\phi)$. The extremal arc

$$x_1(t) = \tfrac{1}{2}, \quad x_2(t) = 0, \quad u(t) = \tfrac{1}{2}, \quad \lambda_1(t) = \lambda_2(t) = 0 \tag{3.7}$$

is singular. The increment in cost along (3.7) is zero and the necessary conditions indicate that the optimal solution from any initial point $(\xi_1, \xi_2) \neq (\frac{1}{2}, 0)$ comprises a nonsingular arc from (ξ_1, ξ_2) reaching $(\frac{1}{2}, 0)$ at time t_c, followed by the singular arc (3.7) for $t_c \leqq t \leqq t_1$. It will be shown that ξ_1, ξ_2, t_0 and t_1 can be chosen so that such a trajectory is extremal and has a piecewise analytic control.

The problem is autonomous so choose $t_c = 0$ with $t_0 < 0 < t_1$. The singular arc is on the interval $(0, t_1]$, the nonsingular on $[t_0, 0)$. On $[t_0, 0)$ let us attempt to construct an extremal with $\phi = \lambda_1 x_2 + \lambda_2 < 0$; then $u = -\mathrm{sgn}\,\phi = 1$ and (3.2) and (3.5) become:

$$\begin{aligned} \dot{x}_1 &= x_2, & \dot{x}_2 &= 1 - x_1,\\ \dot{\lambda}_1 &= \lambda_2 - 2(x_1 - \tfrac{1}{2}), & \dot{\lambda}_2 &= -\lambda_1 \end{aligned}$$

or

$$(d^4/dt^4 + 2d^2/dt^2 + 1)\lambda_2 = 1. \tag{3.8}$$

Thus $\lambda_2(t) = A \sin t + Bt \sin t + C \cos t + Dt \cos t + 1$, $t \in [t_0, 0)$. From the boundary conditions (3.7) at $t_c = 0$,

$$\lambda_2(t) = 1 - \cos t - (t/2) \sin t$$

whence

$$\begin{aligned} \lambda_1(t) &= (t/2) \cos t - (1/2) \sin t,\\ x_2(t) &= (1/2) \sin t \quad \text{and} \quad x_1(t) = 1 - (1/2) \cos t,\\ \phi(t) &= \lambda_1(t) x_2(t) + \lambda_2(t)\\ &= 1 - \cos t - (t/2) \sin t + (1/2) \sin t((t/2) \cos t - \tfrac{1}{2} \sin t)\\ &= 1 - \cos t - (t/2) \sin t + (t/8) \sin (2t) + (1/8)(\cos (2t) - 1),\\ (d/dt)\phi(t) &= (1/2)(\sin t - t \cos t) - (1/8)(\sin (2t) - 2t \cos (2t)). \end{aligned} \tag{3.9}$$

Now

$$\begin{aligned} \sin r - r \cos r &= (r - r^3/3! + r^5/5! - \cdots) - r(1 - r^2/2! + r/4! - \cdots)\\ &= r^3/3 - r^5/30 + \cdots; \end{aligned}$$

hence

$$\begin{aligned} (d/dt)\phi(t) &= (1/2)(t^3/3 - t^5/30 + \cdots) - (1/8)(8t^3/3 - 32t^5/30 + \cdots)\\ &= -t^3/6 + 7t^5/60 - \cdots. \end{aligned}$$

For t negative but sufficiently near zero then $(d/dt)\phi(t) > 0$ and $\phi(0) = 0$ implies $\phi(t) < 0$. We conclude that there is an interval $[t_0, 0)$ on which $\phi(t) < 0$ and therefore the control $u(t) = 1$ and trajectory $x_1(t) = 1 - (1/2)\cos t$, $x_2 = (1/2)\sin t$ are extremal. For such a t_0, taking $\xi_1 = 1 - (1/2)\cos t_0$, $\xi_2 = (1/2)\sin t_0$ and any $t_1 > 0$ the control problem admits an extremal of the stated form.

Comments. The above detail is necessitated by the fact that problems very similar to the one given do not admit extremals with piecewise analytic controls. Indeed, if we replace the bilinear form $\dot{x}_1 = x_2 u$ above by $\dot{x}_1 = x_2$, the singular arc remains the same. Assuming $\phi < 0$ on the nonsingular arc, u, x_1, x_2, λ_1 and λ_2 are as above but

$$\begin{aligned}\phi = \lambda_2 = 1 - \cos t - (t/2)\sin t &= 1 - (1 - t^2/2! + t^4/4! - \cdots)\\ &\qquad - (t/2)(t - t^3/3! + t^5/5! - \cdots),\\ \phi = t^4/12 - O(t^5),&\end{aligned}$$

i.e. $\phi > 0$ near $t = 0$, a contradiction and it transpires that the switching function switches infinitely often in a neighborhood of $t = 0$. The corresponding control is measurable but not piecewise analytic; cf. [5], [6].

Let us determine the intrinsic order of the problem and the local order and degree of the singular arc.

$$\begin{aligned}H_u &= \lambda_1 x_2 + \lambda_2,\\ (d/dt)H_u &= (\lambda_2 - 2(x_1 - \tfrac{1}{2}))x_2 + \lambda_1(u - x_1) - \lambda_1 u\\ &= (\lambda_2 - 2(x_1 - \tfrac{1}{2}))x_2 - \lambda_1 x_1,\\ (d^2/dt^2)H_u &= (-\lambda_1 u - 2x_2 u)x_2 + (\lambda_2 - 2(x_1 - \tfrac{1}{2}))(u - x_1) - (\lambda_2 - 2(x_1 - \tfrac{1}{2}))x_1 - \lambda_1 x_2 u\\ &= -2(\lambda_2 - 2(x_1 - \tfrac{1}{2}))x_1 + (\lambda_2 - 2(x_1 - \tfrac{1}{2}) - 2(\lambda_1 + x_2)x_2)u.\end{aligned}$$

Hence $(\partial/\partial u)(d^2/dt^2)H_u = \lambda_2 - 2(x_1 - \tfrac{1}{2}) - 2(\lambda_1 + x_2)x_2$ and the intrinsic order is $q = 2/2 = 1$.

However, along the singular arc $(\partial/\partial u)(d^2/dt^2)H_u = 0$ so the local order is greater than 1. Since $(d^k/dt^k)u(t) = 0$, $k = 1, 2, \cdots$, along both arcs of the solution, these terms are neglected in higher order derivatives of H_u. We find:

$$(d^3/dt^3)H_u = ((4(\lambda_1 + 2x_2)x_1 - 4(\lambda_2 - 2(x_1 - \tfrac{1}{2}))x_2) + (-3\lambda_1 - 6x_2)u)u$$

and as expected $(\partial/\partial u)(d^3/dt^3)H_u = 0$ along the singular arc.

$$\begin{aligned}(d^4/dt^4)H_u = ((8\lambda_2 x_1 - 16(x_1 - \tfrac{1}{2})x_1 - 8x_1^2)&\\ + (14x_1 + 8(\lambda_1 + 2x_2)x_2 - 7(\lambda_2 - 2(x_1 - \tfrac{1}{2})))u + (-6)u^2)u.&\end{aligned}$$

Along the singular arc

$$(\partial/\partial u)(d^4/dt^4)H_u = -8x_1^2 + 2(14x_1)u + 3(-6)u^2 = 2(1/4) = \tfrac{1}{2}$$

as $u = x_1 = \tfrac{1}{2}$.

The local order is therefore $4/2 = 2$. Note that $(-1)^2(\partial/\partial u)(d^4/dt^4)H_u = \tfrac{1}{2} > 0$, i.e. the strengthened GLC condition is satisfied along the singular arc.

The adjoint multipliers associated with the singular arc are unique: we require $H_u = \lambda_1 x_2 + \lambda_2 = \lambda_2 = 0$ and $\dot{\lambda}_2 = -\lambda_1/2$ which implies $\lambda_1 = \lambda_2 = 0$. Therefore the degree of the singular arc, as defined by Definition 2.9, is 4, twice the local order.

Now, in terms of local order p, this example satisfies the conditions of the junction theorem but $p=2$ and $r=0$ (the control is discontinuous at $t=0$) so $p-r=2$, an even integer. This contradiction shows that the theorem is invalid in terms of p.

4. Problems nonlinear in the control. Singularity of first order necessary conditions can also occur in problems nonlinear in the control, where the functions $L_0(t, x)+L_1(t, x)u$ and $f_0(t, x)+f_1(t, x)u$ in (1.1) and (1.2) are replaced by analytic functions $L(t, x, u)$ and $f(t, x, u)$ respectively. The minimum principle takes the same form as in § 1, with the Hamiltonian defined by

$$H(t, x, \lambda, u)=\lambda^T f(t, x, u)+L(t, x, u). \tag{4.1}$$

An extremal arc is singular if

$$\frac{\partial^i}{\partial u^i} H(t, x, \lambda, u)=0 \quad \text{along the arc,} \quad i=1,2,\cdots. \tag{4.1}$$

We note that if a problem is strictly nonlinear in the control then there exist (t, x, λ, u) such that $(\partial^2/\partial u^2)H(t, x, \lambda, u)\neq 0$ and therefore the intrinsic order of such a problem must be zero. Hence only Definitions 2.6 (local order) and 2.9 are useful here. The interpretations of these are the same as for the linear case. Theorems 2.7 and 2.10 remain true (see [3] and [1]).

An interesting alternative procedure is given by Robbins [2]. He shows that when replacing the nonlinear Hamiltonian $H(t, x, \lambda, u)$ by $\bar{H}(t, x, \lambda, u)=H(t, x, \lambda, u^*(t))+(u-u^*(t))H_u(t, x, \lambda, u^*(t))$ the effect of a second order control variation along the extremal $(x^*(\cdot), \lambda^*(\cdot), u^*(\cdot))$ is the same; hence $\bar{H}$ can be used in place of H in the GLC test for optimality. Using $\bar{H}$ the degree of the extremal can be determined as in the linear case but this is not intrinsic as $\bar{H}$ depends on the extremal; moreover it may not correspond to the local order either as $\bar{H}$ may be an "atypical case". It therefore seems better to use local order ab initio.

In § 3 we noted that the validity of the junction condition, Theorem 3.1, depends upon the linearity of $(d^{2q}/dt^{2q})H_u$ with respect to u (q as in the statement of Theorem 3.1). In nonlinear problems this will not hold and neither therefore will the junction conditions.

5. Conclusions. Loosely worded definitions of order have led to some confusion and incorrect claims about the nature of singular control problems. Several important ones have been studied here and it has been shown that their authors intended one of the basic interpretations (a third, different definition is not yet widely used).

The junction conditions of McDanell and Powers are valid only in terms of the weaker form, i.e. intrinsic order, and it appears that there are problems for which no junction condition can be given. The frequency of occurrence of such examples is of interest since various authors have either implied they do not exist [5] or called them atypical cases [2]. The example given is a two dimensional bilinear one, not reducible to a lower dimensional canonical form and not having special boundary conditions. Until at least the class of bilinear problems has been exhaustively studied, it might be advisable to refrain from any claims involving genericity.

The first results in this direction are given in [8], where the time optimal behavior of systems linear in the control is considered. The functions $f_0(x, t)$ and $f_1(x, t)$ are required to be C^∞ and the set of control systems is given a Whitney topology. For systems of dimension 2 it is shown that singular extremals cannot be generic, i.e. given a system which admits a singular extremal and any open neighborhood of that system, there is a system in this neighborhood which does not admit singular extremals. However, singular extremals can be generic among systems of dimension 3 or greater. With regard to order, when dim = 3 only 1st order extremals can be generic. It is tempting to suppose that with increasing dimension higher order extremals can be generic but this has not yet been proved. Then, since the order used in [8] is the local variety, we might be able to make useful statements regarding the "atypical cases."

The problems considered here have all involved a scalar control variable $u(\cdot)$. With vector controls $u(\cdot)=[u_1(\cdot), \cdots, u_m(\cdot)]^T$, problems can be singular of rank r for any $1 \leqq r \leqq m$, by which we mean, in the linear case, that $(\partial/\partial u_i)H=0$, on the extremal arc, for r indices i. Clearly the arc can have different order with respect to each control u_i, whatever definitions of order are used, and this complicates application of the optimality tests. The simplest case, when for each u_i the local and intrinsic orders coincide, is dealt with in [2]. Junction conditions for vector control problems are not yet available.

Acknowledgments. I wish to thank Dr. D. Bell for discussions which led to the discovery of the example in § 4, and anonymous referees for suggesting an improved presentation of this material.

REFERENCES

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

[2] H. M. ROBBINS, *A generalized Legendre–Clebsch condition for the singular cases of optimal control*, IBM J. Res. Develop., 11 (1967), pp. 361–372.

[3] R. E. KOPP AND H. G. MOYER, *Necessary Conditions for Singular Extremals*, AIAA Journal, 3 (1965), pp. 1439–1444.

[4] D. J. BELL AND D. H. JACOBSON, *Singular Optimal Control Problems*, Academic Press, New York, 1975.

[5]. J. P. MCDANELL AND W. F. POWERS, *Necessary conditions for joining optimal singular and non-singular subarcs*, this Journal, 9 (1971), pp. 161–173.

[6] A. T. FULLER, *Study of an optimum nonlinear control system*, J. Electronics Control, 15 (1963), pp. 63–71.

[7] H. J. KELLEY, R. E. KOPP AND H. G. MOYER, *Singular extremals*, Topics in Optimization, G. Leitmann, ed., Academic Press, New York, 1967, Chap. 3.

[8] D. REBHUHN, *On the stability of the existence of singular controls under perturbation of the control system*, this Journal, 16 (1978), pp. 463–472.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 1, January 1980

0363-0129/80/1801-0003$01.00/0

REPRESENTATION AND APPROXIMATION OF NONCOOPERATIVE SEQUENTIAL GAMES*

WARD WHITT†

Abstract. Noncooperative sequential games, including the noncooperative stochastic game of Rogers (1969) and Sobel (1971), are investigated in the monotone contraction operator framework of Denardo (1967). Sufficient conditions are determined for the existence of equilibrium points in this setting. Techniques for comparing and approximating dynamic programs previously developed by the author are then applied to these sequential games, yielding conditions for the existence of ε-equilibrium points.

1. Introduction and Summary. It is now widely recognized in economics and several other fields that there is a need for mathematical models which can represent the behavior of several competing decision makers interacting over time, possibly under uncertainty. A natural model for this purpose is the sequential game, which combines the dynamic properties of dynamic programming with the competitive properties of game theory. The purpose of the present paper is to provide a general framework for analyzing and approximating a large class of noncooperative sequential games. We focus on noncooperative equilibrium points in the sense of Nash (1951), i.e., we look for policies or strategies for all players with the property that no single player acting alone can do better by changing. We consider the important questions of existence and approximation. Approximation seems particularly worth studying because it opens the way to computation and existence proofs for larger games.

The framework we suggest is the monotone contraction operator model introduced by Denardo (1967). He showed that this model encompasses the two-person zero-sum discounted stochastic game of Shapley (1953) plus many dynamic programming models. In this paper, we consider N-person nonzero-sum noncooperative sequential games in the same framework. The motivating special case is the noncooperative discounted stochastic game studied by Rogers (1969), Sobel (1971), Parthasarathy (1973), Himmelberg, Parthasarathy, Raghavan and Van Vleck (1976) and Federgruen (1978). As with Denardo (1967), the generality and abstraction here is useful to identify the essential structure. The contraction operator framework is also very natural because it emphasizes the reduction of the initial dynamic sequential game to a static one-period game. The final payoff to all players associated with a specification of all strategies is the unique fixed point of the contraction operator; the static game involves the choice of the fixed point. However, the sequential game is not immediately covered by the existing theory of static one-period noncooperative games because, as will be developed, the payoff (fixed-point) is a function of the state.

The contraction assumption means that the criterion for evaluating a payoff stream is discounted present value. However, it is well known that in many instances the average cost criterion can be reduced to a discounting criterion, cf. p. 149 of Ross (1970). Moreover, as in Section 5 of Denardo (1967), we use the N-stage contraction assumption, which covers a larger class of models, including many finite-stage models, cf. Whitt (1977).

A primary purpose of this paper is to apply to noncooperative sequential games the approximation techniques developed for dynamic programs and two-person zero-sum stochastic games in Whitt (1978). The idea is to replace the original state and action

* Received by the editors September 5, 1978, and in revised form March 25, 1979.

† Operations Research Center, Bell Laboratories, Holmdel, New Jersey 07733.

spaces with smaller sets and define a new transition and reward structure to approximate the original. In this way, we show that the extension of an ε_1-equilibrium policy vector in the smaller model is an ε_2-equilibrium policy vector in the original model, where ε_2 is a function of ε_1 and an appropriate measure of oscillation, cf. Theorem 4.2. The approximation results are in turn used to provide conditions under which a noncooperative sequential game has an ε-equilibrium point for each $\varepsilon > 0$, cf. Theorem 5.1.

As special cases, we obtain new results for stochastic games. Of particular interest is the application of the approximation procedure to provide conditions for the existence of ε-equilibrium points for all $\varepsilon > 0$ in the noncooperative discounted stochastic game when the state space is uncountable, cf. Theorem 6.4. The only other results for uncountable state space seem to be in Himmelberg, Parthasarathy, Raghavan and Van Vleck (1976). We also suggest what appears to be a promising procedure for finding ε-equilibrium points in many large noncooperative stochastic games, namely combining the approximation procedure here with an algorithm for finding approximate fixed-points of a continuous function mapping a subset of R^n into itself, cf. Remark (3) at the end of § 6.

A good indication of possible economic applications can be obtained by looking at the specific stochastic game in Kirman and Sobel (1974). As noted by Federgruen (1978), earlier work by Sobel (1973) on discounted stochastic games with uncountable state space, which is applied in Kirman and Sobel (1974), is not valid. Our results can be applied to obtain conditions for the existence of ε-equilibria in the game studied by Kirman and Sobel (1974).

We now briefly indicate how this paper is organized. We begin in § 2 by defining á la Denardo (1967), noncooperative monotone contraction operator games. Following van Nunen (1976), Wessels (1977) and others, we allow for unbounded rewards. As in § 5 of Denardo (1967), we use the N-stage contraction assumption. In § 3 we apply the Glicksberg (1952)–Fan (1952) generalization of the Kakutani fixed-point theorem to obtain sufficient conditions for the existence of equilibrium points. In § 4 we show how two sequential games can be compared, which provides the basis for approximations. In § 5 the approximation scheme is applied to provide conditions for the existence of ε-equilibrium points for each $\varepsilon > 0$. Finally, the special case of a noncooperative stochastic game is investigated in § 6.

2. Noncooperative monotone contraction operator games. Our model of a noncooperative sequential game is a direct extension of Denardo (1967), with the representation of a noncooperative discounted stochastic game being very similar to the representation of Shapley's (1953) two-person zero-sum stochastic game in Example 2 of § 8 in Denardo (1967). Let the *state space* S and the *player space* I be nonempty sets. For each player $i \in I$ and each state $s \in S$, let the *action space* $A_i(s)$ be a nonempty set. To allow for randomized strategies, $A_i(s)$ is often $\mathcal{P}(B_i(s))$, i.e., the set of all probability measures on an underlying action space $B_i(s)$, but we do not stipulate this yet. Let the space of all possible actions for all players in state s be the product space $A(s) = X_{i\in I}A_i(s)$. For each $i \in I$, let the *policy space for player i* be $\Delta_i = X_{s\in S}A_i(s)$. An element δ_i in Δ_i is called a *stationary policy* for player i because it represents the policy that takes action $\delta_i(s)$ every time the system is in state s. Let $\Delta = X_{i\in I}\Delta_i$ represent the space of policies for all players. Throughout this paper, we consider only stationary policies, but the symmetry argument in § 7 of Denardo (1967) can be used to show that no one player acting alone can do better by employing a more general history-remembering policy. Hence, we show that there exist equilibrium points or ε-equilibrium points consisting of stationary policies within the class of all history-remembering policies. Of course, we do not exclude existence of other equilibrium points and ε-equilibrium points consisting of nonstationary policies. While Δ and Δ_i contain only stationary policies, more general policies such as history-remembering policies can be included in this scheme by enlarging the state space. For example, the stage should usually be included as part of the state description in representations of finite-stage sequential games via monotone contraction operator models, cf. Whitt (1977).

Let the space V of potential return functions be a subset of $R^{S\times I}$. In order to allow for unbounded rewards, let $\alpha : S \to (0, \infty)$ and $\beta : S \to R$ be two functions. (The common choice of α and β is $\alpha(s) = 1$ and $\beta(s) = 0$ for all $s \in S$, which yields bounded rewards.) For any $v_1, v_2 \in R^{S\times I}$, let

$$\|v_1\| = \sup\{|v_1(s, i)| : s \in S, i \in I\}$$

(1) and

$$d(v_1, v_2) = \|\alpha(v_1 - v_2)\|,$$

where we regard $\alpha(s)$ as a function of both s and i which is independent of i. Let the space of potential return functions be

$$V = \{v \in R^{S\times I} | d(v, \beta) < \infty\}. \tag{2}$$

It is easy to see that (V, d) is a complete metric space.

The basic ingredient in the model specification is the *local income function* $h(s, i, \mathbf{a}, v)$, which assigns a real number to each quadruple $(s, i, \mathbf{a}, v)$ with $s \in S$, $i \in I$, $\mathbf{a} \in A(s)$ and $v \in V$. The number $h(s, i, \mathbf{a}, v)$ represents the return to player i beginning in state s when player j uses action a_j for all $j \in I$ and all future returns are described by the function v in V. For each $\delta \in \Delta$, let $(H_\delta v)(s, i) = h(s, i, \delta(s), v)$. We make the following basic boundedness (B), monotonicity (M) and N-stage contraction (NC) assumptions about the collection of operators $\{H_\delta, \delta \in \Delta\}$:

(B) There exist constants K_1 and K_2 such that $\|\alpha(H_\delta v - \beta)\| \leqq K_1 + K_2\|\alpha(v - \beta)\|$ for all $\delta \in \Delta$ and $v \in V$.

(M) If $v_1 \leqq v_2$ in V, i.e., if $v_1(s, i) \leqq v_2(s, i)$ for all $s \in S$ and $i \in I$, then $H_\delta v_1 \leqq H_\delta v_2$ for all $\delta \in \Delta$.

(NC) There exists a positive integer N and nonnegative constants m and c, $0 \leqq c < 1$, such that

$$d(H_\delta v_1, H_\delta v_2) \leqq m\, d(v_1, v_2)$$

and

$$d\,(H_\delta^N v_1, H_\delta^N v_2) \leqq c\, d(v_1, v_2)$$

for all $\delta \in \Delta$ and $v_1, v_2 \in V$, where H_δ^N is the N-fold iterate of H_δ.

Obviously (B) implies that the range of H_δ is contained in V. Property (NC) is the N-stage contraction assumption, cf. § 5 of Denardo (1967). The ordinary contraction assumption occurs when $N = 1$. The contraction modulus c often arises as a discount factor. Properties (M) and (NC) imply that each operator H_δ has a unique fixed point v_δ in V which we call the *return function* associated with policy vector δ. Note that the monotone contraction operator model reduces a sequential game to a one-stage game; the set of strategies available to player i is Δ_i and the return to player i from a specification of strategies by all players, i.e., δ, is the fixed point $v_\delta(\cdot, i)$. This differs from the usual static noncooperative game, however, because the return to each player is not a real number, but a function of the state.

A slight modification of Theorem 4 in Denardo (1967) yields

$$d(v_\delta, v) \leqq (1+m+\cdots+m^{N-1})(1-c)^{-1}\, d(H_\delta v, v) \tag{3}$$

for all $\delta \in \Delta$ and $v \in V$. The N-stage contraction assumption covers many N-stage sequential games with $c=0$, cf. Whitt (1977).

It should be noted that it is often possible to transform an N-stage contraction into a 1-stage contraction by modifying the bounding function α. A transformation for Markov programs, which also applies to the stochastic games in § 6 here, was constructed in § 8 of van Nunen (1976). However, it appears that such a transformation is not always possible for the more general monotone contraction operator models here. Moreover, even when such a transformation is possible, the new distance d is different from the old one and may be difficult to compute. Hence, we keep the N-stage contraction assumption.

For any $\delta \in \Delta$ and $\gamma_i \in \Delta_i$, let $[\delta^{-i}, \gamma_i]$ represent the policy vector δ' in Δ with $\delta'_j = \delta_j$ for $j \neq i$ and $\delta'_i = \gamma_i$. Let f_δ represent the *optimal return function* given that the other players are using δ^{-i} for each i, defined by

$$f_\delta(s, i) = \sup\{v_{[\delta^{-1},\gamma_i]}(s, i) : \gamma_i \in \Delta_i\}.$$

Let F_δ be the associated *maximal return operator*, defined by

$$(F_\delta v)(s, i) = \sup\{(H_{[\delta^{-1},\gamma_i]}v)(s, i) : \gamma_i \in \Delta_i\}$$

for each $s \in S$, $i \in I$, $\delta \in \Delta$ and $v \in V$.

Note that property (B) insures that the range of F_δ is in V for each $\delta \in \Delta$. A slight modification of Theorem 4 in Denardo (1967) shows that f_δ is the unique fixed point of F_δ. It is natural to define a *disequilibrium function* $\eta : \Delta \times S \times I \to R$ as $\eta_\delta(s, i) = f_\delta(s, i) - v_\delta(s, i)$. Call a policy δ an *ε-equilibrium point* (ε-EP) if $\eta_\delta(s, i) \leqq \varepsilon/\alpha(s)$ for all i and s, i.e., if $d(f_\delta, v_\delta) \leqq \varepsilon$. Call a policy δ an *equilibrium point* (EP) if it is an ε-EP for $\varepsilon = 0$.

3. Existence of equilibria. The existence of equilibrium points in noncooperative sequential games can be established by applying classical fixed point theorems, following the original line of reasoning used by Nash (1951) to treat static games. This approach has been applied to stochastic games by Rogers (1969), Sobel (1971), Parthasarathy (1973), Himmelberg et al. (1976) and Federgruen (1978). In this paper, we indicate how to apply the Kakutani fixed-point theorem for point-to-set functions as generalized by Glicksberg (1952) and Fan (1952) to the monotone contraction operator games. An alternate approach would be to apply the Brouwer fixed point theorem as generalized by Schauder and Tychonoff, cf. Theorem 1 of Sobel (1971).

Let 2^Y represent the set of all nonempty closed subsets of a Hausdorff topological space Y. Let X be a Hausdorff topological space. A set-valued function $\Phi : X \to 2^Y$ is called upper-semicontinuous (u.s.c.) if $y \in \Phi(x)$ for each $x \in X$, net $\{x_j, j \in J\}$ in X and net $\{y_j, j \in J\}$ in Y such that $x_j \to x$, $y_j \to y$ and $y_j \in \Phi(x_j)$ for each j. (Since X and Y need not be first countable, we use nets instead of sequences, cf. Chapter X of Dugundji (1966).)

THEOREM 3.1 (Kakutani, Glicksberg and Fan). *If X is a convex compact subset of a Hausdorff locally convex topological vector space* (*LCTVS*) *and $\Phi : X \to 2^X$ is convex-valued and u.s.c., then $x \in \Phi(x)$ for some $x \in X$.*

For our application, we want $X = \Delta$ and $\Phi = \psi_\varepsilon$, where $\psi_\varepsilon(\delta) = X_{i \in I}\psi_\varepsilon(\delta)_i$ and

$$\psi_\varepsilon(\delta)_i = \{\gamma_i \in \Delta_i : f_\delta(s, i) \leqq v_{[\delta^{-i},\gamma_i]}(s, i) + \varepsilon/\alpha(s) \quad \text{for} \quad \text{all } s\}. \tag{4}$$

The rest of this section is devoted to providing conditions on the monotone contraction operator game in order for (Δ, ψ_0) to satisfy the conditions of Theorem 3.1. The obvious modification (to account for the metric d in (2)) of Corollary 1 together with Theorem 4 of Denardo (1967) shows that $\psi_\varepsilon(\delta)_i$ is nonempty for each $\varepsilon > 0$. Throughout this paper, let $\Delta = X_{i\in I}\Delta_i$ and $\Delta_i = X_{s\in S}A_i(s)$ be given the product topology, cf. p 98 of Dugandji (1966).

THEOREM 3.2. *There exists an EP if*

(i) *$A_i(s)$ is a convex compact subset of a LCTVS for each $i \in I$ and $s \in S$,*

(ii) *$h(s, i, \mathbf{a}, v)$ is a concave function of a_i for each s, i, $\mathbf{a}$, v, and*

(iii) *$v_\delta(s, i)$ and $f_\delta(s, i)$ are continuous functions of δ for each $s \in S$ and $i \in I$.*

Proof. Since the properties of convexity, Hausdorff, compactness, TVS and LCTVS are preserved under arbitrary products, cf. pp. 138 and 224 of Dugundji (1966) and pp. 19 and 52 of Schaefer (1966), the product spaces Δ_i and Δ are convex compact subsets of a LCTVS. Condition (ii) implies that ψ_ε is convex-valued. Conditions (i) and (ii) plus Corollary 2 of Denardo (1967) show that $\psi_\varepsilon(\delta)_i$ is nonempty for $\varepsilon = 0$ as well as $\varepsilon > 0$. To see that ψ_ε is u.s.c., suppose $\{\delta_j, j \in J\}$ and $\{\delta'_j, j \in J\}$ are nets in Δ with $\delta_j \to \delta$, $\delta'_j \to \delta'$ and $\delta'_j \in \psi_\varepsilon(\delta_j)$ for each $j \in J$. Let δ_{ji} and δ'_{ji} be the ith coordinate in Δ_i of δ_j and δ'_j in Δ. Apply the triangle inequality to obtain

$$\begin{aligned}|v_{[\delta^{-i},\delta'_i]}(s, i) - f_\delta(s, i)| \leqq {} & |v_{[\delta^{-i},\delta'_i]}(s, i) - v_{[\delta_j^{-i},\delta'_{ji}]}(s, i)| \\ & + |v_{[\delta_j^{-i},\delta'_{ji}]}(s, i) - f_{\delta_j}(s, i)| + |f_{\delta_j}(s, i) - f_\delta(s, i)|\end{aligned}$$

for each s and i. The first and third term converge to zero by condition (iii) and the second term is less than or equal to ε for each j because $\delta'_{ji} \in \psi_\varepsilon(\delta_j)$ for each j. Hence, $\delta' \in \psi_\varepsilon(\delta)$, so ψ_ε is u.s.c. and the conditions of Theorem 3.1 are satisfied with $\varepsilon = 0$.

LEMMA 3.1. *If*

(i) *$A_i(s)$ is a compact metric space for each $i \in I$ and $s \in S$,*

(ii) *S is countable, and*

(iii) *$v_\delta(s, i)$ is a continuous function of δ for each $s \in S$ and $i \in I$, then $f_\delta(s, i)$ is a continuous function of δ for each s and i.*

Proof. Suppose $\{\delta_j, j \in J\}$ is a net in Δ with $\delta_j \to \delta$. Let s and i be given. For any $\varepsilon_1, \varepsilon_2 > 0$ there is a $\gamma_1 \in \Delta_i$ and a j_0 such that

$$\begin{aligned} f_\delta(s, i) &\leqq v_{[\delta^{-i}, \gamma_i]}(s, i) + \varepsilon_1 \\ &\leqq v_{[\delta_j^{-i},\gamma_i]}(s, i) + \varepsilon_1 + \varepsilon_2 \quad \text{for } j \geqq j_0 \\ &\leqq f_{\delta_j}(s, i) + \varepsilon_1 + \varepsilon_2 \quad \text{for } j \geqq j_0. \end{aligned}$$

Moreover, there is a net $\{\gamma_{ji}, j \in J\}$ in Δ_i such that

$$f_{\delta_j}(s, i) \leqq v_{[\delta_j^{-i},\gamma_{ji}]}(s, i) + \varepsilon_1 \quad \text{for all } j,$$

so that

$$\limsup_{j\in J} f_{\delta_j}(s, i) \leqq \limsup_{j\in J} v_{[\delta_j^{-i},\gamma_{ji}]}(s, i) + \varepsilon_1.$$

Choose a countable totally ordered subset J' of the directed set J so that the lim sup is attained on the left. Then, using the fact that Δ_i is compact metric space, by virtue of conditions (i) and (ii), choose a convergent subsequence $\{\gamma_{j_k i}\}$ of $\{\gamma_{ji}, j \in J'\}$

with limit γ_i. Hence

$$\begin{aligned}\limsup_{j\in J} f_{\delta_j}(s, i) &\leqq \limsup_{k\to\infty} v_{[\delta_{j_k}^{-i}, \gamma_{j_k i}]}(s, i)+\varepsilon_1\\ &\leqq v_{[\delta^{-i}, \gamma_i]}(s, i)+\varepsilon_1\\ &\leqq f_\delta(s, i)+\varepsilon_1.\end{aligned}$$

LEMMA 3.2. *If $H_\delta v : \Delta\to V$ is a continuous function of δ for each $v\in W$, where W is a subset of V containing v_δ for all $\delta\in\Delta$, then $\nu_\delta : \Delta\to V$ is a continuous function of δ, so that $v_\delta(s, i) : \Delta\to R$ is a continuous function of δ for each $s\in S$ and $i\in I$.*

Proof. By (3),

$$d(v_{\delta_j}, v_\delta)\leqq(1+m+\cdots+m^{N-1})(1-c)^{-1}\, d(H_{\delta_j}v_\delta, v_\delta),$$

where $d(H_{\delta_j}v_\delta, v_\delta)=d(H_{\delta_j}v_\delta, H_\delta v_\delta)\to 0$ as $\delta_j\to\delta$.

The continuity condition in Lemma 3.2 is more likely to hold if W is a subset of V with convenient properties. For example, if $H_\delta v$ is continuous (concave, monotone) for each δ and each continuous (concave, monotone) v in V, then H_δ maps the closed subset of all continuous (concave, monotone) functions in V into itself, so the fixed point v_δ is continuous (concave, monotone). However, even if W has convenient properties, the continuity condition in Lemma 3.2 is quite strong because it requires

$$(5)\qquad d(H_{\delta_n}v, H_\delta v)=\sup_{\substack{s\in S\\ i\in I}}|\alpha(s)(h(s, i, \delta_n(s), v)-h(s, i, \delta(s), v))|\to 0$$

whenever $\delta_n\to\delta$. Since Δ has the product topology, the metric convergence in (5) is difficult to achieve unless S and I are finite. More useful conditions are contained in

LEMMA 3.3. *Suppose $\{\delta_j, j\in J\}$ is a net in Δ converging to δ. If*

(i) *$h(s, i, \delta_j(s), v_j)\to h(s, i, \delta(s), v)$ whenever $v_j(s, i)\to v(s, i)$ for all $s\in S$, $i\in I$ and v_j, $v\in V$; and*

(ii) $\sup_{j\in J} d(H_{\delta_j}^k v_0, v_{\delta_j})\to 0$ as $k\to\infty$ *for some $v_0\in V$; then $v_{\delta_j}(s, i)\to v_\delta(s, i)$.*

Proof. By (i), $(H_{\delta_j}v_0)(s, i)\to(H_\delta v_0)(s, i)$ for all s, i.

By (i) again and mathematical induction,

$$(H_{\delta_j}^k v_0)(s, i)=[H_{\delta_j}(H_{\delta_j}^{k-1}v_0)](s, i)\to[H_\delta(H_\delta^{k-1}v_0)](s, i)=(H_\delta^k v_0)(s, i)$$

as $j\to\infty$ for each $k\geqq 1$. As a consequence of this and (ii), $v_{\delta_j}(s, i)\to v_\delta(s, i)$.

The standard way to make Δ convex and $h(s, i, \mathbf{a}, v)$ concave in a_i is to introduce the *mixed extension*, i.e., let $A_i(s)=\mathcal{P}(B_i(s))$, the set of all probability measures on an underlying action space $B_i(s)$, and let the local income function applied to probability measures be defined via expectation:

$$(6)\qquad h(s, i, \mathbf{a}, v)=\int h(s, i, \mathbf{b}, v)\, d\mu_{\mathbf{a}}(\mathbf{b}),$$

where μ_a is the product probability measure on the product σ-field of $X_{i\in I}B_i(s)$ with one-dimensional marginal probability distributions a_i and the integral is an upper integral if $h(s, i, \mathbf{b}, v)$ is not measurable in $\mathbf{b}$, cf. Example 3 in § 8 of Denardo (1967).

It is well known that if $B_i(s)$ is a topological space and $\mathcal{P}(B_i(s))$ is endowed with the topology of weak convergence, then $\mathcal{P}(B_i(s))$ tends to inherit the topological properties of $B_i(s)$. For completely regular spaces, the weak convergence topology is naturally characterized by the continuity of $\int f\, dP$ in P for each bounded continuous real-valued f. The basic inheritance properties here can be found in § II.6 of Parthasarathy (1967), Varadarajan (1958) and footnote 10 in Fan (1952). Call a measure μ regular [Radon] if

$\mu(A)=\sup\{\mu(C): C\subseteq A\}$ for all measurable subsets A, where the supremum is over all closed [compact] subsets. Obviously regular and Radon are equivalent in compact Hausdorff spaces. The LCTVS that appears below is the space of finite signed measures.

LEMMA 3.4. *Let $A_i(s)=\mathcal{P}(B_i(s))$ with the topology of weak convergence.*

(a) *If $B_i(s)$ is a separable [compact] metric space, then $A_i(s)$ is a separable [compact] metrizable convex subset of a* LCTVS.

(b) *If $B_i(s)$ is a compact Hausdorff space, then the subset of regular probability measures in $\mathcal{P}(B_i(s))$ is a compact convex subset of a LCTVS.*

There is still a major stumbling block—the integral in (6). There is no problem if the set I is countable and the set $B_i(s)$ has a countable base (i.e., is second countable, which is true if $B_i(s)$ is a separable metric space); then the product σ-field on $X_{i\in I}B_i(s)$ will coincide with the Borel σ-field with respect to the product topology. However, if either I is uncountable or if $B_i(s)$ does not have a countable base, then there can be complications. Henceforth, we make the assumptions to avoid the complications. We can combine this observation with Theorem 3.2 and Lemmas 3.1–3.4 to obtain the following result for the mixed extension.

THEOREM 3.3. *If*

(i) *S and I are countable,*

(ii) *$A_i(s)=\mathcal{P}(B_i(s))$ with the topology of weak convergence, where $B_i(s)$ is a compact metric space,*

(iii) *$h(s,i,\mathbf{b}_n,v_n)\to h(s,i,\mathbf{b},v)$ whenever $b_{ni}\to b_i$ and $v_n(s,i)\to v(s,i)$ for each $s\in S$ and $i\in I$,*

(iv) *$h(s,i,\mathbf{a},v)=\int h(s,i,\mathbf{b},v)\,d\mu_{\mathbf{a}}(\mathbf{b})$, where $\mu_{\mathbf{a}}$ is the product measure on $X_{i\in I}B_i(s)$ with marginal measures $a_i\in A_i(s)$,*

(v) *$\sup_n d(H^k_{\delta_n}v_0, v_{\delta_n})\to 0$ as $k\to\infty$ for some v_0 in V and any convergent sequence $\{\delta_n\}$ in Δ,*

then there exists an EP, i.e., there exists $\delta^*\in\Delta$ such that $\delta^*\in\psi_0(\delta^*)$.

Proof. By conditions (i) and (ii) and Lemma 3.4(a), Δ is a convex compact metrizable subset of a LCTVS. By (i) and (ii), the Borel σ-field on $X_{i\in I}B_i(s)$ with the product topology coincides with the product σ-field. By (iii), the integral in (iv) is well defined. By (iii) and the almost-surely convergent representation of weak convergence, cf. Dudley (1968), $h(s,i,\delta_n(s),v_n)\to h(s,i,\delta(s),v)$ whenever $\delta_n(s)\to\delta(s)$ and $v_n(s,i)\to v(s,i)$ for each (s,i). This and (v) plus Lemma 3.3 imply that $v_\delta(s,i)$ is continuous in δ for each (s,i). Lemma 3.1 implies that $f_\delta(s,i)$ is continuous in δ for each (s,i). By (iv), ψ_0 is convex-valued. Hence, all conditions of Theorem 3.2 are satisfied.

Remark. The difficult condition in Theorem 3.3 is (iii). Since the convergence $\mathbf{b}_n\to\mathbf{b}$ and $v_n\to v$ is pointwise in s and i, in order to satisfy (iii) it will often be convenient to have I and/or S finite.

4. Comparing sequential games. Following Whitt (1978), let $(S, I, \{A_i(s), s\in S, i\in I\}, h, \alpha, \beta, c)$ and $(\tilde{S}, \tilde{I}, \{\tilde{A}_i(s), s\in\tilde{S}, i\in\tilde{I}\}, \tilde{h}, \tilde{\alpha}, \tilde{\beta}, \tilde{c})$ be two sequential games as defined in § 2. In order to compare these games, we require that several comparison functions be defined. These comparison functions arise naturally in deliberate approximations, which can be constructed by selecting partitions of subsets of the sets S, I and $A_i(s)$ for each $i\in I$ and $s\in S$, with one point selected in each partition subset, cf. Section 4 of Whitt (1978). In that setting the mappings below correspond to projections and extensions, which is the motivation for the notation. The comparison functions are:

(i) a mapping p of S *onto* $\tilde{S}$;

(ii) a *one-to-one* mapping p of I *onto* $\tilde{I}$;

(iii) a mapping p of $A_i(s)$ *onto* $\tilde{A}_{p(i)}(p(s))$ for each $i\in I$ and $s\in S$;

(iv) a mapping e of $\tilde{S}$ *into* S such that $p(e[\tilde{s}])=\tilde{s}$ for each $\tilde{s}\in\tilde{S}$;

(v) a mapping $e_{s,i}$ of $\tilde{A}_{p(i)}(p(s))$ *into* $A_i(s)$ such that $p(e_{s,i}[\tilde{a}])=\tilde{a}$ for each $\tilde{a}\in\tilde{A}_{p(i)}(p(s))$, $i\in I$ and $s\in S$.

(vi) $e:\tilde{V}\to R^{S\times I}$ with $e(\tilde{v})(s,i)=\tilde{v}(p(s),p(i))$ for each $S\in S$ and $i\in I$;

(vii) $p\colon V\to\tilde{V}$ with $p(v)(\tilde{s},\tilde{i})=v(e(\tilde{s}),e(\tilde{i}))$ for each $\tilde{s}\in\tilde{S}$ and $\tilde{i}\in\tilde{I}$;

(viii) $e\colon\tilde{I}\to I$ with $p[e(\tilde{i})]=\tilde{i}$ for all $\tilde{i}\in\tilde{I}$;

(ix) $e\colon\tilde{\Delta}_{p(i)}\to\Delta_i$ with $e(\tilde{\delta}_{p(i)})(s)=e_{s,i}(\tilde{\delta}_{p(i)}[p(s)])$ for each $s\in S$ and $i\in I$, and

(x) $p\colon\Delta_i\to\tilde{\Delta}_{p(i)}$ with $p(\delta_i)(\tilde{s})=p(\delta_i[e(\tilde{s})])$ for each $\tilde{s}\in\tilde{S}$ and $\tilde{i}\in\tilde{I}$.

Let e and p also map product spaces onto product spaces in the obvious way, e.g., $e:\tilde{\Delta}\to\Delta$ with $e(\tilde{\delta})_i=e(\tilde{\delta}_{p(i)})$ and $p:X_{i\in I}A_i(s)\to X_{p(i)\in\tilde{I}}\tilde{A}_{p(i)}(p(s))$ with $p(\{a_i(s)\})_{p(i)}=p(a_i(s))$ for $a_i(s)\in A_i(s)$ for each $i\in I$ and $s\in S$. Note that $e(\tilde{\delta})\in\Delta$ for each $\tilde{\delta}\in\tilde{\Delta}$.

Assume that $e(\tilde{v})\in V$ for each $\tilde{v}\in\tilde{V}$. Note that this is automatic if $\alpha(s)\leqq\tilde{\alpha}(p(s))$ and $\beta(s)-\tilde{\beta}(p(s))=0$ for all $s\in S$, but might fail in general.

We expect interest to be focused on approximating the action spaces $A_i(s)$, because these spaces—usually being sets of probability measures—are often large. Thus the map $p\colon S\to\tilde{S}$ might often be one-to-one as is the map $p\colon I\to\tilde{I}$, but we do not require it. The "distance" between these models can be expressed in terms of the measure of oscillation

$$\begin{aligned}K(\tilde{v})&=\sup_{\delta\in\Delta} d(H_\delta e(\tilde{v}),e(\tilde{H}_{p(\delta)}\tilde{v}))\\ &=\sup_{\substack{s\in S\\ \delta\in\Delta\\ i\in I}}|\alpha(s)[h(s,i,\delta(s),e(\tilde{v}))-\tilde{h}(p(s),p(i),p[\delta(s)],\tilde{v})]|.\end{aligned}\tag{7}$$

Obviously $p\colon I\to\tilde{I}$ should usually be one-to-one, as already assumed, in order for $K(\tilde{v})$ to have any chance of being small, but the following results hold even if $p\colon I\to\tilde{I}$ were not required to be one-to-one.

THEOREM 4.1. *For any* $\tilde{\delta}\in\tilde{\Delta}$,

$$d(e(\tilde{v}_{\tilde{\delta}}),v_{e(\tilde{\delta})})\leqq(1+m+\cdots+m^{N-1})(1-c)^{-1}K(\tilde{v}_{\tilde{\delta}}).$$

Proof. Just as in Theorem 3.2 of Whitt (1978), substitute $e(\tilde{\delta})$ for δ and $\tilde{v}_{\tilde{\delta}}$ for $\tilde{v}$ in (7) to obtain

$$d(H_{e(\tilde{\delta})}e(\tilde{v}_{\tilde{\delta}}),e(\tilde{v}_{\tilde{\delta}}))\leqq K(\tilde{v}_{\tilde{\delta}}).$$

Finally, apply formula (3) recalling that we have assumed that $e(\tilde{v})\in V$ for each $\tilde{v}\in\tilde{V}$.

THEOREM 4.2. *If* $\tilde{\delta}$ *is an* ε*-EP, then* $e(\tilde{\delta})$ *is an* $(1+m+\cdots+m^{N-1})(1-c)^{-1}(\varepsilon+2K(\tilde{v}_{\tilde{\delta}}))$*-EP.*

Proof. As in the proof of Theorem 3.1 of Whitt (1978),

$$\begin{aligned}\alpha(s)[H_{[e(\tilde{\delta})^{-i},\gamma_i]}\,e(\tilde{v}_{\tilde{\delta}})](s,i)&=\alpha(s)h(s,i,[e(\tilde{\delta})^{-i},\gamma_i](s),e(\tilde{v}_{\tilde{\delta}}))\\ &\leqq\alpha(s)\tilde{h}(p(s),p(i),p([e(\tilde{\delta})^{-i},\gamma_i](s)),\tilde{v}_{\tilde{\delta}})+K(\tilde{v}_{\tilde{\delta}})\\ &\leqq\alpha(s)\tilde{h}(p(s),p(i),p([e(\tilde{\delta})^{-i},\gamma_i](s)),\tilde{f}_{\tilde{\delta}})+K(\tilde{v}_{\tilde{\delta}})\\ &\leqq\alpha(s)\tilde{f}_{\tilde{\delta}}(p(s),p(i))+K(\tilde{v}_{\tilde{\delta}})\\ &\leqq\alpha(s)e(\tilde{v}_{\tilde{\delta}})(s,i)+(K(\tilde{v}_{\tilde{\delta}})+\varepsilon)\end{aligned}$$

for each $s\in S$, $\gamma_i\in\Delta_i$ and $i\in I$. As a consequence of properties (M) and (NC),

$$\begin{aligned}&\alpha(s)[H^N_{[e(\tilde{\delta})^{-i},\gamma_i]}\,e(\tilde{v}_{\tilde{\delta}})](s,i)\\ &\quad\leqq\alpha(s)\,e(\tilde{v}_{\tilde{\delta}})(s,i)+(1+m+\cdots+m^{N-1})(K(\tilde{v}_{\tilde{\delta}})+\varepsilon)\end{aligned}$$

and, by induction,

$$\begin{aligned}&\alpha(s)[H^{Nk}_{[e(\tilde{\delta})^{-i},\gamma_i]}\,e(\tilde{v}_{\tilde{\delta}})](s,i)\\&\qquad\leqq \alpha(s)\,e(\tilde{v}_{\tilde{\delta}})(s,i)+(1+c+\cdots+c^{k-1})(1+m+\cdots+m^{N-1})(K(\tilde{v}_{\tilde{\delta}})+\varepsilon)\end{aligned}$$

for all $k \geqq 1$. Since $d(H^{Nk}_{\delta}v, v_{\delta}) \to 0$ as $k \to \infty$,

$$\alpha(s)v_{[e(\tilde{\delta})^{-i},\gamma_i]}(s,i) \leqq \alpha(s)\,e(\tilde{v}_{\tilde{\delta}})(s,i)+(1-c)^{-1}(1+m+\cdots+m^{N-1})(K(\tilde{v}_{\tilde{\delta}})+\varepsilon)$$

for all $\gamma_i \in \Delta_i$, so that

$$\alpha(s)f_{e(\tilde{\delta})}(s,i) \leqq \alpha(s)e(\tilde{v}_{\tilde{\delta}})(s,i)+(1-c)^{-1}(1+m+\cdots+m^{N-1})(K(\tilde{v}_{\tilde{\delta}})+\varepsilon).$$

Apply Theorem 4.1 to obtain

$$a(s)f_{e(\tilde{\delta})}(s,i) \leqq \alpha(s)v_{e(\tilde{\delta})}(s,i)+(1-c)^{-1}(1+m+\cdots+m^{N-1})(2K(\tilde{v}_{\tilde{\delta}})+\varepsilon)$$

or

$$d(f_{e(\tilde{\delta})}, v_{e(\tilde{\delta})}) \leqq (1-c)^{-1}(1+m+\cdots+m^{N-1})(2K(\tilde{v}_{\tilde{\delta}})+\varepsilon).$$

5. Existence of ε-equilibria. We now combine Theorems 3.3 and 4.2 to obtain sufficient conditions for the existence of ε-EP's in sequential games with uncountable state spaces and noncompact action spaces. The ε-EPs obtained are also mixtures of only finitely many actions for each player in each state. Throughout this section, let m represent several different metrics and let the set I be countable. For any subset C in a metric space (B, m), let

$$C^{\varepsilon} = \{b \in B : m(b, b') < \varepsilon \text{ for some } b' \in C\}.$$

THEOREM 5.1. *If*

(i) *S is a separable metric space*;

(ii) *$\beta(s) = 0$ and $\alpha(s)$ is bounded over any finite sphere $\{s\}^{\varepsilon}$ in S*;

(iii) *for each (i, s), $A_i(s) = \mathcal{P}(B_i(s))$ with the topology of weak convergence, where $B_i(s)$ is a subset with compact closure in a metric space B*;

(iv) *for each i, the set-valued function mapping s into $B_i(s)$ is uniformly continuous: for each $\varepsilon_1 > 0$ there is an $\varepsilon_2 > 0$ such that $B_i(s_1) \subseteq B_i(s_2)^{\varepsilon_1}$ whenever $m(s_1, s_2) \leqq \varepsilon_2$*;

(v) *for each (i, s), $h(s, i, \mathbf{a}, v) = \int h(s, i, \mathbf{b}, v)\, d\mu_{\mathbf{a}}(\mathbf{b})$, where $\mu_{\mathbf{a}}$ is the product measure on the product σ-field on $X_{i\in I}B_i(s)$ with marginal measures $a_i \in A_i(s)$ and the integral is an upper integral if $h(s, i, \mathbf{b}, v)$ is not measurable in $\mathbf{b}$*;

(vi) *for any $\varepsilon_1 > 0$, there is an $\varepsilon_2 > 0$ such that*

$$\sup_{\substack{v\in V\\ i\in I}} |\alpha(s')[h(s', i, \mathbf{b}', v) - h(s'', i, \mathbf{b}'', v)]| < \varepsilon_1$$

if $m(s', s'') \leqq \varepsilon_2$ and $m(b_i', b_i'') \leqq \varepsilon_2$ for all i;

(vii) *$h(s, i, \mathbf{b}, v_n) \to h(s, i, \mathbf{b}, v)$ whenever $b_{ni} \to b_i$ and $v_n(s, i) \to v(s, i)$ for all s, i*;

(viii) *$\sup_n d(H^k_{\delta_n}v_0, v_{\delta_n}) \to 0$ for some v_0 in V and any convergent sequence $\{\delta_n\}$ in Δ*;

then, for any $\varepsilon > 0$, there exists an ε-EP δ^* with $\delta^*_i(s)$ being a probability measure on a finite subset of $B_i(s)$ for each s and i.

Proof. We construct a sequence of approximate models

$$\{(\tilde{S}_n, \tilde{I}_n, \{\tilde{A}_{ni}(s), s \in \tilde{S}_n, i \in \tilde{I}_n\}, h_n, \alpha_n, \beta_n, c), n \geqq 1\}$$

according to the scheme in § 4, each of which satisfies the conditions of Theorem 3.3. Let $\tilde{I}_n = I$ for each $n \geqq 1$. Let $\{s_k\}$ be a countable dense subset of S, which exists by virtue of condition (i). For each $n \geqq 1$, form a countable partition of subsets of S by setting

$$S_{n1} = \{s \in S : m(s, s_1) \leqq n^{-1}\}$$

and

$$S_{nk} = \{s \in S : m(s, s_k) \leqq n^{-1}, s \notin \bigcup_{j=1}^{k-1} S_{nj}\}, \qquad k \geqq 2.$$

For each $n \geqq 1$, let $\tilde{S}_n$ be obtained by selecting one point s_{nk} from each nonempty subset in the partition $\{S_{nk}\}$. (Henceforth, omit all empty partition subsets.) Select the point s_{nk} so that $\alpha(s_{nk}) \geqq \alpha(s)/2$ for all $s \in S_{nk}$. This can be done by condition (ii).

For each $n \geqq 1$, $k \geqq 1$ and $s \in S_{nk}$, form finite partitions $\{B_{nkij}(s), 1 \leqq j \leqq K_{nk}\}$ of nonempty measurable subsets of $B_i(s)$ of common cardinality K_{nk} such that $m(b_1, b_2) \leqq \nu(n)$ if $b_1 \in B_{nkil}(s_1)$ and $b_2 \in B_{nkil}(s_2)$, where $s_1, s_2 \in S_{nk}$ and $\nu(n) \to 0$ as $n \to \infty$. These properties can be satisfied because of conditions (iii) and (iv).

For each $i \in I$, $n \geqq 1$, $k \geqq 1$ and $s \in S_{nk}$, let $B_{ni}(s)$ be a finite subset of $B_i(s)$ obtained by selecting one point from each partition subset $B_{nkij}(s)$, $1 \leqq j \leqq K_{nk}$. Let $\tilde{A}_{nk}(s) = \mathscr{P}(B_{ni}(s))$ for each $i \in I$, $s \in \tilde{S}_n$ and $n \geqq 1$.

We now define the five basic comparison functions. Let $p_n : I \to \tilde{I}_n$ be defined by $p_n(i) = i$. Let $p_n : S \to \tilde{S}_n$ be defined by $p_n(s) = s_{nk}$ if $s, s_{nk} \in S_{nk}$ and $s_{nk} \in \tilde{S}_n$. This obviously yields $m(s, p_n(s)) \leqq n^{-1}$ for all s and n. Let $p_n : A_i(s) \to \tilde{A}_{n,p_n(i)}(p_n(s))$ be defined by

$$p_n(a_i(s))(\{b\}) = a_i(s)(B_{nkij}(s)),$$

where $b \in B_{n,p_n(i)}(p_n(s))$ and $b \in B_{nkp_n(i)j}(p_n(s))$, which requires that $s \in S_{nk}$. This obviously means that $p_n(a_i(s))$ is the probability measure in $\tilde{A}_{n,p_n(i)}(p_n(s))$ assigning mass to each point in $B_{n,p_n(i)}(p_n(s))$ equal to the mass the probability measure $a_i(s)$ assigns to the corresponding partition subset $B_{nkij}(s)$ in $B_i(s)$.

Let the mapping $e_n : \tilde{S}_n \to S$ be defined in the obvious way: $e_n(\tilde{s}_n) = \tilde{s}_n$. Let the mappings $e_{nsi} : \tilde{A}_{n,p_n(i)}(p_n(s)) \to A_i(s)$ be defined by setting

$$e_{nsi}(\tilde{a}_{n,p_n(i)}(p_n(s)))(\{b\}) = \tilde{a}_{n,p_n(i)}(p_n(s))(\{b'\})$$

for $b \in B_{ni}(s)$, $b \in B_{nkij}(s)$ and $b' \in B_{nkp_n(i)j}(p_n(s))$. This implies that $e_n(\tilde{\delta}_{np_n(i)})(s)$ is a probability measure on a finite set for each i, s, n and $\tilde{\delta}_{ni} \in \tilde{\Delta}_{ni}$.

Let the approximate local income functions be defined as

$$h_n(\tilde{s}_n, \tilde{i}_n, \tilde{a}_n, \tilde{v}_n) = h(\tilde{s}_n, \tilde{i}_n, \tilde{a}_n, e_n(\tilde{v}_n))$$

for all n, $\tilde{s}_n \in \tilde{S}_n$, $\tilde{i}_n \in \tilde{I}_n$, $\tilde{a}_{n\tilde{i}_n} \in \tilde{A}_{n\tilde{i}_n}(\tilde{s}_n)$ and $\tilde{v}_n \in \tilde{V}_n$, just as in (4.1) of Whitt (1978). Then, by condition (vi), the measure of oscillation $K_n(\tilde{v}_n)$ in (7) is

$$K_n(\tilde{v}_n) = \sup_{\substack{s \in S \\ i \in I \\ \delta \in \Delta}} |\alpha(s)[h(s, i, \delta(s), e_n(\tilde{v}_n)) - h(p_n(s), i, p_n[\delta(s)], e_n(\tilde{v}_n))]|$$

$$\leqq \sup |\alpha(s')[h(s', i, \mathbf{b}', v) - h(s'', i, \mathbf{b}'', v)]|$$

where the second supremum is over all $v \in V$, all s', $s'' \in S$ with $m(s', s'') \leqq n^{-1}$, $b'_i \in B_i(s')$ and $b''_i \in B_i(s'')$ with $m(b'_i, b''_i) \leqq \nu(n) \to 0$ as $n \to \infty$ and all $i \in I$. Hence, condition (vi) implies that $\sup_{\tilde{\delta}_n \in \tilde{\Delta}_n} K(\tilde{v}_{\tilde{\delta}_n}) \to 0$ as $n \to \infty$.

The construction above plus conditions (vii) and (viii) imply that the conditions in Theorem 3.2 are satisfied in each approximate model, so there exists an EP in each approximate model. Theorem 4.2 then implies that, for each $\varepsilon > 0$, there is an n_0 such that the extension of each EP in the nth approximate model is an ε-EP in the original model for all $n \geqq n_0$.

Remarks. (1) Note that conditions (vii) and (viii) are only applied to establish the existence of an EP in each approximate model. If the existence is already known, these conditions can be omitted. The conditions could also be stated for each approximate model. For example, if I is finite, then (vi) can be replaced with $h(s, i, \mathbf{b}, v_n) \to h(s, i, \mathbf{b}, v)$ whenever $v_n(s, i) \to v(s, i)$.

(2) If S is a subset with compact closure in a metric space, then $\tilde{S}_n$ can be finite for each n.

6. Stochastic games. We now consider the special case of a noncooperative stochastic game. As before, let the set I of players be finite or countably infinite. Let the sets S and $B_i(s)$ be separable metric spaces endowed with their Borel σ-fields. Let $A_i(s)$ be the space $\mathcal{P}(B_i(s))$ with the topology of weak convergence. A stochastic game is obtained by letting the local income function be

$$h(s, i, \mathbf{b}, v) = r(s, i, \mathbf{b}) + \int_S v(x, i) q(dx|s, \mathbf{b}), \tag{8}$$

where $r(s, i, \mathbf{b})$ is a measurable real-valued function of $s \in S$, $i \in I$ and $\mathbf{b} \in X_{i \in I} B_i(s)$, $q(C|s, \mathbf{b})$ is a subprobability measure on S for each $s \in S$ and $\mathbf{b} \in X_{i \in I} B_i(s)$ and a measurable function of $(s, \mathbf{b})$ for each measurable subset C, and the integral in (8) is an abstract Lebesgue integral if v is measurable and an upper integral otherwise. (We assume the integral is well defined, i.e., the integral of $|v|$ is finite.)

Also let

$$r_\delta(s, i) = r(s, i, \delta(s)) = \int r(s, i, \mathbf{b})\, d\mu_{\delta(s)}(\mathbf{b}),$$

and

$$q_\delta(C|s) = q(C|s, \delta(s)) = \int q(C|s, \mathbf{b})\, d\mu_{\delta(s)}(\mathbf{b}),$$

where $\mu_{\delta(s)}$ is the probability measure on the product space $X_{i \in I} B_i(s)$ with marginal measures $\delta_i(s)$ for each i.

Let the associated return operator H_δ be defined by

$$\begin{aligned}(H_\delta v)(s, i) &= h(s, i, \delta(s), v) \\ &= \int h(s, i, \mathbf{b}, v)\, d\mu_{\delta(s)}(\mathbf{b}) \\ &= r_\delta(s, i) + \int_S v(x, i) q_\delta(dx|s).\end{aligned}$$

Let the space (V, d) of potential return functions be as in (1) and (2). Let $(q_\delta w)(s) = \int w(x) q_\delta(dx|s)$ for any function w for which the integral is defined. Following van Nunen (1976), with the obvious modification to include N-stage contractions, we make the following assumptions:

$$\|\alpha(r_\delta - (1-c)\beta)\| \leqq M_1, \tag{9}$$

$$\|\alpha(q_\delta \beta - c\beta)\| \leqq M_2, \tag{10}$$

$$\begin{aligned}&\|\alpha q_\delta \alpha^{-1}\| = \sup_{s \in S} \left| \alpha(s) \int [1/\alpha(x)] q_\delta(dx|s) \right| \leqq m, \\ &\|\alpha q_\delta^N \alpha^{-1}\| \leqq c < 1\end{aligned} \tag{11}$$

for all $\delta \in \Delta$, where q_δ^N is the N-step transition kernel associated with q_δ, defined as usual by

$$q_\delta^N(C|s) = \int_S q_\delta^{N-1}(C|s')q_\delta(ds'|s).$$

Conditions under which (11) hold with $N=1$ are discussed by van Hee and Wessels (1977). If $\alpha(s)=1$ for all s, then (11) holds with $N=1$ if $q_\delta(S|s) \leqq c$, which arises naturally if a discount factor c has been incorporated into the probability transition function. As a straightforward extension of Lemma 3.2.2 of van Nunen (1976), we have

THEOREM 6.1. *Under* (9)–(11), *the return operators* H_δ, $\delta \in \Delta$, *satisfy properties* (B), (M) *and* (NC). *Moreover,* H_δ^N *maps* V_0 *into itself, where*

$$V_0 = \{v \in V: \|\alpha(v-\beta)\| \leqq (1+m+\cdots+m^{N-1})(1-c)^{-1}(M_1+M_2)\}. \tag{12}$$

Proof. (B) Note that

$$\begin{aligned}\|\alpha(H_\delta v-\beta)\| &= \|\alpha(r_\delta+q_\delta v-\beta)\| \\ &= \|\alpha(r_\delta-(1-c)\beta)+\alpha(q_\delta(v-\beta))+\alpha(q_\delta\beta-c\beta)\| \\ &\leqq \|\alpha(r_\delta-(1-c)\beta)\|+\|\alpha q_\delta(v-\beta)\|+\|\alpha(q_\delta\beta-c\beta)\| \\ &\leqq M_1+\|\alpha q_\delta\alpha^{-1}\|\cdot\|\alpha(v-\beta)\|+M_2 \\ &\leqq M_1+M_2+m\|\alpha(v-\beta)\|.\end{aligned}$$

(M) This is straightforward.

(NC) For any v₁, v₂ ∈ V,

$$\begin{aligned}d(H_\delta v_1, H_\delta v_2) &= \|\alpha q_\delta(v_1-v_2)\| \\ &\leqq \|\alpha q_\delta\alpha^{-1}\|\cdot\|\alpha(v_1-v_2)\| \leqq md(v_1,v_2)\end{aligned}$$

and

$$\begin{aligned}d(H_\delta^N v_1, H_\delta^N v_2) &= \|\alpha q_\delta^N(v_1-v_2)\| \\ &\leqq \|\alpha q_\delta^N\alpha^{-1}\|\cdot\|\alpha(v_1-v_2)\| \leqq cd(v_1,v_2).\end{aligned}$$

(V_0) First note that

$$\begin{aligned}\|\alpha q_\delta^k w\| &\leqq \|\alpha q_\delta^k\alpha^{-1}\|\cdot\|\alpha w\| \\ &\leqq \|\alpha q_\delta(q_\delta^{k-1}\alpha^{-1})\|\cdot\|\alpha w\| \\ &\leqq \|\alpha q_\delta\alpha^{-1}\|\cdot\|\alpha q_\delta^{k-1}\alpha^{-1}\|\cdot\|\alpha w\| \leqq m^k\|\alpha w\|.\end{aligned}$$

For any $v \in V$,

$$\begin{aligned}\alpha(H_\delta^N v-\beta) &= \alpha[r_\delta+q_\delta r_\delta+q_\delta^2 r_\delta+\cdots+q_\delta^{N-1}r_\delta+q_\delta^N v-\beta] \\ &= \alpha[r_\delta-(1-c)\beta]+\alpha q_\delta[r_\delta-(1-c)\beta] \\ &\quad+\cdots+\alpha q_\delta^{N-1}[r_\delta-(1-c)\beta]+\alpha q_\delta^N[v-\beta] \\ &\quad+\alpha[q_\delta\beta-c\beta]+\alpha q_\delta[q_\delta\beta-c\beta] \\ &\quad+\cdots+\alpha q_\delta^{N-1}[q_\delta\beta-c\beta],\end{aligned}$$

so that

$$\begin{aligned}\|\alpha(H_\delta^N v-\beta)\| &\leqq (1+m+\cdots+m^{N-1})(M_1+M_2)+\|\alpha q_\delta^N \alpha^{-1}\|\cdot\|\alpha(v-\beta)\| \\ &\leqq (1+m+\cdots+m^{N-1})(M_1+M_2)+c\|\alpha(v-\beta)\|,\end{aligned}$$

which implies that

$$\|\alpha(H_\delta^N v-\beta)\|\leqq(1+m+\cdots+m^{N-1})(1-c)^{-1}(M_1+M_2)$$

if

$$\|\alpha(v-\beta)\|\leqq(1+m+\cdots+m^{N-1})(1-c)^{-1}(M_1+M_2).$$

We now determine sufficient conditions for the existence of an EP by applying Theorem 3.3.

THEOREM 6.2. *The stochastic game defined above has an EP if*

(i) *S is countable*;

(ii) *$B_i(s)$ is a compact metric space for each i and s*;

(iii) *$r(s, i, \mathbf{b})$ and $q(\{s'\}|s, \mathbf{b})$ are continuous functions of $\mathbf{b}$ in $X_{i\in I}B_i(s)$ for each s, s' and i*; *and*

(iv) *for any $\varepsilon>0$ and convergent sequence $\{\mathbf{b}_n\}$, there exists a finite subset C of S such that*

$$\sum_{s'\in S-C}(|\beta(s')|+\alpha^{-1}(s'))q(\{s'\}|s, \mathbf{b}_n)<\varepsilon \quad \textit{for all } n.$$

Remarks. (1) Condition (iv) follows from condition (iii) if $\beta(s)=0$ and $\alpha(s)=1$ because the convergence $q(\{s'\}|s, \mathbf{b}_n)\to q(\{s'\}|s, \mathbf{b})$ implies uniform tightness, cf. p. 47 of Parthasarathy (1967).

(2) Conditions (iii) and (iv) are both satisfied automatically if I is finite and $B_1(s)$ is countable and discrete for each i and s.

(3) Theorem 6.2 reduces to Theorem 1 of Federgruen (1976) when $N=1$, $\beta(s)=0$, $\alpha(s)=1$ and I is finite—which in turn reduces to Theorem 1 of Sobel (1971)—when, in addition, S and $B_i(s)$ for each i and s are finite.

Proof. We show that the five conditions of Theorem 3.3 are satisfied. By direct assumption, conditions (i), (ii) and (iv) hold here. By condition (iii) and (iv) here

$$\begin{aligned}h(s, i, \mathbf{b}_n, v_n) &= r(s, i, \mathbf{b}_n)+\sum_{s'\in S} v_n(s', i)q(\{s'\}|s, \mathbf{b}_n) \\ &\to r(s, i, \mathbf{b})+\sum_{s'\in S} v(s', i)q(\{s'\}|s, \mathbf{b}) = h(s, i, \mathbf{b}, v),\end{aligned}$$

which is condition (iii) of Theorem 3.3. Finally, condition (v) holds because, for any $\delta\in\Delta$,

$$\begin{aligned}d(H_\delta^{Nk}v_0, v_\delta) &= d(H_\delta^{Nk}v_0, H_\delta^{Nk}v_\delta) \\ &\leqq c^k d(v_0, v_\delta) \\ &\leqq c^k(\|\alpha(v_0-\beta)\|+\|\alpha(v_\delta-\beta)\|) \\ &\leqq 2c^k(1+m+\cdots+m^{N-1})(1-c)^{-1}(M_1+M_2).\end{aligned}$$

We now consider comparisons between the stochastic game model $(S, I, \{B_i(s), i\in I, s\in S\}, h, \alpha, \beta, c)$ and a "smaller" stochastic game model $(\tilde{S}, \tilde{I}, \{\tilde{B}_i(s), i\in\tilde{I}, s\in\tilde{S}\}, \tilde{h}, \tilde{\alpha}, \tilde{\beta}, \tilde{c})$ which are both assumed to satisfy (8)–(11). Assume that the comparison functions in § 4 have been defined. Let $\tilde{S}$, $\tilde{B}_i(s)$ for each $i \in \tilde{I}$ and $s \in \tilde{S}$, and $\tilde{I}$ be countable sets. Assume that

$$S_n = p^{-1}(\tilde{s}_n) = \{s \in S : p(s) = \tilde{s}_n\}, \qquad \tilde{s}_n \in \tilde{S}$$

and

$$\begin{aligned} B_{ni}(s) &= p^{-1}(\tilde{b}) \cap B_i(s) \\ &= \{b \in B_i(s) : p(i) = \tilde{i}, p(s) = \tilde{s}, p(b) = \tilde{b}\}, \qquad \tilde{b} \in \tilde{B}_i(\tilde{s}), \end{aligned}$$

are measurable subsets for each n, i and s.

As in § 4, assume that $e(\tilde{v}) \in V$ for each $\tilde{v} \in \tilde{V}$. In this setting, the comparison results in § 4 can be expressed in terms of the measures of oscillation

$$K_r = \sup_{\substack{s \in S \\ i \in I \\ \mathbf{b} \in XB_i(s) \\ i \in I}} |\alpha(s)[r(s, i, \mathbf{b}) - \tilde{r}(p(s), p(i), p(\mathbf{b}))]|$$

$$K_q(v) = \sup_{\substack{s \in S \\ \mathbf{b} \in X \in B_i(s) \\ i \in I}} \left\{ \alpha(s) \sum_{n=1}^{\infty} (|\tilde{v}(s_n)|) \Big| q(S_n | s, \mathbf{b}) - \tilde{q}(\{s_n\} | p(s), p(\mathbf{b})) \Big| \right\}$$

and $K_q = K_q(\tilde{v}^*)$, where

$$\tilde{v}^*(s_n) = \sup_{s \in S_n} \{\tilde{\beta}(s_n) + \tilde{\alpha}^{-1}(s_n)(1 + m + \cdots + m^{N-1})(1 - \tilde{c})^{-1}(\tilde{M}_1 + \tilde{M}_2)\}, \qquad n \geqq 1. \tag{14}$$

THEOREM 6.3. *For any* $\tilde{\delta} \in \tilde{\Delta}$, $K(\tilde{v}_\delta) \leqq K_r + K_q(\tilde{v}_{\tilde{\delta}}) \leqq K_r + K_q$.

Proof. By (7) and the triangle inequality,

$$\begin{aligned} K(\tilde{v}_{\tilde{\delta}}) &= \sup_{\substack{s \in S \\ i \in I \\ \delta \in \Delta}} |\alpha(s)[h(s, i, \delta(s), e(\tilde{v}_{\tilde{\delta}})) - \tilde{h}(p(s), p(i), p[\delta(s)], \tilde{v}_{\tilde{\delta}})]| \\ &\leqq \sup_{\substack{s \in S \\ \delta \in \Delta \\ i \in I}} |\alpha(s)[r(s, i, \delta(s)) - \tilde{r}(p(s), p(i), p[\delta(s)])]| \\ &\qquad + \sup_{\substack{s \in S \\ \delta \in \Delta}} \alpha(s) \sum_{n=1}^{\infty} (|\tilde{v}_{\tilde{\delta}}(s_n)|) |q(S_n | s, \delta(s)) - \tilde{q}(\{s_n\} | p(s), p[\delta(s)])| \\ &\leqq K_r + K_q(\tilde{v}_{\tilde{\delta}}) \leqq K_r + K_q, \end{aligned}$$

where the last step follows because $|\tilde{v}_{\tilde{\delta}}(s_n)| \leqq \tilde{v}^*(s_n)$ for all n by (12).

Remarks. When $\alpha(s) = \tilde{\alpha}(p[s]) = 1$ and $\beta(s) = \tilde{\beta}(p[s]) = 0$ for all s, Theorem 6.3 reduces to Theorem 6.1(a) of Whitt (1978). For further refinements, see § 6 of Whitt (1978).

We now present sufficient conditions for the stochastic game to have an ε-EP for each $\varepsilon > 0$. For simplicity, we assume I is finite and $\alpha(s) = 1$ and $\beta(s) = 0$ for all s.

THEOREM 6.4. *The stochastic game has an ε-EP for each $\varepsilon > 0$ if*

(i) *I is finite;*

(ii) *$B_i(s)$ is a subset with compact closure in a separable metric space for each i and s;*

(iii) *the point-to-set function mapping s into $B_i(s)$ is uniformly continuous for each i: for each $\varepsilon_1 > 0$, there exists an $\varepsilon_2 > 0$ such that $B_i(s_1) \subseteq B_i(s_2)^{\varepsilon_1}$ if $m(s_1, s_2) < \varepsilon_2$;*

(iv) *$\alpha(s) = 1$ and $\beta(s) = 0$ for all s;*

(v) *$r(s, i, \mathbf{b})$ and $q(C|s, b)$ are uniformly continuous in s and $\mathbf{b}$, uniformly in C.*

Proof. Construct a sequence of approximating models as in the proof of Theorem 5.1. Note that conditions (i)–(v) of Theorem 5.1 have been assumed again here and condition (vi) of Theorem 5.1 holds because of conditions (iv) and (v) here. For this purpose, it suffices to consider only those v with $|v(s, i)| \leqq (1 + \cdots + m^{N-1})(1-c)^{-1}M_1$. Alternatively, it is easy to see that $K_n(\tilde{v}_{\tilde{\delta}_n}) \to 0$ by applying Theorem 6.3. Theorem 6.2 implies that each approximate game has an EP.

Remarks. (1) The transition kernel q satisfies condition (5) in Theorem 6.4 if $q(C|s, \mathbf{b}) = \int_C f(x|s, \mathbf{b})\lambda(dx)$, for all measurable subsets C, where λ is a finite measure on S and $f(x|s, \mathbf{b})$ is uniformly continuous in s and $\mathbf{b}$, uniformly in x.

(2) To see that it is not sufficient in Theorem 6.4 to have $q(\cdot\,|s, \mathbf{b})$ be uniformly continuous in the space of probability measures on S with the topology of weak convergence, let S be the unit circle, i.e., $S = [0, 1)$ with the metric $m(s_1, s_2) = \min\{s_2 - s_1, 1 - s_2 + s_1\}$ for $s_1 \leqq s_2$. Let $T: S \to S$ be defined by $T(s) = s + \lambda \pmod 1$ where λ is a fixed irrational number. Let $q(\{T_s\}|s, \mathbf{b}) = c$ and $q(S - \{T_s\}s, \mathbf{b}) = 0$ for all $s, \mathbf{b}$. Then $q(\cdot\,|s, \mathbf{b})$ is a uniformly continuous function of $(s, \mathbf{b})$ into the space of probability measures on S with the weak convergence topology. However, since the transformation T is ergodic, it is impossible to have $K_q < c$ for K_q in (6.9) and any countable partition of S.

(3) If, in addition to the assumptions of Theorem 6.4, S is a subset of a compact metric space, then there is a natural algorithm to find an ε-EP. Since each approximate model then can have S as well as I and $B_i(s)$ finite, the EP's in each approximate model can be found by applying Brouwer's fixed point theorem, as shown in Theorem 1 of Sobel (1971). Hence, it suffices to apply one of the algorithms for finding an approximate fixed point of a continuous function mapping a subset of R^n into itself, cf. Karmardian (1976).

(4) We have yet to determine interesting sufficient conditions for the existence of an EP (rather than an ε-EP) when S is uncountable. For example, suppose $S = [0, 1]$, $I = \{1, 2\}$, $B_i(s) = \{1, 2\}$ and $A_i(s) = \mathcal{P}(B_i(s))$ for each i and s. For simplicity consider either (1)$N = 1$ or (2) $N = 2$ and $c = 0$.

Acknowledgment. Part of this research was conducted in the School of Organization and Management at Yale University. I am grateful to Eric Denardo, Martin Shubik and Matthew Sobel for stimulating my interest in this area. Eric Denardo made a very significant contribution to this whole effort by suggesting improvements in my earlier paper Whitt (1978). The first drafts of Whitt (1978) were cast in the setting of Markov programs and the previous version of Theorem 3.1 there combined several tight bounds in a suboptimal way. Eric Denardo showed that the analysis could be carried out in the more general framework of monotone contraction operators and, as acknowledged in Whitt (1978), provided the current form of Theorem 3.1 there. I am also grateful to R. Emerson Thomas and the referee for helpful comments on this paper.

REFERENCES

E. V. DENARDO (1967), *Contraction mappings in the theory underlying dynamic programming*, SIAM Rev., 9, pp. 165–177.

R. M. DUDLEY (1968), *Distances of probability measures and random variables*, Ann. Math. Statist., 39, pp. 1563–1572.

J. DUGUNDJI (1966), *Topology*, Allyn and Bacon, Boston.

K. FAN (1952), *Fixed-point and minimax theorems in locally convex topological linear spaces*, Proc. Nat. Acad. Sci. U.S.A., 38, pp. 121–126.

A. FEDERGRUEN (1978), *On N-person stochastic games with denumerable state space*, Advances Appl. Probability, 10, pp. 452–471.

I. GLICKSBERG (1952), *A further generalization of the Kakutani fixed point theorem with application to Nash equilibrium points*, Proc. Amer. Math. Soc., 3, pp. 170–174,

C. J. HIMMELBERG, T. PARTHASARATHY, T. E. S. RAGHAVAN AND F. S. VAN VLECK (1976), *Existence of p-equilibrium and optimal stationary strategies in stochastic games*, Proc. Amer. Math. Soc., 60, pp. 245–251.

S. KARMADIAN (1976), *Fixed points: Algorithms and Applications*, Academic Press, New York.

A. P. KIRMAN AND M. J. SOBEL (1974), *Dynamic oligolopy with inventories*, Econometrica, 42, pp. 279–287.

J. F. NASH (1951), *Noncooperative games*, Ann. of Math., 54, pp. 286–295.

K. R. PARTHASARATHY (1967), *Probability Measures on Metric Spaces*, Academic Press, New York.

T. PARTHASARATHY (1973), *Discounted, positive and non-cooperative stochastic games.* Internat. J. Game Theory, 2, pp. 25–37.

P. D. ROGERS (1969), *Non-zero sum stochastic games*, Ph.D. dissertation, University of California, Berkeley.

S. M. ROSS (1970), *Applied Probability Models with Optimization Applications*, Holden-Day, San Francisco.

H. H. SCHAEFER (1966) *Topological Vector Spaces*, Macmillan, London.

L. S. SHAPLEY (1953), *Stochastic games*, Proc. Nat. Acad. Sci. U.S.A., 39, pp. 1095–1100.

M. J. SOBEL (1971), *Noncooperative stochastic games*, Ann. Math. Statist., 42, pp. 1930–1935.

——— (1973), *Continuous stochastic games*, J. Appl. Probability, 10, pp. 597–604.

K. M. VAN HEE AND J. WESSELS (1978), *Markov decision processes and strongly excessive functions*, Stochastic Processes Appl., 8, pp. 59–76.

J. A. E. E. VAN NUNEN (1976), *Contracting Markov decision processes*, Ph.D. dissertation, Eindhoven University of Technology (Mathematical Center Tract Number 71, Amsterdam.)

V. VARADARAJAN (1958), *Weak convergence of measures on separable metric spaces*, Sankhyā Ser. A, pp. 15–22.

J. WESSELS (1977), *Markov programming by successive approximations with respect to weighted supremum norms*, J. Math. Anal. Appl., 58, pp. 326–335.

W. WHITT (1978), *Approximation of dynamic programs I*, Math. Operations Res., 3, pp. 231–243.

——— (1977), *Respresentation and approximation of finite-stage dynamic programs*, School of Organization and Management, Yale University.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 1, January 1980

0363-0129/80/1801-0004$01.00/0

A STABILITY THEORY FOR THE LINEAR-QUADRATIC-GAUSSIAN PROBLEM FOR SYSTEMS WITH DELAYS IN THE STATE, CONTROL, AND OBSERVATIONS*

R. H. KWONG†

Abstract. The estimation and control of linear stochastic systems with delays in the state, control, and observations are studied. First, the deterministic optimal control problem with quadratic cost over an infinite time interval is examined. Using an extended notion of stabilizability, the existence and characterization of the optimal control law is obtained. Using the additional assumption of detectability the optimal closed-loop system is shown to be L^2-stable. Next, the stochastic filtering problem is studied. A new version of the duality relations between optimal control and filtering is developed. This combined with a suitable notion of detectability, is exploited to show convergence of the filter gains. Under the additional assumption of stabilizability, the optimal stationary filter is shown to be L^2-stable. Finally, by putting together the optimal control and filtering results, a stable constant stochastic control law is obtained for the linear-quadratic-Gaussian problem.

1. Introduction. Recently, there have been many investigations on control and filtering problems for linear systems with delays in the state [1]–[13]. Both finite as well as infinite time problems have been treated, and various viewpoints and techniques have been developed. In [11], [12], we have given a complete linear-quadratic-Gaussian theory for linear systems with a single delay in the state, although the same methods can be extended to cover linear systems with multiple and distributed state delays. The situation is quite different when there are also delays in the control and observations. Koivo and Lee [14] studied the quadratic control problem for linear systems with delays in the state and control and derived the optimal feedback law. Bagchi [10] and Kwong and Willsky [11], [12] obtained the optimal filter for linear systems with delays in the state and observations. Lindquist, in a series of papers [8], [15], [16] discussed the stochastic control problem and proved versions of the separation theorem. However, all the above papers are concerned only with finite time problems (the infinite time problems treated in [11] and [12] were for systems with no control or observation delays). Thus, qualitative properties such as stability of the optimal control law or optimal filter have not been studied. In this paper, we shall present a linear-quadratic-Gaussian theory for systems with delays in the state, control, and observations, with particular emphasis on infinite time problems, stability of control laws and filters, and relationship to the notions of stabilizability and detectability. We shall first discuss the finite time quadratic control problem for linear systems with delays in the state and control. Although this problem has been studied earlier by Koivo and Lee [14], their results are incomplete as the expression for the optimal cost was not given. We complete the picture by presenting the expression for the optimal cost and deriving differential equations satisfied by the gains. Next, we study the infinite time quadratic control problem. A stabilizability notion is formulated which enables us to solve the infinite time problem and obtain the optimal control law. Under the additional assumption of detectability, the optimal closed-loop system is shown to be L^2-stable. We then turn to the problem of optimal filtering of linear stochastic systems with delays in the state and observations. Duality relations between optimal control and filtering, and between the notions of stabilizability and detectability are proved. These relations enable us to exploit the infinite time optimal control results to prove

* Received by the editors December 30, 1977, and in revised form May 29, 1979. This research was supported by the National Research Council of Canada under Grants A4786 and A0875.

† Department of Electrical Engineering, University of Toronto, Toronto, Ontario, Canada M5S 1A4.

convergence of the filter gains. The resulting stationary optimal filter is shown to be the "adjoint" system to the optimal stationary closed-loop system for the infinite time control problem. Using this fact together with an additional assumption of stabilizability, the optimal stationary filter is shown to be L^2-stable. By combining the results for deterministic optimal control and optimal filtering, we obtain a stochastic control law for the linear-quadratic-Gaussian problem which is L^2-stable. The approach used here parallels the one used in [12]. Other approaches to the problem are certainly possible, and in fact, Ichikawa [23] has independently studied the finite time problem using the method of evolution equations. It would be interesting to see how his approach can be used in the finite time problem and how it would compare with the results obtained here. Numerical and implementation aspects of the theory have not been considered here at all. These are important in their own right, but must be left for future investigations. A summary of the results here was presented in [17].

2. Finite time quadratic optimal control for systems with delays in the state and control. We begin our investigation with the deterministic optimal control problem for linear systems with delays in the state and control. The system under consideration is given by

$$\begin{aligned} &\dot{x}(t)=A_0x(t)+A_1x(t-h)+B_0u(t)+B_1u(t-h),\\ (2.1)\qquad &x(\theta)=x_0(\theta), \qquad \theta\in[-h,0],\\ &u(\theta)=u_0(\theta), \qquad \theta\in[-h,0). \end{aligned}$$

The state vector $x(t)$ takes value in R^n, the control vector in R^m. The constant matrices A_0 and A_1 are $n\times n$, while B_0 and B_1 are $n\times m$. The positive fixed number h is the length of the delay interval. The initial trajectory piece x_0 is taken to be an element in the space $R^n\times L^2[-h,0]$, denoted by $M^2[-h,0]$ (or simply M^2) as in [3]. That is, $x_0=(x_0^0,x_0^1)$ when x_0^0 is a vector in R^n, and x_0^1 is an element of $L^2[-h,0]$. For any $\phi=(\phi^0,\phi^1)$ and $\omega=(\omega^0,\omega^1)$ in M^2, their inner product is defined by $\langle\phi,\omega\rangle_{M^2}=\langle\phi^0,\omega^0\rangle_{R^n}+\langle\phi^1,\omega^1\rangle_{L^2[-h,0]}$. The norm on M^2 is the one induced by this inner product. The initial control piece u_0 is taken to be an element of the space $L^2[-h,0]$. The symbol ψ (possibly with subscripts) will be used for elements of the space $M^2[-h,0]\times L^2[-h,0]$. Define also the linear operator M mapping $L^2[-h,0]\times L^2[-h,0]$ into $L^2[-h,0]$ by

$$M(\phi,v)(\theta)=A_1\phi(\theta)+B_1v(\theta), \qquad \theta\in[-h,0].$$

A moment's reflection shows that in order to determine the future state trajectory $x(t)$, $t\geqq s$, we need to know the values of $x(s)$, the function $M(x_s,u_s)$, and the future inputs $u(\sigma)$, $s\leqq\sigma\leqq t$. Here, x_s and u_s are defined as usual by

$$\begin{aligned} x_s(\theta)&=x(s+\theta), \qquad \theta\in[-h,0],\\ u_s(\theta)&=u(s+\theta), \qquad \theta\in[-h,0]. \end{aligned}$$

Thus, we might think of the pair $(x(s),M(x_s,u_s))$ as the true state of the system. In this paper, we shall be primarily interested in the infinite time control problem, particularly stability properties of the optimal closed-loop system and their relation to the properties of stabilizability and detectability. However, we shall have to first discuss the finite time problem as existing results are incomplete. We consider therefore the following cost functional associated with (2.1)

$$J_T(u)=\int_0^T[x'(t)H'Hx(t)+u'(t)Su(t)]\,dt$$

where $T<\infty$, H a $p\times n$ constant matrix and S a positive definite $m\times m$ constant matrix. The problem is to choose u in $L^2[0, T]$ such that J_T is minimized. This problem was studied by Koivo and Lee [14], who obtained the optimal control in feedback form as follows:

$$\begin{aligned} u(t)=S^{-1}B_0'\Big[L(t,\tau,\tau)x(\tau)&+\int_{\tau-h}^{\tau} L(t,\sigma+h,\tau)A_1x(\sigma)\,d\sigma \\ &+\int_{\tau-h}^{\tau} L(t,\sigma+h,\tau)B_1u(\sigma)\,d\sigma\Big] \\ -S^{-1}B_1'\Big[L(t+h,\tau,\tau)x(\tau)& \\ &+\int_{\tau-h}^{\tau} L(t+h,\sigma+h,\tau)A_1x(\sigma)\,d\sigma \\ &+\int_{\tau-h}^{\tau} L(t+h,\sigma+h,\tau)B_1u(\sigma)\,d\sigma\Big] \end{aligned} \tag{2.2}$$

where the function $L(t, s, \tau)$ for $\tau\leqq s$, $t\leqq T$, satisfies a Fredholm integral equation

$$\begin{aligned} L(t,s,\tau)&=M(t,s)-\int_{\tau}^{T} L(t,\sigma,\tau)\Gamma'(s,\sigma)\,d\sigma \\ &=M(t,s)-\int_{\tau}^{T}\Gamma(t,\sigma)L(\sigma,s,\tau)\,d\sigma \\ L(t,s,\tau)&=0 \quad \text{if } t \text{ or } s>T. \end{aligned} \tag{2.3}$$

Here the matrix-valued function $M(t, s)$ is given by

$$M(t,s)=\int_{\max(t,s)}^{T}\Phi'(\sigma,t)H'H\Phi(\sigma,s)\,d\sigma \tag{2.4}$$

where $\Phi(t, s)$ is the fundamental matrix associated with the homogeneous system

$$\dot{x}(t)=A_0x(t)+A_1x(t-h).$$

The function $\Gamma(s, \sigma)$ is given by

$$\begin{aligned} \Gamma(s,\sigma)=M(s,\sigma)B_0S^{-1}B_0'&+M(s,\sigma+h)B_1S^{-1}B_0'\chi_{\tau,T-h}(\sigma) \\ &+M(s,\sigma)B_1S^{-1}B_1'\chi_{\tau+h,T}(\sigma)+M(s,\sigma-h)B_0S^{-1}B_1'\chi_{\tau+h,T}(\sigma) \end{aligned} \tag{2.5}$$

where

$$\chi_{s,t}(\sigma)=\begin{cases}1 & \text{if } s\leqq\sigma\leqq t,\\ 0 & \text{otherwise.}\end{cases}$$

In particular, if we take τ to be t in (2.2), we obtain the optimal control as feedback of the pair $(x(t), M(x_t, u_t))$

$$\begin{aligned} u(t)=-S^{-1}B_0'&\Big[L(t,t,t)x(t)+\int_{-h}^{0}L(t,t+\theta+h,t)M(x_t,u_t)(\theta)\,d\theta\Big] \\ -S^{-1}B_1'&\Big[L(t+h,t,t)x(t)+\int_{-h}^{0}L(t+h,t+\theta+h,t)M(x_t,u_t)(\theta)\,d\theta\Big]. \end{aligned} \tag{2.6}$$

Of crucial importance for our later development is the expression for the optimal cost associated with the optimal control law (2.6), which was not obtained in [14]. Motivated by an idea of Datko in [4], we introduce the following functions:

For any initial conditions $\psi_i = (\phi_i, v_i)$ in $M^2 \times L^2[-h, 0]$, $i = 1, 2$, let

$$p_i(t) = \int_t^T \Phi'(\sigma, t)H'Hx_i^m(\sigma)\,d\sigma \tag{2.7}$$

where $x_i^m(\cdot)$ and $u_i^m(\cdot)$ are the optimal state and control trajectories associated with the initial condition ψ_i. It was shown in [14] that the function $p_i(t)$ satisfies the equation

$$p_i(t) = L(t, t, t)x_i^m(t) + \int_{-h}^0 L(t, t+\theta+h, t)[A_1x_i^m(t+\theta) + B_1u_i^m(t+\theta)]\,d\theta \tag{2.8}$$

and that the optimal control u_i^m satisfies

$$u_i^m(t) = -S^{-1}B_0'\int_t^T \Phi'(s, t)H'Hx_i^m(s)\,ds - S^{-1}B_1'\int_{t+h}^T \Phi'(s, t+h)H'Hx_i^m(s)\,ds. \tag{2.9}$$

Introduce the bilinear form $\langle\!\langle \cdot, \cdot \rangle\!\rangle$ on $(M^2 \times L^2[-h, 0]) \times (M^2 \times L^2[-h, 0])$ defined by

$$\begin{aligned} \langle\!\langle \psi_1, \psi_2 \rangle\!\rangle &\equiv \langle\!\langle (\phi_1, v_1), (\phi_2, v_2) \rangle\!\rangle \\ &= \phi_2'(0)p_1(0) + \int_{-h}^0 \phi_2'(s)A_1'p_1(s+h)\,ds + \int_{-h}^0 v_2'(s)B_1'p_1(s+h)\,ds. \end{aligned} \tag{2.10}$$

We then have the following lemma.

LEMMA 2.1. *The optimal cost J_T^m associated with the initial trajectory piece ϕ and initial control piece v is given by $\langle\!\langle (\phi, v), (\phi, v) \rangle\!\rangle$.*

Proof. Using (2.7) and the variation of constants formula for the solution of (2.1), we find

$$\begin{aligned} \phi_2'(0)p_1(0) &= \int_0^T \phi_2'(0)\Phi'(s, 0)H'Hx_1^m(s)\,ds \\ &= \int_0^T x_2^{m'}(s)H'Hx_1^m(s)\,ds - \int_0^T\int_{-h}^0 \phi_2'(\sigma)A_1'\Phi'(s, \sigma+h)H'Hx_1^m(s)\,d\sigma\,ds \\ &\quad - \int_0^T\int_0^s u_2^{m'}(\sigma)B_0'\Phi'(s, \sigma)H'Hx_1^m(s)\,d\sigma\,ds \\ &\quad - \int_0^T\int_0^s u_2^m(\sigma-h)B_1'\Phi'(s, \sigma)H'Hx_1^m(s)\,d\sigma\,ds. \end{aligned} \tag{2.11}$$

Using the fact that $\Phi(t, s) = 0$, $t < s$, and applying Fubini's theorem, the first three terms on the right hand side of (2.11) can be written as

$$\begin{aligned} &\int_0^T x_2^{m'}(s)H'Hx_1^m(s)\,ds - \int_{-h}^0\int_{\sigma+h}^T \phi_2'(\sigma)A_1'\Phi'(s, \sigma+h)H'Hx_1^m(s)\,ds\,d\sigma \\ &\qquad - \int_0^T\int_\sigma^T u_2^{m'}(\sigma)B_0'\Phi'(s, \sigma)H'Hx_1^m(s)\,ds\,d\sigma. \end{aligned} \tag{2.12}$$

For the fourth term on the right hand side of (2.11), we obtain

$$\int_0^T \int_0^s u_2^{m'}(\sigma-h)B_1'\Phi'(s,\sigma)H'Hx_1^m(s)\,d\sigma\,ds$$

$$(2.13)\qquad = \int_{-h}^0 \int_{\sigma+h}^T v_2'(\sigma)B_1'\Phi'(s,\sigma+h)H'Hx_1^m(s)\,ds\,d\sigma$$

$$+\int_0^T \int_{\sigma+h}^T u_2^{m'}(\sigma)B_1'\Phi'(s,\sigma+h)H'Hx_1^m(s)\,ds\,d\sigma$$

Combining (2.7), (2.9), (2.11)–(2.13), we get

$$(2.14)\qquad \begin{aligned}\phi_2'(0)p_1(0) &= \int_0^T x_2^{m'}(s)H'Hx_1^m(s)\,ds - \int_{-h}^0 \phi_2'(\sigma)A_1'p_1(\sigma+h)\,d\sigma \\ &\quad + \int_0^T u_2^{m'}(s)Su_1^m(s)\,ds - \int_{-h}^0 v_2'(\sigma)B_1'p_1(\sigma+h)\,d\sigma.\end{aligned}$$

On substituting this into (2.10), we see that

$$(2.15)\qquad \langle\!\langle(\phi_1, v_1), (\phi_2, v_2)\rangle\!\rangle = \int_0^T x_2^{m'}(s)H'Hx_1^m(s)\,ds + \int_0^T u_2^{m'}(s)Su_1^m(s)\,ds.$$

This proves the lemma.

Using the above lemma and (2.8), we establish the following theorem.

THEOREM 2.1. *For any $\tau < T$, and fixed initial trajectory and control functions ϕ_τ and v_τ defined on $[\tau-h, \tau]$ and $[\tau-h, \tau)$ respectively, the optimal cost J_T^m for the control problem*

$$J_T(\tau, \phi_\tau, v_\tau) = \int_\tau^T [x'(t)H'Hx(t) + u'(t)Su(t)]\,dt$$

is given by

$$(2.16)\qquad \begin{aligned} J_T^m(\tau, \phi_\tau, v_\tau) &= \phi'(\tau)L(\tau,\tau,\tau)\phi(\tau) + \int_{-h}^0 \phi'(\tau)L(\tau, \tau+\theta+h, \tau)M(\phi_\tau, v_\tau)(\theta)\,d\theta \\ &\quad + \int_{-h}^0 M(\phi_\tau, v_\tau)'(\theta)L(\tau+\theta+h, \tau, \tau)\phi(\tau)\,d\theta \\ &\quad + \int_{-h}^0\int_{-h}^0 M(\phi_\tau, v_\tau)'(\theta)L(\tau+\theta+h, \tau+\xi+h, \tau)M(\phi_\tau, v_\tau)(\xi)\,d\theta\,d\xi.\end{aligned}$$

Proof. By Lemma 2.1, the optimal cost $J_T^m(\tau, \phi_\tau, v)$ is given by

$$(2.17)\qquad \phi'(\tau)p(\tau) + \int_{\tau-h}^\tau \phi'(s)A_1'p(s+h)\,ds + \int_{\tau-h}^\tau v'(s)B_1'p(s+h)\,ds.$$

Substituting the expression (2.8) for p into (2.17) yields (2.16) after some straightforward computations.

Remark 2.1. The expression (2.16) for the optimal cost is similar in form to the one for systems with state delays only. The role of the "state" here is played by the pair $(x(t), M(x_t, u_t))$, and the optimal cost is a quadratic form on $(x(t), M(x_t, u_t))$.

Remark 2.2. We can define the bounded linear operator $\pi_T(\tau)$ mapping M^2 into M^2 as follows:

$$\pi_T(\tau)(\phi^0, \phi^1) = (k^0, k^1),$$

$$\begin{aligned} k^0 &= L(\tau, \tau, \tau)\phi^0 + \int_{-h}^{0} L(\tau, \tau+\theta+h, \tau)\phi^1(\theta)\, d\theta, \\ k^1(\theta) &= L(\tau+\theta+h, \tau, \tau)\phi^0 + \int_{-h}^{0} L(\tau+\theta+h, \tau+\xi+h, \tau)\phi^1(\xi)\, d\xi. \end{aligned} \tag{2.18}$$

Using the definition, the optimal cost $J_T^m(\tau, \phi_\tau, v_\tau)$ can also be written as

$$J_T^m(\tau, \phi_\tau, v_\tau) = \langle(\phi(\tau), M(\phi_\tau, v_\tau)), \pi_T(\tau)(\phi(\tau), M(\phi_\tau, v_\tau))\rangle_{M^2}. \tag{2.19}$$

We can also derive differential equations satisfied by the kernel $L(t, s, \tau)$. The following equations were given in [14].

$$\begin{aligned} \frac{\partial}{\partial\tau}L(t, s, \tau) &= [L(t, \tau+h, \tau)B_1S^{-1}B_0' + L(t, \tau, \tau)B_0S^{-1}B_0']L(\tau, s, \tau) \\ &\quad + [L(t, \tau+h, \tau)B_1S^{-1}B_1' + L(t, \tau, \tau)B_0S^{-1}B_1']L(\tau+h, s, \tau); \end{aligned} \tag{2.20}$$

$$\begin{aligned} \frac{\partial}{\partial\tau}L(t, \tau, \tau) &= -L(t, \tau, \tau)A_0 - L(t, \tau+h, \tau)A_1 + L(t, \tau, \tau)B_0S^{-1}B_0'L(\tau, \tau, \tau) \\ &\quad + L(t, \tau+h, \tau)B_1S^{-1}B_0'L(\tau, \tau, \tau) + L(t, \tau, \tau)B_0S^{-1}B_1'L(\tau+h, \tau, \tau) \\ &\quad + L(t, \tau+h, \tau)B_1S^{-1}B_1'L(\tau+h, \tau, \tau); \end{aligned} \tag{2.21}$$

$$\begin{aligned} \frac{d}{d\tau}L(\tau, \tau, \tau) &= -A_0'L(\tau, \tau, \tau) - L(\tau, \tau, \tau)A_0 - A_1'L(\tau+h, \tau, \tau) \\ &\quad - L(\tau, \tau+h, \tau)A_1 - H'H + L(\tau, \tau, \tau)B_0S^{-1}B_0'L(\tau, \tau, \tau) \\ &\quad + L(\tau, \tau+h, \tau)B_1S^{-1}B_0'L(\tau, \tau, \tau) + L(\tau, \tau, \tau)B_0S^{-1}B_1'L(\tau+h, \tau, \tau) \\ &\quad + L(\tau, \tau+h, \tau)B_1S^{-1}B_1'L(\tau+h, \tau, \tau). \end{aligned} \tag{2.22}$$

In fact, following the method given in [6], we can directly show that $L(t, s, \tau)$ also has partial derivatives with respect to t and s. This involves simply differentiating the integral equation (2.3) and using the fact that the function $M(t, s)$ has the following derivatives:

For $t \neq s$,

$$\frac{\partial}{\partial t}M(t, s) = -A_0'M(t, s) - A_1'M(t+h, s) - H'H\Phi(t, s), \tag{2.23}$$

$$\frac{\partial}{\partial s}M(t, s) = -M(t, s)A_0 - M(t, s+h)A_1 - \Phi'(s, t)H'H \tag{2.24}$$

and

$$\frac{d}{dt}M(t, t) = -A_0'M(t, t) - M(t, t)A_0 - A_1'M(t+h, t) - M(t+h, t)A_1 - H'H. \tag{2.25}$$

We find that

$$\frac{\partial}{\partial t}L(t, s, \tau) = -A_0'L(t, s, \tau) - A_1'L(t+h, s, \tau) - H'HS(t, s, \tau) \tag{2.26}$$

and

$$\frac{\partial}{\partial s}L(t,s,\tau)=-L(t,s,\tau)A_0-L(t,s+h,\tau)A_1-S'(s,t,\tau)H'H \tag{2.27}$$

where

$$\begin{aligned} S(t,s,\tau)=\Phi(t,s)&-\int_\tau^{\min(T,t)}\Phi(t,\sigma)B_0S^{-1}B_0'L(\sigma,s,\tau)\,d\sigma\\ &-\int_\tau^{\min(T-h,t-h)}\Phi(t,\sigma+h)B_1S^{-1}B_0'L(\sigma,s,\tau)\,d\sigma\\ &-\int_{\tau+h}^{\min(T,t+2h)}[\Phi(t,\sigma)B_1S^{-1}B_1'+\Phi(t+h,\sigma-h)B_0S^{-1}B_1']L(\sigma,s,\tau)\,d\sigma \end{aligned} \tag{2.28}$$

Define $L_t(\theta,\xi)=L(t+\theta,t+\xi,t)$, $0\leqq\theta$, $\xi\leqq h$. Then using (2.20), (2.26) and (2.27), we can derive the following set of differential equations for $L_t(\theta,\xi)$.

$$\begin{aligned} \frac{d}{dt}L_t(0,0)=-[A_0'-L_t(0,h)B_1S^{-1}B_0']L_t(0,0)&\\ -L_t(0,0)[A_0-B_0S^{-1}B_1'L_t(h,0)]-A_1'L_t(h,0)-L_t(0,h)A_1&\\ +L_t(0,0)B_0S^{-1}B_0'L_t(0,0)+L_t(0,h)B_1S^{-1}B_1'L_t(h,0)-H'H& \end{aligned} \tag{2.29}$$

$$\begin{aligned} \left(\frac{\partial}{\partial t}-\frac{\partial}{\partial \xi}\right)L_t(0,\xi)=-[A_0'-L_t(0,0)B_0S^{-1}B_0']L_t(0,\xi)&\\ +L_t(0,h)B_1S^{-1}B_0'L_t(0,\xi)+L_t(0,0)B_0S^{-1}B_1'L_t(h,\xi)&\\ +L_t(0,h)B_1S^{-1}B_1'L_t(h,\xi)-A_1'L_t(h,\xi)& \end{aligned} \tag{2.30}$$

$$\begin{aligned} \left(\frac{\partial}{\partial t}-\frac{\partial}{\partial \theta}-\frac{\partial}{\partial \xi}\right)L_t(\theta,\xi)=L_t(\theta,0)B_0S^{-1}B_0'L_t(0,\xi)+L_t(\theta,h)B_1S^{-1}B_0'L_t(0,\xi)&\\ +L_t(\theta,0)B_0S^{-1}B_1'L_t(h,\xi)+L_t(\theta,h)B_1S^{-1}B_1'L_t(h,\xi)& \end{aligned} \tag{2.31}$$

The optimal feedback control can now be written as

$$\begin{aligned} u(t)=-S^{-1}&[B_0'L_t(0,0)+B_1'L_t(h,0)]x(t)\\ -S^{-1}&\int_{-h}^{0}[B_0'L_t(0,\theta+h)+B_1'L_t(h,\theta+h)]M(x_t,u_t)(\theta)\,d\theta \end{aligned} \tag{2.32}$$

3. Infinite time quadratic optimal control. In this section, we study the infinite time quadratic control problem for linear systems with delays in the state and control. The system considered is again (2.1), and the cost functional to be optimized is given by

$$J_\infty(x_0,u_0,u)=\int_0^\infty[x'(t)H'Hx(t)+u'(t)Su(t)]\,dt.$$

For the infinite time problem to be well-posed, we need some condition which guarantees that for each initial condition (x_0,u_0), the cost $J_\infty(x_0,u_0,u)$ can be made finite by some choice of $u\in L^2[0,\infty)$. This involves the notion of stabilizability, which we shall now discuss.

Stabilizability for linear system with delays in the state only has been studied by Manitius and Triggiani (see [18] and references therein). In that situation, we have the system

$$\dot{x}(t)=A_0x(t)+A_1x(t-h)+Bu(t). \tag{3.1}$$

If a bounded linear operator $K: M^2\rightarrow R^m$ can be found such that on putting $u(t)=-Kx_t$ in (3.1), the resulting closed-loop system is L^2-stable, then the control u itself will be an element of $L^2[0, \infty)$, and the system (3.1) is said to be stabilizable. The point to notice here is that the feedback system

$$\dot{x}(t)=A_0x(t)+A_1x(t-h)-BKx_t \tag{3.2}$$

is again a delay differential equation. In the case of systems with delays also in the control, we have already seen in § 2 that the optimal control law for the finite time quadratic control problem has not only a feedback term on x_t, but also a feedback term on the past control u_t. As such, the optimal closed-loop system under the law (2.2) is no longer a single delay differential equation. The reasoning used in formulating the notion of stabilizability for (3.1) cannot be used here directly, and we have to extend it somewhat to accommodate this additional complication.

One possible formulation would be to say that (2.1) is stabilizable if there exist a constant matrix G_0, and measurable and essentially bounded functions $G_1(\cdot)$ and $G_2(\cdot)$, both defined on $[-h, 0]$ and taking values in $R^{m\times n}$ and $R^{m\times m}$ respectively, such that the control law

$$u(t)=G_0x(t)+\int_{-h}^{0} G_1(\theta)x(t+\theta)\,d\theta+\int_{-h}^{0} G_2(\theta)u(t+\theta)\,d\theta \tag{3.3}$$

stabilizes (2.1) in the sense that the resulting solution x and control u are both elements of $L_2[0,\infty)$. Indeed, this type of definition has been used by Olbrot [24] who also derived algebraic conditions for stabilizability in this sense. However, it turns out that for the study of duality between optimal control and filtering, this is not the most convenient definition. We shall instead adopt the following definition essentially given in [17].

Definition 3.1. The system (2.1) is said to be *stabilizable* if there exist a constant matrix K_0, a matrix function $K_1(\cdot)$ defined on $[-h, 0]$ which is measurable and essentially bounded, and a measurable matrix function $K_2(\cdot)$ defined on $[0, \infty)$ which is integrable on every compact subset of $[0, \infty)$ and which generates a Volterra integral operator mapping $L^2[0, \infty)$ into $L^2[0, \infty)$, such that the control law

$$u(t)=-\left[K_0x(t)+\int_{-h}^{0} K_1(\theta)x(t+\theta)\,d\theta+\int_{-h}^{t} K_2(t-s)x(s)\,ds\right] \tag{3.4}$$

gives rise to a state trajectory $x(\cdot)$ which is an element of $L^2[0, \infty)$, i.e., the system process x is L^2-stable. We then also say that (A_0, A_1, B_0, B_1) is stabilizable.

It is easy to see that controls of the form (3.4) include controls of the form (3.3). In fact, in the case that the initial control segment is zero, we can view (3.3) as a Volterra integral equation in u. Solving it for u yields precisely a feedback law solely in terms of x of the form (3.4). Notice that in contrast to systems with no control delays, we now require feedback not only on x_t, but also on the entire past history $x(s)$, $-h\leqq s\leqq t$. Note also that if the system is stabilizable in the sense of Definition 3.1, the resulting control is an element of $L^2[0, \infty)$. Thus, if the system is stabilizable, there exists a control $\bar{u}$ in $L_2[0, \infty)$ such that the corresponding cost $J(x_0, u_0, \bar{u})$ is finite.

We now return to the infinite time optimal control problem and give the following.

Theorem 3.1. *Assume that* (2.1) *is stabilizable. Then* $\lim_{T\rightarrow\infty}\langle(\phi(\tau), M(\phi_\tau, v_\tau)), \pi_T(\tau)(\phi(\tau), M(\phi_\tau, v_\tau))\rangle$ *exists, is finite and independent of* τ. *Furthermore the kernel*

$L_\tau(\cdot,\cdot)$ has the following convergence behavior:

$$
(3.5)\quad
\begin{array}{ll}
\text{(i)} & L_\tau(0,0)\to L_0; \\
\text{(ii)} & \left.\begin{array}{l} L_\tau(0,\cdot)A_1\to L_1(\cdot)A_1 \\ \textit{and } L_\tau(0,\cdot)B_1\to L_1(\cdot)B_1 \end{array}\right\} \textit{strongly in } L_2[-h,0]; \\
\text{(iii)} & \left.\begin{array}{l} A_1'L_\tau(\cdot,\cdot)A_1\to A_1'L_2(\cdot,\cdot)A_1 \\ A_1'L_\tau(\cdot,\cdot)B_1\to A_1'L_2(\cdot,\cdot)B_1 \\ B_1'L_\tau(\cdot,\cdot)B_1\to B_1'L_2(\cdot,\cdot)B_1 \end{array}\right\} \begin{array}{l}\textit{strongly with respect to each} \\ \textit{variable with the other variable} \\ \textit{fixed.}\end{array}
\end{array}
$$

Proof. We shall follow closely the arguments of [4] and [5]. By the assumption of stabilizability, there exist maps K_0, $K_1(\cdot)$ and $K_2(\cdot)$ such that $u(t)=K_0x(t)+\int_{-h}^{0}K_1(\theta)x(t+\theta)\,d\theta+\int_{-h}^{t}K_2(t-s)x(s)\,ds$ is a stabilizing control law in the sense of Definition 3.1. This implies that $J_\infty^m<\infty$. Since $\langle(\phi(\tau),M(\phi_\tau,v_\tau)),\pi_T(\tau)(\phi(\tau),M(\phi_\tau,v_\tau))\rangle$ is monotone in T, it follows that its limit as $T\to\infty$ exists, is finite and independent of τ (see the arguments in [4] and [5]). To show the convergence behavior of the kernel function, define the map $Z\colon M^2\times L^2\to M^2$ by

$$Z((\phi^0,\phi^1),v)=(\phi^0,M(\phi^1,v)).$$

By the above considerations, the operator $Z^*\pi_T(\tau)Z$ converges strongly to an operator $\tilde{\pi}$ on $M^2\times L^2$. Since for every $((\phi^0,\phi^1),v)$ in $M^2\times L^2$, we have

$$Z^*\pi_T(\tau)Z((\phi^0,\phi^1),v)=((\phi_T^0,\phi_T^1),v_T)$$

where

$$\phi_T^0=L_\tau(0,0)\phi^0+\int_{-h}^{0}L_\tau(0,\xi+h)M(\phi^1,v)(\xi)\,d\xi,$$

$$\phi_T^1(\theta)=A_1'L_\tau(\theta+h,0)\phi^0+\int_{-h}^{0}A_1'L_\tau(\theta+h,\xi+h)M(\phi^1,v)(\xi)\,d\xi,$$

$$v_T(\theta)=B_1'L_\tau(\theta+h,0)\phi^0+\int_{-h}^{0}B_1'L_\tau(\theta+h,\xi+h)M(\phi^1,v)(\xi)\,d\xi$$

the strong convergence of $Z^*\pi_T(\tau)Z$ implies that ϕ_T^0 converges, and that ϕ_T^1 and v_T converge strongly in $L^2[-h,0]$ for every $((\phi^0,\phi^1),v)$ in $M^2\times L^2$. This immediately implies the convergence properties stated in the theorem.

Consider now a sequence $T_n\to\infty$ as $n\to\infty$. From Theorem 3.1, we can find a subsequence T_{n_i} such that as $i\to\infty$,

$$\left.\begin{array}{l} L_\tau(0,\cdot)B_1\to L_1(\cdot)B_1 \\ B_1'L_\tau(h,\cdot)A_1\to B_1'L_{11}(\cdot)A_1 \\ B_1'L_\tau(h,\cdot)B_1\to B_1'L_{11}(\cdot)B_1 \end{array}\right\}\text{ pointwise a.e.}$$

Define $\lim_{i\to\infty}B_1'L_\tau(h,0)=B_1'L_{01}$. We then have:

THEOREM 3.2. *The optimal control law for the infinite time problem is given by*

$$
(3.6)\quad
\begin{aligned}
u(t)=&-S^{-1}(B_0'L_0+B_1'L_{01})x(t)\\
&-S^{-1}\int_{-h}^{0}[B_0'L_1(\theta+h)+B_1'L_{11}(\theta+h)]M(x_t,u_t)(\theta)\,d\theta.
\end{aligned}
$$

The optimal cost J_∞^m is given by

$$(3.7)\qquad \langle((x(0),x_0),u_0),\tilde{\pi}((x(0),x_0),u_0)\rangle.$$

Proof. Let x and u_x denote the solution of the closed-loop system defined on $[0, \infty)$ under the stationary law (3.6), and y and u_y denote the solution of the closed-loop equation defined on $[0, T_{n_i}]$ under the time-varying law (2.32).

Let

$$\tilde{x}(t)=y(t)-x(t) \quad \text{and} \quad \tilde{u}(t)=u_y(t)-u_x(t)$$

Then

$$\dot{\tilde{x}}(t)=A_0\tilde{x}(t)+A_1\tilde{x}(t-h)+B_0\tilde{u}(t)+B_1\tilde{u}(t-h), \qquad t\in[0, T_{n_i}], \tag{3.8}$$

with initial conditions

$$\tilde{x}(\theta)=0, \qquad \theta\in[-h, 0],$$
$$\tilde{u}(\theta)=0, \qquad \theta\in[-h, 0].$$

Applying the variations of constants formula, we obtain

$$\tilde{x}(t)=\int_0^t [\Phi(t-s)B_0+\Phi(t-s-h)B_1]\tilde{u}(s)\,ds. \tag{3.9}$$

Let

$$[B_0'L_t^i(0, \xi+h)+B_1'L_t^i(h, \xi+h)]=P_i(t, \xi)$$

and

$$B_0'L_t^i(0, 0)+B_1L_t^i(h, 0)=W_i(t)$$

where we have added the superscript i to indicate explicitly the dependence of $L_t(\cdot, \cdot)$ on T_{n_i}. Similarly, define $P(\xi)A_1=B_0'L_1(\xi+h)A_1+B_1'L_{11}(\xi+h)A_1$, $P(\xi)B_1=B_0'L_1(\xi+h)B_1+B_1'L_{11}(\xi+h)B_1$ and $W=B_0'L_0+B_1'L_{01}$. Then

$$\begin{aligned}\tilde{u}(t)=&-S^{-1}W_i(t)\tilde{x}(t)-S^{-1}[W_i(t)-W]x(t)-S^{-1}\int_{t-h}^t P_i(t, s-t)A_1\tilde{x}(s)\,ds\\ &-S^{-1}\int_{t-h}^t [P_i(t, s-t)-P(s-t)]A_1x(s)\,ds-S^{-1}\int_{t-h}^t P_i(t, s-t)B_1\tilde{u}(s)\,ds\\ &-S^{-1}\int_{t-h}^t [P_i(t, s-t)-P(s-t)]B_1u_x(s)\,ds.\end{aligned} \tag{3.10}$$

Let $q_i(t)=-S^{-1}[W_i(t)-W]x(t)-S^{-1}\int_{t-h}^t[P_i(t, s-t)-P(s-t)][A_1x(s)+B_1u_x(s)]\,ds$. We can then combine (3.9) and (3.10) into the equation

$$\begin{aligned}&\begin{bmatrix} I & 0\\ S^{-1}W_i(t) & I\end{bmatrix}\begin{bmatrix}\tilde{x}(t)\\ \tilde{u}(t)\end{bmatrix}\\ &=\begin{bmatrix}0\\ q_i(t)\end{bmatrix}+\int_0^t\begin{bmatrix}0 & \Phi(t-s)B_0+\Phi(t-s-h)B_1\\ -S^{-1}P_i(t, s-t)A_1 & -S^{-1}P_i(t, s-t)B_1\end{bmatrix}\begin{bmatrix}\tilde{x}(s)\\ \tilde{u}(s)\end{bmatrix}ds,\\ &\qquad\qquad t\in[0, T_{n_i}].\end{aligned} \tag{3.11}$$

Equation (3.11) is, with obvious notation, of the form

$$\xi_i(t)=\alpha_i(t)+\int_0^t R_i(t, s)\xi_i(s)\,ds. \tag{3.12}$$

Let $c_i(t)=\sup_{s\leq t}|R_i(t, s)|$. Then from the properties of $\Phi(t)$ and $P_i(t, \xi)$, we see that $c_i(t)$

is uniformly integrable over any finite interval. Now

$$|\xi_i(t)| \leqq |\alpha_i(t)| + c_i(t) \int_0^t |\xi_i(s)|\, ds. \tag{3.13}$$

By Gronwall's inequality, we obtain

$$|\xi_i(t)| \leqq |\alpha_i(t)| + \int_0^t |\alpha_i(s)| \exp\left[\int_s^t c_i(\sigma)\, d\sigma\right] ds.$$

Furthermore, $\alpha_i(t)$ is uniformly bounded in i, and for each $t < T_{n_i}$, $\alpha_i(t) \to 0$ as $i \to \infty$. By dominated convergence, we obtain that for each $t < T_{n_i}$, $|\xi_i(t)| \to 0$ as $i \to \infty$.

Next, let

$$g_i(t) = \begin{cases} y'(t)H'Hy(t) + u_y'(t)Su_y(t), & t \in [0, T_{n_i}] \\ 0 & \text{otherwise} \end{cases}$$

and

$$g(t) = x'(t)H'Hx(t) + u_x'(t)Su_x(t), \qquad t \in [0, \infty).$$

The above development shows that $g_i(t) \to g(t)$ pointwise in $[0, \infty)$ as $i \to \infty$. By Fatou's lemma

$$\begin{aligned} \int_0^\infty g(t)\, dt &\leqq \liminf_{i\to\infty} \int_0^\infty g_i(t)\, dt \\ &= \liminf_{i\to\infty} \langle (x(\theta), M(x_0, u_0)), \pi_{T_{n_i}}(0)(x(0), M(x_0, u_0))\rangle \\ &= \langle ((x(0), x_0), u_0), \tilde{\pi}((x(0), x_0), u_0)\rangle \end{aligned} \tag{3.14}$$

from previous considerations. On the other hand, for all $T \geqq 0$, optimality considerations give

$$\begin{aligned} &\int_0^T [x'(t)H'Hx(t) + u_x'(t)Su_x(t)]\, dt \\ &\qquad \geqq \langle (x(0), M(x_0, u_0)), \pi_T(0)(x(0), M(x_0, u_0))\rangle. \end{aligned} \tag{3.15}$$

The two inequalities (3.14), (3.15) establish that the optimal cost is given by $\langle ((x(0), x_0), u_0), \tilde{\pi}((x(0), x_0), u_0)\rangle$, and that this cost is attained with the law (3.6). The proof is completed.

Theorems 3.1 and 3.2 are generalizations of the results of [4] and [5] to systems with delays in the state and control. It gives the existence and characterization of the optimal control for the infinite time quadratic cost problem, but as usual, does not guarantee that the closed-loop system is stable. For that, we need the additional assumption of detectability of (A_0, A_1, H) (see [12] for the definition and further discussions).

THEOREM 3.3. *Let* (A_0, A_1, B_0, B_1) *be stabilizable and* (A_0, A_1, H) *be detectable. Then the closed-loop system generated by the control law* (3.6) *is* L^2*-stable.*

Proof. By detectability of (A_0, A_1, H) (see the definition in [12]), there exist matrices F_0 and F_1, and a measurable and essentially bounded matrix-valued function $F_2(\cdot)$, such that the system

$$\dot{y}(t) = A_0 y(t) + A_1 y(t-h) - F_0 Hy(t) - F_1 Hy(t-h) - \int_{-h}^0 F_2(\theta) Hy(t+\theta)\, d\theta \tag{3.16}$$

is exponentially stable. Now the solution x satisfies the following equation

$$\begin{aligned}\dot{x}(t)&=(A_0-F_0H)x(t)+(A_1-F_1H)x(t-h)-\int_{-h}^{0}F_2(\theta)Hx(t+\theta)\,d\theta\\ &\quad+F_0Hx(t)+F_1Hx(t-h)+\int_{-h}^{0}F_2(\theta)Hx(t+\theta)\,d\theta+B_0u(t)+B_1u(t-h).\end{aligned}\tag{3.17}$$

Letting $\Phi_F(\cdot)$ be the fundamental matrix associated with (3.16), we obtain, using the variation of constants formula,

$$\begin{aligned}|x(t)|\leqq{}&\left|\Phi_F(t)x_0(0)+\int_{-h}^{0}\Phi_F(t-s-h)[(A_1-F_1H)x_0(s)+B_1u_0(s)]\,ds\right.\\ &\left.+\int_{-h}^{0}\int_{\theta}^{0}\Phi_F(t-s+\theta)F_2(\theta)Hx_0(s)\,ds\,d\theta\right|\\ &+\int_{0}^{t}|\Phi_F(t-s)|\left|F_0Hx(s)+F_1Hx(s-h)\right.\\ &\left.+\int_{\neg h}^{0}F_2(\theta)Hx(s+\theta)\,d\theta\right|\,ds+\int_{0}^{t}|\Phi_F(t-s)B_0+\Phi_F(t-s-h)B_1|\,|u(s)|\,ds.\end{aligned}\tag{3.18}$$

Let $k_1=\max\{|F_0|, |F_1|, \operatorname{ess\,sup}_{-h\leqq\theta\leqq0}|F_2(\theta)|\}$. We obtain

$$\begin{aligned}|x(t)|\leqq{}&\left|\Phi_F(t)x_0(0)+\int_{-h}^{0}\Phi_F(t-s-h)[(A_1-F_1H)x_0(s)+B_1u_0(s)]\,ds\right.\\ &\left.+\int_{-h}^{0}\int_{\theta}^{0}\Phi_F(t-s+\theta)F_2(\theta)Hx_0(s)\,ds\,d\theta\right|\\ &+k_1\int_{0}^{t}|\Phi_F(t-s)|\left|Hx(s)+Hx(s-h)\right.\\ &\left.+\int_{-h}^{0}Hx(s+\theta)\,d\theta\right|\,ds+\int_{0}^{t}|\Phi_F(t-s)B_0+\Phi_F(t-s-h)B_1|\,|u(s)|\,ds.\end{aligned}\tag{3.19}$$

Since for u corresponding to the optimal control, we have $u\in L^2[0,\infty)$ and $Hx\in L^2[0,\infty)$, we have, by application of Young's inequality,

$$\begin{aligned}\left[\int_0^\infty|x(t)|^2\,dt\right]^{1/2}\leqq{}&k_2\left\{\int_0^\infty\left[\left|\Phi_F(t,0)x_0(0)+\int_{-h}^{0}\Phi_F(t-s-h)((A_1-F_1H)x_0(s)\right.\right.\right.\\ &\left.\left.\left.+B_1u_0(s))\,ds+\int_{-h}^{0}\int_{\theta}^{0}\Phi_F(t-s+\theta)F_2(\theta)Hx_0(s)\,ds\,d\theta\right|\right]^2dt\right\}^{1/2}\\ &+k_3\int_0^\infty|\Phi_F(t)|\,dt\left[\int_0^\infty\left|Hx(t)+Hx(t-h)+\int_{-h}^{0}Hx(t+\theta)\,d\theta\right|^2dt\right]^{1/2}\\ &+k_4\int_0^\infty|\Phi_F(t-s)B_0+\Phi_F(t-s-h)B_1|\,dt\left[\int_0^\infty|u(t)|^2\,dt\right]^{1/2}\end{aligned}\tag{3.20}$$

for appropriate constants k_2, k_3 and k_4. Since $|\Phi_F(t)|\leqq\beta\,e^{-\alpha t}$ for some $\beta\geqq1$, $\alpha>0$, we have that

$$\left[\int_0^\infty |x(t)|^2\,dt\right]^{1/2} < \infty$$

and the theorem is proved.

Theorems 3.1, 3.2 and 3.3 together describe the structure of the optimal control law for the infinite time quadratic control problem. Note that we have not shown that the gains of the optimal stationary control law can be obtained from the unique solution of a Riccati-like differential equation. Thus the justification of the existence and uniqueness of solutions to the steady state version of equations (2.29)–(2.31) is still an open question. This is in contrast to the case of linear systems with delays only in the state where complete results on the infinite time quadratic control problem are available [12].

4. Optimal filtering over a finite time interval for linear systems with delays in the state and observations. In this section, we shall discuss optimal filtering, in the minimum mean square error sense, for the stochastic delay system

$$\begin{aligned} &dx(t) = [A_0x(t) + A_1x(t-h)]\,dt + F\,dw(t), \\ &x(\theta) = 0, \qquad \theta \in [-h, 0], \end{aligned} \tag{4.1}$$

$$dz(t) = [C_0x(t) + C_1x(t-h)]\,dt + N\,dv(t). \tag{4.2}$$

Here $w(t)$ and $v(t)$ are independent standard Wiener processes in R^m and R^p respectively. The matrix N is assumed to be nonsingular. We shall also denote NN' by R. The nonsingularity of N then implies that $R > 0$.

For any $s \geqq t$, we denote the conditional expectation $E\{x(t)|z(\sigma), 0 \leqq \sigma \leqq s\}$ by $\hat{x}(t|s)$. Let the estimation error $x(t) - \hat{x}(t|s)$ be denoted by $\tilde{x}(t|s)$, and let $P(t, \sigma, s)$ denote the error covariance function $E[\tilde{x}(t|s)\tilde{x}(\sigma|s)']$ for $t \leqq s$, $\sigma \leqq s$. The filter equations characterizing $\hat{x}(t|t)$ have been derived independently in [10] and [12], and are stated below:

$$\begin{aligned} d\hat{x}(t|t) = {} & [A_0\hat{x}(t|t) + A_1\hat{x}(t-h|t)]\,dt + [P(t, t, t)C_0' + P(t, t-h, t)C_1']R^{-1} \\ & \cdot [dz(t) - C_0\hat{x}(t|t)\,dt - C_1\hat{x}(t-h|t)\,dt] \end{aligned} \tag{4.3}$$

$$\begin{aligned} \hat{x}(t-h|t) = \hat{x}(t-h\,|\,t-h) + \int_{t-h}^{t} & [P(t-h, s, s)C_0' + P(t-h, s-h, s)C_1']R^{-1} \\ & \cdot [dz(s) - C_0\hat{x}(s|s)\,ds - C_1\hat{x}(s-h|s)\,ds]. \end{aligned} \tag{4.4}$$

A set of differential equations was also derived for $P(t, \sigma, s)$ [10], [12]. The existence and uniqueness of solutions to (4.3)–(4.4) can be obtained using standard techniques, an outline of which is given in the Appendix.

It turns out that the equations for $P(t, \sigma, s)$ in [10], [12] are not quite convenient for the development of duality results for linear control and filtering in our problem. Thus, in the following, we will derive an integral equation for $P(t, \sigma, s)$ using the projection theorem characterization of $\hat{x}(t|s)$ [20].

We know that the process $\hat{x}(t|s)$ is Gaussian [10]. By the projection theorem, we can write

$$\hat{x}(t|s) = \int_0^s G_s(t, r)\,dz(r) \tag{4.5}$$

for some L^2 kernel $G_s(t, r)$. We now characterize the kernel $G_s(t, r)$. For any $0 \leqq \sigma \leqq s$,

$$\begin{aligned} 0 &= E[\tilde{x}(t|s)z'(\sigma)\} \\ &= \int_0^\sigma \{E[\tilde{x}(t|s)x'(\alpha)]C_0' + E[\tilde{x}(t|s)x'(\alpha - h)]C_1'\}\, d\alpha - E\{\tilde{x}(t|s)v'(\sigma)\}N'. \end{aligned} \tag{4.6}$$

But, by the projection theorem again, $E[\tilde{x}(t|s)x'(\alpha)] = E[\tilde{x}(t|s)\tilde{x}'(\alpha|s)] = P(t, \alpha, s)$ and

$$E\{\tilde{x}(t|s)v'(\sigma)\}N' = -\int_0^\sigma G_s(t, r)R\, dr. \tag{4.7}$$

This yields

$$G_s(t, \alpha) = [P(t, \alpha, s)C_0' + P(t, \alpha - h, s)C_1']R^{-1}. \tag{4.8}$$

Thus the error covariance function $P(t, \alpha, s)$ satisfies

$$\begin{aligned} P(t, \alpha, s) &= E\{\tilde{x}(t|s)x'(\alpha)\} \\ &= E[x(t)x'(\alpha)] - \int_0^s [P(t, \sigma, s)C_0' + P(t, \sigma - h, s)C_1']R^{-1} \\ &\qquad E\{[C_0x(\sigma) + C_1x(\sigma - h)]x'(\alpha)\}\, d\sigma. \end{aligned} \tag{4.9}$$

Let $\Sigma(\sigma, \alpha) = E[x(\sigma)x'(\alpha)]$. Then (4.9) can be written as

$$\begin{aligned} P(t, \alpha, s) = \Sigma(t, \alpha) &- \int_0^s P(t, \sigma, s)C_0'R^{-1}[C_0\Sigma(\sigma, \alpha) + C_1\Sigma(\sigma - h, \alpha)]\, d\sigma \\ &- \int_0^{s-h} P(t, \sigma, s)C_1'R^{-1}[C_0\Sigma(\sigma + h, \alpha) + C_1\Sigma(\sigma, \alpha)]\, d\sigma \end{aligned} \tag{4.10}$$

where we have used the assumption that $x(\theta) = 0$, $\theta \in [-h, 0]$. Define the kernel $K(\sigma, \alpha)$ by

$$\begin{aligned} K(\sigma, \alpha) = C_0'R^{-1}C_0\Sigma(\sigma, \alpha) &+ C_0'R^{-1}\Sigma(\sigma - h, \alpha)\chi_{h,s}(\sigma) \\ &+ C_1'R^{-1}C_0\Sigma(\sigma + h, \alpha)\chi_{0,s-h}(\sigma) + C_1'R^{-1}C_1\Sigma(\sigma, \alpha)\chi_{0,s-h}(\sigma). \end{aligned} \tag{4.11}$$

Then we see that $P(t, \alpha, s)$ satisfies the integral equation

$$P(t, \alpha, s) = \Sigma(t, \alpha) - \int_0^s P(t, \sigma, s)K(\sigma, \alpha)\, d\sigma. \tag{4.12}$$

Furthermore, since by definition,

$$P(t, \alpha, s) = P'(\alpha, t, s)$$

we get

$$\begin{aligned} P(t, \alpha, s) &= \Sigma'(\alpha, t) - \int_0^s K'(\sigma, t)P'(\alpha, \sigma, s)\, d\sigma \\ &= \Sigma(t, \alpha) - \int_0^s K'(\sigma, t)P(\sigma, \alpha, s)\, d\sigma. \end{aligned} \tag{4.13}$$

From the variation of constants formula for (4.1), it is easy to see that the matrix $\Sigma(t, \alpha)$ is given by

$$\Sigma(t, \alpha) = \int_0^{\min(t,\alpha)} \Phi(t, \sigma)FF'\Phi'(\alpha, \sigma)\, d\sigma. \tag{4.14}$$

For each fixed s and t, (4.13) is a Fredholm integral equation satisfied by $P(t, \alpha, s)$. Since the kernel $K(\sigma, \alpha)$ is readily seen to be positive, we can apply standard Fredholm theory to conclude that there exists a unique L^2 solution $P(t, \alpha, s)$. Moreover, $P(t, \alpha, s)$ is continuous, as discussed previously in [12], so that $P(t, \alpha, s)$ is also defined pointwise. Equations (4.3), (4.4) and (4.12)–(4.14) thus give an alternative characterization of the optimal filter.

Notice that the form of the integral equation defined by (4.12)–(4.14) is very similar to the one satisfied by $L(t, s, \tau)$, i.e. (2.3)–(2.4). In the next section, we will make this precise by giving a duality theorem relating optimal linear filtering to quadratic optimal control of a dual system.

5. A theorem on the duality between estimation and control. Let us consider the following system

$$\begin{aligned} &\dot{y}(t) = -A_0'y(t) - A_1'y(t+h) - C_0'u(t) - C_1'u(t+h), \qquad t \in [0, T], \\ (5.1) \qquad &y(\theta) = \phi(\theta), \qquad T \leqq \theta \leqq T+h, \\ &u(\theta) = v(\theta), \qquad T \leqq \theta \leqq T+h. \end{aligned}$$

The control problem is to optimize the functional

$$(5.2) \qquad J^a = \int_0^T [y'(t)FF'y(t) + u'(t)Ru(t)]\, dt$$

by some choice of $u \in L^2[0, T]$. The advanced system (5.1) runs backwards in time. By a change of variables, we can convert the problem into the standard control problem for linear systems with delays in the state and control.

Define $s = T - t$, $\bar{y}(s) = y(T-s) = y(t)$, $\bar{u}(s) = u(T-s) = u(t)$. Then we have

$$\begin{aligned} &\frac{d}{ds}\bar{y}(s) = A_0'\bar{y}(s) + A_1'\bar{y}(s-h) + C_0'\bar{u}(s) + C_1'\bar{u}(s-h), \\ (5.3) \qquad &\bar{y}(\theta) = \phi(T-\theta), \qquad \theta \in [-h, 0], \\ &\bar{u}(\theta) = v(T-\theta), \qquad \theta \in [-h, 0], \end{aligned}$$

$$(5.4) \qquad J^a = \int_0^T [\bar{y}'(s)FF'\bar{y}(s) + \bar{u}'(s)R\bar{u}(s)]\, ds.$$

The results of § 2 can then be applied directly to the problem defined by (5.3)–(5.4). This gives rise to a kernel $\bar{L}(s, \alpha, \tau)$ satisfying the integral equation

$$(5.5) \qquad \bar{L}(s, \alpha, \tau) = \bar{M}(s, \alpha) - \int_\tau^T \bar{L}(s, \beta, \tau)\bar{\Gamma}'(\alpha, \beta)\, d\beta$$

where

$$\begin{aligned} &\Gamma'(\alpha, \beta) = C_0'R^{-1}C_0\bar{M}(\beta, \alpha) + C_0'R^{-1}C_1\bar{M}(\beta+h, \alpha)\chi_{\tau, T-h}(\beta) \\ (5.6) \qquad &\qquad + C_1'R^{-1}C_0\bar{M}(\beta-h, \alpha)\chi_{\tau+h, T}(\beta) + C_1'R^{-1}C_1\bar{M}(\beta, \alpha)\chi_{\tau+h, T}(\beta). \end{aligned}$$

Here $\bar{M}(s, \alpha)$ is given by

$$(5.7) \qquad \bar{M}(s, \alpha) = \int_{\max(s,\alpha)}^T \bar{\Phi}'(\beta, s)FF'\Phi(\beta, \alpha)\, d\beta$$

where $\bar{\Phi}(s, \alpha)$ satisfies

$$\begin{aligned} &\frac{d}{ds}\bar{\Phi}(s, \sigma) = A_0'\bar{\Phi}(s, \sigma) + A_1'\bar{\Phi}(s-h, \sigma), \\ &\Phi(\sigma, \sigma) = I, \\ &\Phi(s, \sigma) = 0, \qquad s \in [\sigma - h, \sigma). \end{aligned} \tag{5.8}$$

To transform the equations back to the original time variable, we let $s = T - t$, $\alpha = T - \gamma$, $\tau = T - \sigma$, and define the function

$$\Lambda(t, \gamma, \sigma) = \bar{L}(T-t, T-\gamma, T-\sigma). \tag{5.9}$$

Then the function $\Lambda(t, \gamma, \sigma)$ satisfies the integral equation

$$\begin{aligned} \Lambda(t, \gamma, \sigma) &= \bar{M}(T-t, T-\gamma) - \int_{T-\sigma}^{T} \Lambda(t, T-\beta, \sigma)\bar{\Gamma}'(T-\gamma, \beta)\, d\beta \\ &= \bar{M}(T-t, T-\gamma) - \int_0^{\sigma} \Lambda(t, \phi, \sigma)\bar{\Gamma}'(T-\gamma, T-\phi)\, d\phi. \end{aligned} \tag{5.10}$$

On the other hand,

$$\begin{aligned} \bar{M}(T-t, T-\gamma) &= \int_{\max(T-t, T-\gamma)}^{T} \bar{\Phi}'(\beta, T-t)FF'\bar{\Phi}(\beta, T-\gamma)\, d\beta \\ &= \int_0^{\min(t,\gamma)} \bar{\Phi}'(T-\beta, T-t)FF'\bar{\Phi}(T-\beta, T-\gamma)\, d\beta. \end{aligned} \tag{5.11}$$

Now recall that the homogeneous equation

$$\dot{x}(t) = A_0x(t) + A_1x(t-h) \tag{5.12}$$

has as its adjoint equation

$$\dot{y}(t) = -A_0'y(t) - A_1'y(t+h). \tag{5.13}$$

The fundamental matrix $Y(t, \sigma)$ to (5.13) satisfies

$$\begin{aligned} &\frac{d}{dt}Y(t, \sigma) = -A_0'Y(t, \sigma) - A_1'Y(t+h, \sigma), \\ &Y(\sigma, \sigma) = I, \\ &Y(t, \sigma) = 0 \quad \text{if } t > \sigma. \end{aligned} \tag{5.14}$$

It is easily verified that $Y(t, \sigma)$ and $\bar{\Phi}(T-t, T-\sigma)$ satisfy the same differential equation and boundary condition. By uniqueness, we conclude that

$$Y(t, \sigma) = \bar{\Phi}(T-t, t-\sigma). \tag{5.15}$$

Hence

$$\begin{aligned} \bar{M}(T-t, T-\gamma) &= \int_0^{\min(t,\gamma)} Y'(\beta, t)FF'Y(\beta, \gamma)\, d\beta \\ &= \int_0^{\min(t,\gamma)} \Phi(t, \beta)FF'\Phi'(\gamma, \beta)\, d\beta \end{aligned} \tag{5.16}$$

where $\Phi(t, \beta)$ is the fundamental matrix associated with (5.12). We have here used the well-known result that $\Phi(t, \beta) = Y'(\beta, t)$ (see, for example, [19]).

We can now state the duality theorem.

THEOREM 5.1. *The function* $\Lambda(t, \gamma, \sigma)$ *in the dual control problem defined by* (5.10) *is the same function as the error covariance function* $P(t, \gamma, \sigma)$ *defined by* (4.12).

Proof. From equations (4.14) and (5.16), we see that

$$\bar{M}(T-t, T-\gamma) = \Sigma(t, \gamma).$$

It is now readily verified that

$$\bar{\Gamma}'(T-\gamma, T-\beta) = K(\beta, \gamma)$$

where $K(\beta, \gamma)$ is as defined in (4.11). Hence $\Lambda(t, \gamma, \sigma)$ satisfies the integral equation

$$\Lambda(t, \gamma, \sigma) = \Sigma(t, \gamma) - \int_0^{\sigma} \Lambda(t, \phi, \sigma) K(\phi, \gamma)\, d\phi$$

which is the same integral equation as the one satisfied by $P(t, \gamma, \sigma)$ (equation (4.12)). By uniqueness of solutions to the Fredholm integral equation, we obtain the conclusion of the theorem.

In addition to its theoretical interest, the above theorem enables us to apply the infinite-time optimal control results to the filtering problem. Various versions of the duality theorem have been given in the literature [15], [19], but Theorem 5.1 seems to us to be the most convenient for the purpose at hand.

6. Detectability and adjoint systems. In § 3, we have obtained the infinite-time control results under the hypotheses of stabilizability of (A_0, A_1, B_0, B_1) and detectability of (A_0, A_1, H). In order to apply these results to the filtering problem, we study the notion of detectability which is dual to the stabilizability of (A_0, A_1, B_0, B_1). We make the following.

DEFINITION 6.1. Consider the system

$$\dot{x}(t) = A_0 x(t) + A_1 x(t-h), \qquad x(\theta) = x_0(\theta), \tag{6.1}$$

$$z(t) = C_0 x(t) + C_1 x(t-h). \tag{6.2}$$

We say that the system (6.1)–(6.2) is detectable if its dual system

$$y(t) = -A_0' y(t) - A_1' y(t+h) - C_0' u(t) - C_1' u(t+h) \tag{6.3}$$

is stabilizable. We then also say that (A_0, A_1, C_0, C_1) is detectable.

Note that if we construct a stabilizing control law for (6.3), we will obtain for the closed-loop system a linear Volterra integrodifferential equation running backwards in time. For our study of filter stability, we need to use the fact that such a system is in a certain sense adjoint to a linear Volterra integrodifferential system running forwards in time. The precise notion of adjoint systems of linear Volterra integrodifferential equations is given in the following proposition.

PROPOSITION 6.1. *For* $K(\cdot)$ *a locally* L^1 *matrix function, and* $\tau < T$, *the systems*

$$\dot{x}(t) = A_0 x(t) + A_1 x(t-h) + \int_{\tau}^{t} K(t-s) x(s)\, ds, \qquad \tau \leqq t < \infty, \tag{6.4}$$

and

$$\dot{y}(t) = -A_0' y(t) - A_1' y(t+h) - \int_{t}^{T} K'(s-t) y(s)\, ds, \qquad -\infty < t \leqq T, \tag{6.5}$$

are adjoints of each other in the sense that if $x(t)$ *and* $y(t)$ *are any solutions of* (6.4) *and* (6.5), *then for any* $t \in [\tau, T]$, *the form* $\langle y^t, x^t, t\rangle$ *defined by*

$$\begin{aligned}\langle y^t, x^t, t\rangle &= y'(t)x(t) + \int_t^{t+h} y'(s)A_1 x(s-h)\,ds \\ &\quad + \int_t^T y'(s)\int_\tau^s K(s-\sigma)x(\sigma)\,d\sigma\,ds + \int_\tau^t \int_\sigma^T y'(s)K(s-\sigma)\,ds\,x(\sigma)\,d\sigma\end{aligned} \tag{6.6}$$

is constant in t. Furthermore, if $V(s, t)$ *is the fundamental matrix satisfying* (6.5) *in s with* $V(t, t) = I$, $V(s, t) = 0$ *for* $s > t$, *and* $U(t, s)$ *is the fundamental matrix satisfying* (6.4) *in t with* $U(s, s) = I$, $U(t, s) = 0$ *for* $t < s$, *then* $V'(s, t) = U(t, s)$.

Proof. We compute

$$\begin{aligned}\frac{d}{dt}\langle y^t, x^t, t\rangle = &\Big[-A_0'y(t) - A_1'y(t+h) - \int_t^T K'(s-t)y(s)\,ds\Big)' x(t) \\ &+ y'(t)\Big[A_0 x(t) + A_1 x(t-h) + \int_\tau^t K(t-s)x(s)\,ds\Big] \\ &+ y'(t+h)A_1 x(t) - y'(t)A_1 x(t-h) \\ &- y'(t)\int_\tau^t K(t-\sigma)x(\sigma)\,d\sigma + \int_t^T y'(s)K(s-t)\,ds\,x(t) \\ = &\,0.\end{aligned}$$

Now for equations defined on $(-\infty, t+h]$, we have that

$$\langle V^s, x^s, s\rangle = \langle V^t, x^t, t\rangle$$

where, for each fixed t, we are considering $V(s, t)$ as a function of s. This yields

$$\begin{aligned}V'(s, t)x(s) &+ \int_s^{s+h} V'(\sigma, t)A_1 x(\sigma-h)\,d\sigma + \int_s^t V'(\sigma, t)\int_s^\sigma K(\sigma-\alpha)x(\alpha)\,d\alpha\,d\sigma \\ &= x(t) + \int_s^t \int_\sigma^t V'(\alpha, t)K(\alpha-\sigma)\,d\alpha\,x(\sigma)\,d\sigma.\end{aligned} \tag{6.7}$$

After simplification, (6.7) becomes

$$x(t) = V'(s, t)x(s) + \int_s^{s+h} V'(s, t)A_1 x(\sigma-h)\,d\sigma. \tag{6.8}$$

If we now take $x(t) = U(t, s)$ for fixed s, we get

$$U(t, s) = V'(s, t) \tag{6.9}$$

as claimed.

Remark 6.1. From the constancy of (6.4) and (6.5), we see that $U(t, s)$ is a function of $t-s$ and $V(s, t)$ is a function of $s-t$. Hence if $U(t, 0)$ is an element of $L^2[0, \infty)$, $V(0, t)$ is an element of $L^2(-\infty, 0]$. Using (6.8) we see that this implies that if the solution $y(t)$ to (6.5) is an element of $L_2(-\infty, 0]$, for each initial condition, the solution $x(t)$ to (6.4) is an element of $L_2[0, \infty)$, for each initial condition.

Remark 6.2. Using Proposition 6.1, we can show that (6.1)–(6.2) is detectable if and only if there exist a constant matrix N_0, a measurable and essentially bounded matrix function $N_1(\cdot)$ defined on $[-h, 0]$, and a matrix function $N_2(\cdot)$ defined on $[0, \infty)$, which are measurable and integrable on every compact subset of $[0, \infty)$, and which generate a Volterra integral operator mapping $L^2[0, \infty)$ into $L^2[0, \infty)$, such that the system

$$(6.10)\quad \dot{x}(t) = A_0x(t) + A_1x(t-h) - N_0z(t) - \int_{-h}^{0} N_1(\theta)z(t+\theta)\,d\theta - \int_{-h}^{t} N_2(t-s)z(s)\,ds$$

is L^2-stable.

For the sake of completeness, we mention here that the dual notion of detectability of (A_0, A_1, H) is the stabilizability of the dual control system defined by (A_0', A_1', H'). Note that the absence of delays in the observations leads to a dual control system without delays in the control. The reader is referred to [12] for additional discussions.

7. The existence and stability of the stationary filter. Our objective in this section is to show that under suitable hypothesis, the filtering error covariance matrix converges to a constant matrix. This is of great importance in the evaluation of filter performance, for then we can assess the accuracy of the filter by examining the limiting value of the error covariance. Moreover, we will show that the stationary filter thus obtained is L^2-stable. The technique is to relate the filtering problem to a dual control problem, and use the infinite-time optimal control results. First, we state what we mean by filter stability.

DEFINITION 7.1. The optimal filter defined by (4.3) and (4.4) is said to be *stable* if the following two conditions are satisfied:

(i) The estimation error covariance $P(t, t, t)$ is bounded on $[0, \infty)$ and that $\lim_{t\to\infty} P(t, t, t)$ exists and is finite.

(ii) The estimation error $e(t|t) = x(t) - \hat{x}(t|t)$ associated with the stationary version of the optimal filter satisfies an equation whose homogeneous solution is L^2-stable.

Remark 7.1. The boundedness condition for $P(t, t, t)$ should certainly be required in any definition of filter stability. Since the error process is Gaussian, the convergence of $P(t, t, t)$ implies that the distribution associated with $e(t|t)$ converges to a constant Gaussian distribution. If the second requirement is also satisfied, then in view of the linearity of the system, any disturbance on the error process with trajectories in $L^2[0, \infty)$ will still generate an error process with trajectories in $L^2[0, \infty)$.

THEOREM 7.1. *Consider the stochastic delay system*

$$(7.1)\quad \begin{aligned} dx(t) &= [A_0x(t) + A_1x(t-h)]\,dt + F\,dw(t),\\ x(\theta) &= 0, \qquad \theta \in [-h, 0], \end{aligned}$$

$$(7.2)\quad dz(t) = [C_0x(t) + C_1x(t-h)]\,dt + N\,dv(t).$$

Assume that the system (6.1)–(6.2) *is detectable. Then the error covariance matrix* $P(t+\xi, t+\theta, t)$, $-h \leqq \xi, \theta \leqq 0$, *has the following asymptotic behavior as* $t \to \infty$:[1]

$$P(t, t, t) \to P_0;$$

$$\left.\begin{aligned} P(t, t+\theta, t)A_1' &\to P_1(\theta)A_1' \\ P(t, t+\theta, t)C_1' &\to P_1(\theta)C_1' \end{aligned}\right\} \text{ strongly in } L^2[-h, 0];$$

[1] We shall abuse notation and write $P(t, t+\theta, t)A_1' \to P_1(\theta)A_1'$ in $L^2[-h, 0]$, etc., to mean $P(t, t+\cdot, t)A_1' \to P_1(\cdot)A_1'$ in $L^2[-h, 0]$, etc.

$$\left.\begin{aligned} &A_1P(t+\theta, t+\xi, t)A_1' \to A_1P_2(\theta, \xi)A_1' \\ &C_1P(t+\theta, t+\xi, t)A_1' \to C_1P_2(\theta, \xi)A_1' \\ &C_1P(t+\theta, t+\xi, t)C_1' \to C_1P_2(\theta, \xi)C_1' \end{aligned}\right\} \textit{strongly with respect to each variable with the other variable fixed.}$$

Proof. Consider the dual control problem defined by (5.1)–(5.2). We know by Theorem 5.1 that $\Lambda(t, \gamma, \sigma) = P(t, \gamma, \sigma)$. Also by the hypothesis of detectability, the dual system (5.1) is stabilizable. Hence by Theorem 3.1, we have that as $T \to \infty$,

$$\bar{L}(\tau, \tau, \tau) \to \bar{L}_0$$

$$\left.\begin{aligned} &\bar{L}(\tau, \tau+\sigma+h, \tau)A_1' \to \bar{L}_1(\sigma+h)A_1' \\ &\bar{L}(\tau, \tau+\sigma+h, \tau)C_1' \to \bar{L}_1(\sigma+h)C_1' \end{aligned}\right\} \text{strongly in } L^2[-h, 0]$$

$$\left.\begin{aligned} &A_1\bar{L}(\alpha+\tau+h, \sigma+\tau+h, \tau)A_1' \to A_1\bar{L}_2(\alpha+h, \sigma+h)A_1' \\ &A_1\bar{L}(\alpha+\tau+h, \sigma+\tau+h, \tau)C_1' \to A_1\bar{L}_2(\alpha+h, \sigma+h)C_1' \\ &C_1\bar{L}(\alpha+\tau+h, \sigma+\tau+h, \tau)C_1' \to C_1\bar{L}_2(\alpha+h, \sigma+h)C_1' \end{aligned}\right\} \text{strongly in each variable with the other variable fixed.}$$

But since $\bar{L}(T-t, T-\gamma, T-\sigma) = \Lambda(t, \gamma, \sigma) = P(t, \gamma, \sigma)$, a simple change of variables shows that these convergences yield the asymptotic behavior for $P(t+\xi, t+\theta, t)$ as stated.

Henceforth, we shall denote $P_1(-h)$ by P_{01}' and $P_2(-h, \theta)$ by $P_{11}(\theta)$.

We now examine the infinite-time dual control problem. We have the system

$$\begin{aligned} &\dot{y}(t) = -A_0'y(t) - A_1'y(t+h) - C_0'u(t) - C_1'u(t+h), \qquad t \in (-\infty, 0], \\ &y(\theta) = \phi(\theta), \qquad 0 \leqq \theta \leqq h, \\ &u(\theta) = v(\theta), \qquad 0 < \theta \leqq h, \end{aligned} \tag{7.3}$$

with cost functional

$$J_\infty^a = \int_{-\infty}^0 [y'(t)FF'y(t) + u'(t)Ru(t)]\, dt. \tag{7.4}$$

Using Theorem 5.1 again, we obtain that the optimal control law is given by

$$\begin{aligned} u(t) = &-R^{-1}C_0\left[P_0y(t) + \int_0^h P_1(\alpha-h)(A_1'y(t+\alpha) + C_1'u(t+\alpha))\, d\alpha\right] \\ &-R^{-1}C_1\left[P_{01}y(t) + \int_0^h P_{11}(\alpha-h)(A_1'y(t+\alpha) + C_1'u(t+\alpha))\, d\alpha\right]. \end{aligned} \tag{7.5}$$

On the other hand, the stationary filter is given by the equations

$$\begin{aligned} d\hat{x}(t|t) = &[A_0\hat{x}(t|t) + A_1\hat{x}(t-h|t)]\, dt + [P_0C_0' + P_{01}'C_1']R^{-1} \\ &\cdot [dz(t) - C_0\hat{x}(t|t)\, dt - C_1\hat{x}(t-h|t)\, dt] \end{aligned} \tag{7.6}$$

$$\begin{aligned} A_1\hat{x}(t-h|t) = A_1\hat{x}(t-h|t-h) + &\int_{t-h}^t A_1[P_1'(t-s-h)C_0' + P_{11}'(t-s-h)C_1'] \\ &\cdot R^{-1}[dz(s) - C_0\hat{x}(s|s)\, ds - C_1\hat{x}(s-h|s)\, ds], \end{aligned} \tag{7.7}$$

$$C_1\hat{x}(t-h|t)=C_1\hat{x}(t-h|t-h)+\int_{t-h}^{t} C_1[P_1'(t-s-h)C_0'+P_{11}'(t-s-h)C_1'] \tag{7.8}$$
$$\cdot R^{-1}[dz(s)-C_0\hat{x}(s|s)\,ds-C_1\hat{x}(s-h|s)\,ds].$$

We will now show that the stationary control system and the stationary filter are in fact adjoint systems of each other in the sense of Proposition 6.1. Once this is established, we can appeal to Theorem 3.2 to investigate the stability of the stationary filter. To this end, we first express the control law (7.5) solely in terms of feedback on y. Let

$$D_1(t)=\begin{cases} C_0P_1(t)+C_1P_{11}(t), & t\in[-h,0],\\ 0 & \text{otherwise,}\end{cases}$$

and let

$$F(t)=-R^{-1}D_1(t-h)C_1', \qquad t\in[0,h].$$

Also denote $(C_0P_0+C_1P_{01})$ by D_0. Then we can write

$$u(t)=q(t)+\int_t^h F(t-\sigma)u(\sigma)\,d\sigma \tag{7.9}$$

where

$$q(t)=-R^{-1}D_0y(t)-R^{-1}\int_t^{t+h} D_1(\sigma-t-h)A_1'y(\sigma)\,d\sigma. \tag{7.10}$$

We can view (7.9) as a Volterra integral equation in u. Define $H(t)$ to be the resolvent kernel associated with $F(t)$ [22], i.e.,

$$H(t)=F(t)+\int_t^0 F(t-\sigma)H(\sigma)\,d\sigma. \tag{7.11}$$

Then we can solve for $u(t)$ in (7.9) to give

$$u(t)=q(t)+\int_t^h H(t-\sigma)q(\sigma)\,d\sigma. \tag{7.12}$$

Substituting (7.10) into (7.12), we obtain

$$\begin{aligned} u(t)&=-R^{-1}D_0y(t)-R^{-1}\int_t^{t+h} D_1(\sigma-t-h)A_1'y(\sigma)\,d\sigma-\int_t^h H(t-\sigma)R^{-1}D_0y(\sigma)\,d\sigma\\ &\quad-\int_t^h H(t-\sigma)R^{-1}\int_\sigma^{\sigma+h} D_1(\phi-\sigma-h)A_1'y(\phi)\,d\phi\,d\sigma\\ &=-R^{-1}D_0y(t)-R^{-1}\int_t^{t+h} D_1(\sigma-t-h)A_1'y(\sigma)\,d\sigma-\int_t^h H(t-\sigma)R^{-1}D_0y(\sigma)\,d\sigma\\ &\quad-\int_t^h\int_{\max(t,\phi-h)}^{\min(t+h,\phi)} H(t-\sigma)R^{-1}D_1(\phi-\sigma-h)\,d\sigma A_1'y(\phi)\,d\phi. \end{aligned} \tag{7.13}$$

Substituting (7.13) into (7.5), we obtain the closed-loop system as

$$\begin{aligned}\dot{y}(t) = &-(A_0' - C_0'R^{-1}D_0)y(t) - (A_1' - C_1'R^{-1}D_0)y(t+h)\\ &+ C_0'R^{-1}\int_t^{t+h} D_1(\sigma - t - h)A_1'y(\sigma)\,d\sigma + C_0'\int_t^h H(t-\sigma)R^{-1}D_0y(\sigma)\,d\sigma\\ &+ C_0'\int_t^h\int_{\max(t,\phi-h)}^{\min(t+h,\phi)} H(t-\sigma)R^{-1}D_1(\phi-\sigma-h)\,d\sigma\,A_1'y(\phi)\,d\phi\\ &+ C_1'R^{-1}\int_{t+h}^{t+2h} D_1(\sigma-t-2h)A_1'y(\sigma)\,d\sigma + C_1'\int_{t+h}^h H(t+h-\sigma)R^{-1}D_0y(\sigma)\,d\sigma\\ &+ C_1'\int_{t+h}^h\int_{\max(t+h,\phi-h)}^{\min(t+2h,\phi)} H(t+h-\sigma)R^{-1}D_1(\phi-\sigma-h)\,d\sigma\,A_1'y(\phi)\,d\phi.\end{aligned} \tag{7.14}$$

Next, we give the equation satisfied by the estimation error $\hat{e}(t|t)$ associated with the stationary filter. In particular, in our study of filter stability, we are concerned with the homogeneous part of the equation satisfied by $e(t|t)$. For ease of notation, we write $x(t)$ for $e(t|t)$ and $\xi(t)$ for $e(t-h|t)$ in the homogeneous equation for $e(t|t)$. We then obtain the following integral equation from (7.8):

$$\begin{aligned}C_1\xi(t) = C_1x(t-h) &- \int_{t-h}^t C_1D_1'(t-s-h)R^{-1}C_0x(s)\,ds\\ &- \int_{t-h}^t C_1D_1'(t-s-h)R^{-1}C_1\xi(s)\,ds.\end{aligned} \tag{7.15}$$

Let $W(t) = -C_1D_1'(t-h)R^{-1}$ and define $N(t)$, the Volterra resolvent kernel of $W(t)$, by

$$\begin{aligned}N(t) &= W(t) + \int_0^t W(t-s)N(s)\,ds\\ &= W(t) + \int_0^t N(t-s)W(s)\,ds.\end{aligned} \tag{7.16}$$

Then on solving (7.15), we get

$$\begin{aligned}C_1\xi(t) = C_1x(t-h) &- \int_{t-h}^t C_1D_1'(t-s-h)R^{-1}C_0x(s)\,ds\\ &+ \int_{-h}^t N(t-s)C_1x(s-h)\,ds\\ &- \int_{-h}^t N(t-s)\int_{s-h}^s C_1D_1'(s-\sigma-h)R^{-1}C_0x(\sigma)\,d\sigma\,ds\\ = C_1x(t-h) &+ \int_{-h}^t N(t-s)C_1x(s-h)\,ds\\ &- \int_{-h}^t \Bigg[C_1D_1'(t-\sigma-h)R^{-1}\\ &+ \int_{\max(\sigma,t-h)}^{\min(t,\sigma+h)} N(t-s)C_1D_1'(s-\sigma-h)R^{-1}\,ds\Bigg]C_0x(\sigma)\,d\sigma.\end{aligned} \tag{7.17}$$

But by (7.16),

$$\begin{aligned}C_1D_1'(t-\sigma-h)R^{-1}&+\int_{\max(\sigma,t-h)}^{\min(t,\sigma+h)} N(t-s)C_1D_1'(s-\sigma-h)R^{-1}\,ds\\ &=-W(t-\sigma)-\int_{\max(\sigma,t-h)}^{\min(t,\sigma+h)} N(t-s)W(s-\sigma)\,ds\\ &=-N(t-\sigma).\end{aligned}$$

Thus

$$C_1\xi(t)=C_1x(t-h)+\int_{-h}^{t} N(t-s)[C_0x(s)+C_1x(s-h)]\,ds. \tag{7.18}$$

Combining (7.6), (7.7) and (7.18), we see that the error process $e(t|t)$ satisfies the equation

$$\begin{aligned}\dot{x}(t)=(A_0-D_0'R^{-1}C_0)x(t)&+(A_1-D_0'R^{-1}C_1)x(t-h)\\ &-A_1\int_{t-h}^{t} D_1'(t-s-h)R^{-1}[C_0x(s)+C_1x(s-h)]\,ds\\ &-A_1\int_{-h}^{t}\int_{\max(\sigma,t-h)}^{\min(t,\sigma+h)} D_1'(t-s-h)R^{-1}N(s-\sigma)\,ds[C_0x(\sigma)+C_1x(\sigma-h)]\,d\sigma\\ &-D_0'R^{-1}\int_{-h}^{t} N(t-s)[C_0x(s)+C_1x(s-h)]\,ds.\end{aligned} \tag{7.19}$$

One can also readily see that the homogeneous part of (7.6)–(7.8) (i.e., with $z(s)\equiv 0$) gives rise to an equation in $\hat{x}(t|t)$ identical to (7.19). Thus we shall call (7.19) the homogeneous stationary filter.

We can now give:

THEOREM 7.2. *The stationary closed-loop system* (7.14) *is the adjoint to the homogeneous stationary filter* (7.19) *in the sense of Proposition* 6.1.

Proof. First notice that on comparing (7.11) and (7.16), we obtain

$$N(t)=H'(-t). \tag{7.20}$$

Now the adjoint to (7.19) is given by

$$\begin{aligned}\dot{x}(t)=&-(A_0'-C_0'R^{-1}D_0)x(t)-(A_1'-C_1'R^{-1}D_0)x(t+h)\\ &+\int_{t}^{t+h} C_0'R^{-1}D_1(s-t-h)A_1'x(s)\,ds+\int_{t+h}^{t+2h} C_1'R^{-1}D_1(s-t-2h)A_1'x(s)\,ds\\ &+\int_{t}^{h}\int_{\max(t,\sigma-h)}^{\min(\sigma,t+h)} C_0'N'(s-t)R^{-1}D_1(\sigma-s-h)\,ds\,A_1'x(\sigma)\,d\sigma\\ &+\int_{t+h}^{h}\int_{\max(t+h,\sigma-h)}^{\min(\sigma,t+2h)} C_1'N'(s-t)R^{-1}D_1(\sigma-s-h)A_1'x(\sigma)\,d\sigma\\ &+\int_{t}^{h} C_0'N'(s-t)R^{-1}D_0x(s)\,ds+\int_{t+h}^{h} C_1'N(s-t-h)R^{-1}D_0x(s)\,ds.\end{aligned} \tag{7.21}$$

Using (7.20), we immediately verify that (7.21) and (7.14) are identical. This completes the proof of the theorem.

We have now all the ingredients to prove the central result of the paper, the stability of the stationary filter.

THEOREM 7.3. *Assume that* (A_0, A_1, C_0, C_1) *is detectable and* (A_0, A_1, F) *is stabilizable. Then the stationary filter exists and is* L^2*-stable.*

Proof. By Theorem 7.1, we need only to prove the L^2-stability. The hypothesis gives that (A_0', A_1', F') is detectable. Thus by Theorem 3.2, the optimal stationary law for the dual control problem gives rise to an L^2-stable closed-loop system. Thus, the solutions y of (7.14) are elements of $L_2(-\infty, 0]$. By Proposition 6.1 and Remark 6.1, the solutions of the system adjoint to (7.14) are thus elements of $L_2[0, \infty)$. By Theorem 7.2, the system adjoint to (7.14) is precisely the homogeneous stationary filter (7.19). This implies that the solutions x of (7.19) are elements of $L_2[0, \infty)$ and the theorem is proved.

8. Stochastic control of linear systems with delays in the state, control, and observations. In this section, we put together the theory we have developed for quadratic optimal control and linear filtering to obtain a stable stochastic control scheme. The system we are interested in is of the form

$$\begin{aligned} &dx(t) = [A_0x(t) + A_1x(t-h)]\,dt + [B_0u(t) + B_1u(t-h)]\,dt + F\,dw(t) \\ (8.1)\qquad &x(\theta) = 0, \qquad \theta \in [-h, 0], \\ &u(\theta) = 0, \qquad \theta \in [-h, 0), \end{aligned}$$

$$(8.2)\qquad dz(t) = [C_0x(t) + C_1x(t-h)]\,dt + N\,dv(t).$$

The finite time stochastic control problem has been studied by Lindquist [8], who proved a version of the Separation Theorem in the case where there are no delays in the control, although his methods can be extended to cover that case also (see the remarks in [16]). Here, we provide the final missing element in the linear-quadratic-Gaussian theory for general delay systems: the asymptotic behavior of the stochastic control law given by the cascade of the stationary filter with the stationary deterministic feedback law.

THEOREM 8.1. *Suppose* (A_0, A_1, B_0, B_1) *is stabilizable,* (A_0, A_1, H) *is detectable,* (A_0, A_1, C_0, C_1) *is detectable, and* (A_0, A_1, F) *is stabilizable. Then the control law given by*

$$\begin{aligned} &u(t) = -S^{-1}[B_0'L_0 + B_1'L_{01}]\hat{x}(t|t) \\ (8.3)\qquad &\quad -S^{-1}\int_{-h}^{0} [B_0'L_1(\sigma+h) + B_1'L_{11}(\sigma+h)][A_1\hat{x}(t+\sigma|t) + B_1u(t+\sigma)]\,d\sigma \end{aligned}$$

where L_0*,* L_{01}*,* $L_1(\cdot)$ *and* $L_{11}(\cdot)$ *are obtained as in Theorem* 3.2, *and* $\hat{x}(t+\sigma|t)$, $\sigma \in [-h, 0]$ *is generated by the stationary filter* (7.6)–(7.8), *gives rise to an* L^2*-stable closed-loop system.*

Proof. Let the estimation error $x(s) - \hat{x}(s|t)$, $s \leqq t$, be denoted by $e(s|t)$. Then the error process satisfies the following set of equations:

$$\begin{aligned} de(t|t) &= [A_0e(t|t)+A_1e(t-h|t)-D_0'R^{-1}C_0e(t|t)-D_0'R^{-1}C_1e(t-h|t)]\,dt \\ &\quad +F\,dw(t)-D_0'R^{-1}N\,dv(t), \end{aligned} \tag{8.4}$$

$$\begin{aligned} A_1e(t-h|t) &= A_1e(t-h|t-h)+A_1\int_{t-h}^{t} D_1'(t-\sigma-h)R^{-1}C_0e(\sigma|\sigma)\,d\sigma \\ &\quad +A_1\int_{t-h}^{t} D_1'(t-\sigma-h)R^{-1}C_1e(\sigma-h|\sigma)\,d\sigma \\ &\quad +A_1\int_{t-h}^{t} D_1'(t-\sigma-h)R^{-1}N\,dv(\sigma), \end{aligned} \tag{8.5}$$

$$\begin{aligned} C_1e(t-h|t) &= C_1e(t-h|t-h)+\int_{t-h}^{t} C_1D_1'(t-\sigma-h)R^{-1}C_0e(\sigma|\sigma)\,d\sigma \\ &\quad +\int_{t-h}^{t} C_1D_1'(t-\sigma-h)R^{-1}C_1e(\sigma-h|\sigma)\,d\sigma \\ &\quad +C_1\int_{t-h}^{t} D_1'(t-\sigma-h)R^{-1}N\,dv(\sigma). \end{aligned} \tag{8.6}$$

The homogeneous part of (8.4)–(8.6) is precisely the same as those of (7.6)–(7.8). By Theorem 7.3, we find that the error process $e(t|t)$ is L^2-stable. Now the control law can be rewritten as

$$\begin{aligned} u(t) &= -S^{-1}(B_0'L_0+B_1'L_{01})x(t)-S^{-1}\int_{-h}^{0}[B_0'L_1(\sigma+h)+B_1'L_{11}(\sigma+h)] \\ &\quad \cdot[A_1x(t+\sigma)+B_1u(t+\sigma)]\,d\sigma \\ &\quad +S^{-1}(B_0'L_0+B_1'L_{01})e(t|t)+S^{-1}\int_{-h}^{0}[B_0'L_1(\sigma+h)+B_1'L_{11}(\sigma+h)] \\ &\quad \cdot A_1e(t+\sigma|t)\,d\sigma. \end{aligned} \tag{8.7}$$

Since the error process is decoupled from the x system, we find that the composite system given by (8.1), (8.4)–(8.6) and (8.7) is L^2-stable if the part involving only the x and u processes (with $e(s|t)\equiv 0$) is. As Theorem 3.2 guarantees the L^2-stability of the x and u processes, the theorem is proved.

Appendix. We give here an outline of the proof of existence and uniqueness of solutions to equations (4.3)–(4.4).

Let

$$\eta(t)=\hat{x}(t|t), \qquad \xi(t)=\hat{x}(t-h|t),$$

$$K_1(t)=[P(t,t,t)C_0'+P(t,t-h,t)C_1']R^{-1},$$

$$K_2(t,s)=[P(t-h,s,s)C_0'+P(t-h,s-h,s)C_1']R^{-1}$$

$$z_1(t)=\int_0^t K_1(s)\,dz(s), \qquad z_2(t)=\int_{t-h}^{t} K_2(t,s)\,dz(s).$$

Note that the processes z_1 and z_2 have continuous sample paths almost surely.

Consider the interval $[0, T]$, and let $\phi(t)=\hat{x}(t|t)=Ex(t)$, $-h\leqq t\leqq 0$. Equations (4.3)–(4.4) are equivalent to the following integral equations.

$$\eta(t)=\phi(0)+\int_0^t [A_0-K_1(s)C_0]\eta(s)\,ds+\int_0^t [A_1-K_1(s)C_1]\xi(s)\,ds+z_1(t) \tag{A.1}$$

$$\xi(t)=\begin{cases} \phi(t-h)-\displaystyle\int_{t-h}^0 K_2(t-s)C_0\phi(s)\,ds-\int_0^t K_2(t-s)C_0\eta(s)\,ds \\ \qquad -\displaystyle\int_{t-h}^t K_2(t,s)C_1\xi(s)\,ds+z_2(t) & \text{for } 0\leqq t\leqq h \\ \phi(0)+\displaystyle\int_0^{t-h} [A_0-K_1(s)C_0]\eta(s)\,ds+\int_0^{t-h} [A_1-K_1(s)C_1]\xi(s)\,ds \\ \quad +z_1(t-h)-\displaystyle\int_{t-h}^t K_2(t,s)C_0\eta(s)\,ds-\int_{t-h}^t K_2(t,s)C_1\xi(s)\,ds+z_2(t) \\ & \text{for } h\leqq t\leqq T. \end{cases} \tag{A.2}$$

Now assume the function ϕ is continuous (in (4.3)–(4.4), $\phi\equiv 0$), and for each $\lambda>0$, define a norm on the Banach space $C[0, T]\times C[0, T]$ by

$$\|(\eta, \xi)\|_\lambda=\max\left\{\sup_{0\leqq t\leqq T} e^{-\lambda t}|\eta(t)|;\ \sup_{0\leqq t\leqq T} e^{-\lambda t}|\xi(t)|\right\}$$

Define also the map F: $C[0, T]\times C[0, T]\rightarrow C[0, T]\times C[0, T]$ by

$$F(\eta, \xi)=(\eta_1, \xi_1)$$

where $\eta_1(t)$ is the function on the right hand side of (A.1) and $\xi_1(t)$ is the function on the right hand side of (A.2). Straightforward estimates show that for λ sufficiently large, F is a contraction mapping. We can therefore conclude the existence and uniqueness of solutions to (4.3)–(4.4).

Acknowledgement. I wish to thank A. Manitius and the reviewers for many helpful comments.

REFERENCES

[1] H. J. Kushner and D. I. Barnea, *On the control of a linear functional differential equation with quadratic cost*, this Journal, 8 (1970), pp. 257–272.

[2] Y. Alekal, P. Brunovsky, D. H. Chyung and E. B. Lee, *The quadratic problem for systems with time delays*, IEEE Trans. Automatic Control, AC-16 (1971), pp. 673–688.

[3] M. C. Delfour and S. K. Mitter, *Controllability, observability, and optimal feedback control of affine hereditary differential systems*, this Journal, 10 (1972), pp. 298–328.

[4] R. Datko, *Unconstrained control problems with quadratic cost*, this Journal, 11 (1973), pp. 32–52.

[5] M. C. Delfour, C. McCalla and S. K. Mitter, *Stability and infinite-time quadratic cost problem for linear hereditary differential systems*, this Journal, 13 (1975), pp. 48–88.

[6] A. Manitius, *Optimal control of hereditary systems*, Centre de Recherches Mathématiques Report CRM-472, Université de Montréal, Montreal, P.Q., Canada, December, 1974.

[7] H. Kwakernaak, *Optimal filtering in linear systems with time delays*, IEEE Trans. Automatic Control, AC-12 (1967), pp. 169–173.

[8] A. Lindquist, *Optimal control of linear stochastic systems with applications to time lag systems*, Information Sci., 5 (1973), pp. 81–126.

[9] S. K. MITTER AND R. B. VINTER, *Filtering for linear stochastic hereditary differential systems*, Intern. Symp. Control Theory, Numerical Methods, and Computer Systems Modelling, IRIA (Rocquencourt, France), June 1974.

[10] A. BAGCHI, *A martingale approach to state estimation in delay-differential systems*, J. Math. Anal. Appl., 56 (1976), pp. 195–210.

[11] R. H. KWONG AND A. S. WILLSKY, *Optimal filtering and filter stability of linear stochastic delay systems*, IEEE Trans. Automatic Control, AC-22 (1977), pp. 196–201.

[12] ———, *Estimation and filter stability of stochastic delay systems*, this Journal, 16 (1978), pp. 660–681.

[13] R. B. VINTER, *Filter stability for stochastic evolution equations*, this Journal, 15 (1977), pp. 465–485.

[14] H. KOIVO AND E. B. LEE, *Controller synthesis for linear systems with retarded state and control variables and quadratic cost*, Automatica, 8 (1972), pp. 203–208.

[15] A. LINDQUIST, *A theorem on duality between estimation and control for linear stochastic systems with time delay*, J. Math. Anal. Appl., 37 (1972), pp. 516–536.

[16] ———, *On feedback control of linear stochastic systems*, this Journal, 11 (1973), pp. 323–343.

[17] R. H. KWONG, *The linear quadratic Gaussian problem for systems with delays in the state, control, and observations*, in Proc. 14th Allerton Conf. Circuit and System Theory (Sept. 29–Oct. 1, 1976).

[18] A. MANITIUS, *Controllability, observability, and stabilizability of retarded systems*, Proc. 1976 IEEE Conference on Decision and Control (Clearwater Beach, Florida), pp. 752–758.

[19] J. K. HALE, *Functional Differential Equations*, Springer-Verlag, New York, 1971.

[20] J. L. DOOB, *Stochastic Processes*, John Wiley, New York, 1953.

[21] R. H. KWONG, *Structural properties and estimation of delay systems*, Electronic Systems Lab. Report ESL-R-614, Massachusetts Institute of Technology, Cambridge, MA, September 1975.

[22] R. K. MILLER, *Nonlinear Volterra Integral Equations*, W. A. Benjamin, Menlo Park, 1971.

[23] A. ICHIKAWA, *Optimal control and filtering of evolution equations with delay in control and observation*, Report No. 53, Control Theory Centre, University of Warwick, June 1977.

[24] A. W. OLBROT, *Stabilizability, detectability, and spectrum assignment for linear autonomous systems with general time delays*, IEEE Trans. Automatic Control, AC-23 (1978), pp. 887–890.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 1, January 1980

0363-0129/80/1801-0005$01.00/0

EQUIVALENCE OF LINEAR COMPLEMENTARITY PROBLEMS AND LINEAR PROGRAMS IN VECTOR LATTICE HILBERT SPACES*

C. W. CRYER† AND M. A. H. DEMPSTER‡

Abstract. Let X be a vector lattice Hilbert space with dual X^*. Let M be a continuous linear mapping of X onto X^*. Let $p, q \in X^*$ with $p > 0$. We consider the relationship between the *linear complementarity problem*: Find $x \in X$ such that $x \geqq 0$, $Mx + q \geqq 0$, $\langle x, Mx + q\rangle = 0$, and the *linear programming problem*: Find $x \in X$ which minimizes $\langle x, p\rangle$ subject to $x \geqq 0$, $Mx + q \geqq 0$. For the problem of a cavitating journal bearing, which is used as an example, the linear program requires the minimization of a linear functional which is proportional to the load borne by the bearing.

1. Introduction. The *linear complementarity problem* in real n-dimensional Euclidean space R^n is: Find $x \in R^n$ such that $x \geqq 0$, $Mx + q \geqq 0$, and $x^T(Mx + q) = 0$, where M is a given real $n \times n$-matrix and q is a given vector in R^n. The *linear programming problem in* R^n is: Find $x \in R^n$ which minimizes p^Tx subject to $x \geqq 0$ and $Mx + q \geqq 0$, where M is a given real $n \times n$ matrix and p and q are given vectors in R^n.

Mangasarian (1976) showed that, under certain conditions, each solution of the linear programming problem in R^n is a solution of the linear complementarity problem in R^n. Mangasarian (1977), (1979) has subsequently extended this work. Related work is due to Cottle and Veinott (1972), Moré (1971), Tamir (1973), Cottle, Golub, and Sacher (1978), Cottle and Pang (1976), (1978), Pang (1976), (1977), (1978), and Cottle (1976).

Quite independently, and often not very explicitly, the relationship between certain infinite-dimensional linear programming problems and linear complementarity problems has been noted (Moreau (1971), Durand (1968), Lewy and Stampacchia (1969), Stampacchia (1965), Lions and Stampacchia (1967)).

Here, we consider extensions of some of the results of Mangasarian to infinite-dimensional spaces. Apart from their intrinsic value, our results provide useful ways of interpreting, analyzing, and solving linear programming problems and linear complementarity problems arising in physical situations.

The following abbreviations are used: LP (linear program), LD (dual linear program), LE (least element problem), LC (linear complementarity problem), VI (variational inequality), and UM (unilateral minimization problem).

2. Preliminaries. X denotes a real Hilbert space with norm $\|\cdot\|$ and dual $Y = X^*$. The evaluation of a continuous linear functional $l \in X^*$ at a point $x \in X$ is denoted by $\langle x, l\rangle$.

It is assumed that X is partially ordered by a vector ordering $\geqq$. Let

$$P = \{x \in X: x \geqq 0\}.$$

Then (Kelley and Namioka (1976, p. 224)) P is a convex cone in X with vertex at the origin: that is, $P + P \subset P$ and $\lambda P \subset P$ for all nonnegative real λ. We assume that P is closed. $x \geqq y$ iff $x - y \geqq 0$, that is, $x - y \in P$.

* Received by the editors November 7, 1978, and in revised form May 7, 1979.

† Computer Science Department, University of Wisconsin, Madison, Wisconsin. The work of this author was supported by the National Science Foundation under Grant MCS77-26732 and the University of Wisconsin, Madison, Graduate School Research Committee. The support facilities were provided by the United States Army under Contract DAAG29-75-C-0024.

‡ Balliol College, University of Oxford, Oxford, England, OX1 2JD.

The dual cone $P^* \subset X^*$ is defined by

$$P^* = \{x^* \in X^* : \langle x, x^* \rangle \geqq 0 \text{ for all } x \in P\}. \tag{2.1}$$

We write $x^* \geqq 0$ if $x^* \in P^*$. Since P is closed it follows from the Hahn–Banach theorem that $x \geqq 0$ iff $\langle x, x^* \rangle \geqq 0$ for all $x^* \in P^*$.

It is also assumed that X is a vector lattice (Kelley and Namioka (1976, p. 229)). That is, for all $x, y \in X$, there exists a unique element $\sup(x, y) \in X$ such that $\sup(x, y) \geqq x$ and $\sup(x, y) \geqq y$; furthermore, if $z \in X$ satisfies $z \geqq x$ and $z \geqq y$ then $z \geqq \sup(x, y)$. The assumption that X is a vector lattice has the following consequences. For all $x, y \in X$ there exists a unique element $\inf(x, y)$ such that $x \geqq \inf(x, y)$ and $y \geqq \inf(x, y)$; furthermore, if $z \in X$ satisfies $z \leqq x$ and $z \leqq y$ then $z \leqq \inf(x, y)$. If $x \geqq y$ then $\sup(x, y) = x$, and if $y \geqq x$ then $\sup(x, y) = y$; since $\sup(x, y)$ is unique, it follows that if $x \geqq y$ and $y \geqq x$ then $x = y$. For every $x \in X$, $x = \sup(x, 0) - \inf(x, 0)$ so that $X = P - P$. If $0 = x + y$ where $x, y \in P$ then $x = y = 0$; thus 0 is an extreme point of P, that is, P is a pointed cone.

$M: X \to Y = X^*$ denotes a continuous linear transformation with adjoint $M^*: Y^* \to X^*$ defined by

$$\langle x, M^* y^* \rangle = \langle Mx, y^* \rangle. \tag{2.2}$$

Associated with M we have the continuous bilinear operator $a: X \times X \to R^1$ defined by

$$a(v, u) = \langle u, Mv \rangle; \tag{2.3}$$

a is *symmetric* if $a(u, v) = a(v, u)$, and *coercive* if

$$a(x, x) \geqq \alpha \|x\|^2, \tag{2.4}$$

for some real strictly positive constant α and all $x \in X$.

We will sometimes impose the following conditions upon a and M:

Condition S. If $r \in X^*$ and $u, v \in X$ are such that

$$a(u, \psi) \geqq \langle \psi, r \rangle \text{ and } a(v, \psi) \geqq \langle \psi, r \rangle \text{ for all } \psi \in P,$$

and if $w = \inf(u, v)$, then

$$a(w, \psi) \geqq \langle \psi, r \rangle \text{ for all } \psi \in P.$$

Condition Z. If $u, v \in P$ satisfy $\inf(u, v) = 0$,

then $a(u, v) \leqq 0$.

p and q denote elements of X^*. We assume frequently that $p \in P^*$. We will sometime assume that p is *strictly positive*, that is, if $x \in P$ and $\langle x, p \rangle = 0$ then $x = 0$.

Since

$$a(u, \psi) - \langle \psi, r \rangle = \langle \psi, Mu - r \rangle, \tag{2.5}$$

Condition S may be rewritten as follows: if $Mu \geqq r$, $Mv \geqq r$, and $w = \inf(u, v)$, then $Mw \geqq r$. If $-M$ is the Laplacian operator ∇^2 and $r = 0$ then $Mu \geqq 0$ means, in an appropriate sense, that $-u$ is subharmonic. In this case, Condition S reduces to the well-known fact that the infimum of two superharmonic functions is superharmonic. There is, therefore, a close connection between some of the present results and the theory of subharmonic functions (Rado (1972), Brelot (1945), (1965), Stampacchia (1965), Littman (1963), Moreau (1971)).

Condition S is equivalent to the condition that the set

$$S=\{u\in X\colon a(u,\psi)\geqq\langle\psi,r\rangle \text{ for all } \psi\in P\}$$

is a meet semi-lattice for all $r\in X$.

In the case when M is a square matrix, Condition Z is equivalent to the requirement that the off-diagonal elements of M be nonpositive—that is, that M is a Z-matrix (Fiedler and Ptak (1962)). There is, therefore, also a close connection between some of the present results and the theory of M-matrices and Z-matrices (Poole and Boullion (1974), Plemmons (1976)). Condition Z was implicitly used by Stampacchia (1965, p. 151) with the conclusion $a(u,v)\leqq 0$ replaced by $a(u,v)=0$.

Conditions S and Z are not equivalent because, as shown in § 2.1, the necessary and sufficient conditions for Conditions S and Z are not equivalent in the case of matrices. However, we do have the following

THEOREM 2.1. *Let a be coercive and satisfy Condition Z. Then a satisfies Condition S.*

Proof. Let $u,v\in X$ and $r\in X^*$ satisfy $a(u,\psi)\geqq\langle\psi,r\rangle$ and $a(v,\psi)\geqq\langle\psi,r\rangle$ for all $\psi\geqq 0$. We wish to show that if $w=\inf(u,v)$ then $a(w,\psi)\geqq\langle\psi,r\rangle$ for all $\psi\geqq 0$. To do so, we modify an argument of Stampacchia (1965, p. 205).

Introduce the set $U\subset X$ which consists of all $\zeta\in X$ satisfying $\zeta\geqq w$. $U=P+w$ is closed and convex. From the fundamental theorem on variational inequalities (Stampacchia (1964)) we know that there exists $\eta\in U$ such that

$$a(\eta,z-\eta)\geqq\langle z-\eta,r\rangle, \tag{2.6}$$

for all $z\in U$. In particular, choosing $z=\eta+\psi$, we see that $a(\eta,\psi)\geqq\langle\psi,r\rangle$ for all $\psi\geqq 0$. The theorem will therefore be proved if we can show that $\eta=w$.

Set $\zeta=\inf(\eta,u)\in U$. From (2.6) with $z=\zeta$,

$$a(\eta,\zeta-\eta)\geqq\langle\zeta-\eta,r\rangle. \tag{2.7}$$

On the other hand we know that $\eta-\zeta\geqq 0$, $u-\zeta\geqq 0$, and $\inf(\eta-\zeta,\ u-\zeta)=\inf(\eta,u)-\zeta=0$. Invoking Condition Z we see that

$$\begin{aligned} a(\zeta,\zeta-\eta)&=a(u,\zeta-\eta)+a(\zeta-u,\zeta-\eta)\\ &=a(u,\zeta-\eta)+a(u-\zeta,\eta-\zeta)\\ &\leqq a(u,\zeta-\eta)\\ &\leqq\langle\zeta-\eta,r\rangle. \end{aligned} \tag{2.8}$$

Combining (2.7) and (2.8) we find that

$$a(\zeta-\eta,\zeta-\eta)\leqq 0.$$

Since a is coercive it follows that $\zeta=\inf(u,\eta)=\eta$, so that $\eta\leqq u$. Similarly $\eta\leqq v$. Hence $\eta\leqq\inf(u,v)=w$. But $\eta\in U$ so that $\eta\geqq w$. We conclude that $\eta=w$. □

In the case when M is a real square matrix, it is readily shown from Theorem 2.1.1 below that if a is coercive and satisfies Condition S then a satisfies Condition Z. We do not know whether this is true in general.

We now give two examples of spaces and operators fitting into the above framework.

2.1. Example 1. Let $X=Y=X^*=Y^*=R^n$; $M=(m_{ij})$ an $n\times n$ real matrix; and $p=(p_i)$, $q=(q_i)$, n-vectors. Let P be the set of vectors in R^n with nonnegative components, so that P is closed and $P^*=P$.

P has the additional important property that it has nonempty interior.

Clearly, p is strictly positive iff $p_i > 0$ for all i.

It is readily seen that Condition Z is satisfied iff $m_{ij} \leqq 0$ for $i \neq j$ (that is, M is a Z-matrix).

THEOREM 2.1.1. *Condition S is satisfied iff every row of M has at most one strictly positive coefficient, that is M^T is pre-Leontief* (Cottle and Veinott (1972, p. 244)).

Proof. We first observe that M^T is pre-Leontief iff each row k of the inequality $Mu \geqq r$ can be written in the form

$$(*) \qquad c_s u_s \geqq r_k + \sum_{j=1}^{n} d_j u_j,$$

where $d_j \geqq 0$ for all j; $c_s \geqq 0$; and where the dependence upon k of c_s and d_j has been suppressed.

First let us assume that M^T is pre-Leontief and that $Mu \geqq r$, $Mv \geqq r$. Then inequality $(*)$ holds for u, and a similar inequality holds for v. Since $d_j \geqq 0$ we have that if $w = \inf(u, v)$ then

$$\begin{aligned} c_s w_s &= \inf(c_s u_s, c_s v_s) \\ &\geqq r_k + \sum_{j=1}^{n} d_j \inf(u_j, v_j) = r_k + \sum_{j=1}^{n} d_j w_j, \end{aligned}$$

so that Condition S is satisfied.

Now let us assume that Condition S holds but that M^T is not pre-Leontief. Then there is a row k of M with at least two positive coefficients, m_{ks} and m_{kt} say. Thus, the kth row of the inequality $Mu \geqq r$ takes the form

$$m_{ks} u_s \geqq r_k - m_{kt} u_t - \sum_{\substack{j=1 \\ j \neq s,t}}^{n} m_{kj} u_j,$$

and a similar inequality holds for v. Set $u_s = 1/m_{ks}$, $u_t = -1/m_{kt}$, $v_s = 2u_s$, $v_t = 2u_t$, and $u_j = v_j = 0$ otherwise. Finally, set $w = \inf(u, v)$ and

$$r_j = \min\left[\sum_{l=1}^{n} m_{jl} u_l, \sum_{l=1}^{n} m_{jl} v_l\right],$$

for all j. Then $Mu \geqq r$, and $Mv \geqq r$. But,

$$\begin{aligned} \sum_{l=1}^{n} m_{kl} w_l &= m_{ks} u_s + m_{kt} v_t + \sum_{\substack{j=1 \\ j \neq s,t}}^{n} m_{kj} u_j, \\ &= r_k - 1, \end{aligned}$$

so that the inequality $Mw \geqq r$ does not hold. □

Remark. Cottle and Veinott (1972, Corollary 2, p. 245) show that M^T is pre-Leontief iff the set

$$X_r^+ = \{u \in R^n : Mu \geqq r, u \geqq 0\}$$

has a least element for each r such that X_r^+ is nonempty. This leads to an alternate proof of Theorem 2.1.1.

In conclusion, we note that if the problems in Example 2 below are discretized, the resulting finite difference or finite element matrices usually satisfy Conditions S and Z.

2.2. Example 2. Let $X = H_0^1(\Omega)$, where Ω is a bounded domain (open connected set) in R^n, and $H_0^1(\Omega)$ is the Sobolev space of once-differentiable functions vanishing on $\partial\Omega$ (Adams (1975)). Then $Y = X^* = H^{-1}(\Omega)$. Let M be the linear self-adjoint operator,

$$(Mu)(t) = -\sum \frac{\partial}{\partial t_j}\left(a_{ij}(t)\frac{\partial u}{\partial t_i}(t)\right), \qquad t \in \Omega, \tag{2.2.1}$$

with coefficients $a_{ij}(t)$ which are continuously differentiable, where the indices i and j are summed from 1 to n. It is assumed that $-M$ is uniformly elliptic, so that

$$\sum a_{ij}(t)\xi_i\xi_j \geqq \alpha|\xi|^2, \qquad t \in \Omega, \tag{2.2.2}$$

for all $\xi = (\xi_i) \in R^n$, and some constant $\alpha > 0$.

Every $x \in H_0^1$ has a representation as a measurable function $x(t)$, and any two such representations of x differ only on a set of measure zero. We write $x \geqq 0$ if $x(t) \geqq 0$ a.e. (almost everywhere). $P = \{x \in X : x \geqq 0\}$ is clearly convex.

To show that P is closed, let $\{x_n\}$ be a sequence of points in P which converges to $x \in H_0^1$. Then $x_n(t)$ converges to $x(t)$ in $L^2(\Omega)$, from which it follows that $x_n(t) \to x(t)$ a.e. Hence, $x(t) \geqq 0$ a.e. so that $x \in P$.

Let $x \in H_0^1(\Omega)$. Then $x \geqq 0$ *in the sense of* $H^1(\Omega)$ if there exists a sequence $\{\varphi_m\}$ of functions $\varphi_m \in C^1(\bar{\Omega})$ which satisfy $\varphi_m(t) \geqq 0$ in Ω and which converge to x in $H^1(\Omega)$ (Lewy and Stampacchia (1969, p. 155)). If $x \geqq 0$ in the sense of $H^1(\Omega)$ then it follows immediately that $x(t) \geqq 0$ a.e. Conversely, let $x \in H_0^1(\Omega)$ satisfy $x(t) \geqq 0$ a.e. If $\hat{x}$ denotes the extension of x to R^n obtained by setting $\hat{x}(t) = 0$ for $t \notin \Omega$, we know that $\hat{x} \in H_0^1(R^n)$ (Adams (1975, p. 57)). The averaged functions $\hat{x}_h$ are smooth and nonnegative, and they converge to $\hat{x}$ in $H^1(R^n)$ (Adams (1975, p. 52)). If $\varphi_h = \hat{x}_h|\Omega$ then $\varphi_h \to x$ in Ω, and we can conclude that $x \geqq 0$ in the sense of $H^1(\Omega)$. We have thus shown that if $x \in H_0^1(\Omega)$ then $x \geqq 0$ in the sense of $H^1(\Omega)$ iff $x(t) \geqq 0$ a.e. This is of importance to us because Stampacchia and his colleagues use $\geqq 0$ in the sense of $H^1(\Omega)$.

H_0^1 is a vector lattice: if $x, y \in H_0^1$ then the functions

$$\begin{aligned} \sup(x, y)(t) &= \sup(x(t), y(t)), \qquad t \in \Omega, \\ \inf(x, y)(t) &= \inf(x(t), y(t)), \qquad t \in \Omega, \end{aligned} \tag{2.2.3}$$

are representations of elements in H_0^1. (Lewy and Stampacchia (1969, p. 169) prove that $H^1(\Omega)$ is a vector lattice, and their proof can be readily adapted to the present case.)

Another very useful property of H_0^1 is that if $x \in H_0^1$ and F is a measurable subset of Ω on which $x(t)$ is constant then (Lewy and Stampacchia (1969, p. 169)),

$$\int_F |\text{grad}\, x(t)|^2\, dt = 0. \tag{2.2.4}$$

As defined in (2.2.1), the operator M can only be applied to functions u which are twice differentiable. Let $a: H_0^1 \times H_0^1 \to R^1$ be the symmetric coercive bilinear operator defined by

$$a(u, v) = \sum \int_\Omega a_{ij}(t)\frac{\partial u}{\partial t_i}\frac{\partial v}{\partial t_j}\, dt. \tag{2.2.5}$$

We extend the domain of definition of M by regarding M as the mapping from $X = H_0^1$ to its dual space $X^* = H^{-1}$ defined by

$$\langle v, Mu \rangle = a(u, v) \quad \text{for all } u, v \in H_0^1. \tag{2.2.6}$$

The standard theory of elliptic operators allows us to assert that M is uniquely defined by (2.2.6) and that M is a homeomorphism of $X=H_0^1$ onto $X^*=H^{-1}$ (Lions and Magenes (1972, p. 207)).

THEOREM 2.2.1. *M satisfies Conditions S and Z.*

Proof. To prove Condition Z, let $u, v \in P$ and $\inf(u, v)=0$. Let u vanish on $F \subset \Omega$ and v vanish on $G \subset \Omega$. Then, using (2.2.4), we conclude that

$$\begin{aligned} a(u, v) &= \sum \int_\Omega a_{ij}(t) \frac{\partial u}{\partial t_i} \frac{\partial v}{\partial t_j} dt, \\ &= \sum \int_F a_{ij}(t) \frac{\partial u}{\partial t_i} \frac{\partial v}{\partial t_j} dt + \int_{G-F} a_{ij}(t) \frac{\partial u}{\partial t_i} \frac{\partial v}{\partial t_j} dt = 0 \end{aligned}$$

so that Condition Z is satisfied.

Stampacchia (1965, p. 205) proves that Condition S is satisfied. □

3. The linear program, the dual linear program, and the least element problem. With the notation of § 2, the *linear program* (LP) is:

$$\text{(3.1)} \qquad \text{(LP)} \quad \underset{x \in P}{\text{Minimize}} \ \langle x, p\rangle \ \text{ subject to } \ Mx+q \geqq 0.$$

The dual program (LDF) which is (formally) dual to LP is:

$$\text{(3.2)} \qquad \text{(LDF)} \quad \underset{y^* \in P^{**}}{\text{Maximize}} \ \langle -q, y^*\rangle \ \text{ subject to } \ -M^*y^*+p \geqq 0,$$

where

$$\text{(3.3)} \qquad \begin{aligned} P^{**} &= \{y^* \in Y^*: y^* \geqq 0\}, \\ &= \{y^* \in Y^*: \langle u^*, y^*\rangle \geqq 0 \text{ for all } u^* \in P^*\}. \end{aligned}$$

If x is a solution of LP and y^* is a solution of LDF then,

$$\text{(3.4)} \qquad \langle x, p\rangle - \langle -q, y^*\rangle = \langle x, -M^*y^*+p\rangle + \langle Mx+q, y^*\rangle \geqq 0,$$

so that the value of LP is always greater than or equal to the value of LDF. In particular, if $\langle x, p\rangle + \langle q, y^*\rangle = 0$ for some feasible x and y^* then x and y^* are optimal. It may, however, occur that the two values are never equal, in which case there is a duality gap.

Since X is reflexive we know (Dunford and Schwartz (1966, p. 66)) that there is an isometric isomorphism κ which maps X onto $X^{**} = Y^*$ and which is defined by

$$\text{(3.5)} \qquad \langle x, x^*\rangle = \langle x^*, \kappa x\rangle.$$

Let

$$\text{(3.6)} \qquad y^* = \kappa y,$$

where $y \in X$ (not Y), so that

$$\text{(3.7)} \qquad \langle -q, y^*\rangle = \langle -y, q\rangle.$$

We assert that $y^* \geqq 0$ iff $y \in P$. First assume that $y \in P$. Then, for any $u^* \in P^*$ $\langle y, u^*\rangle \geqq 0$, so that $y^* \geqq 0$. On the other hand, suppose that $y^* \geqq 0$ but that $y \notin P$. Then, since the singleton $\{y\}$ is compact and the cone $P = \{x \in X: x \geqq 0\}$ is closed and convex, these two sets can be separated (Dunford and Schwartz (1966, p. 417)). That is, there exists a linear functional $f \in Y$ and constants $\varepsilon > 0$ and c such that

$$\langle x, f\rangle \geqq c \quad \text{if } x \in P,$$
$$\langle y, f\rangle \leqq c - \varepsilon.$$

Using the properties of P we conclude that $c = 0$, so that $f \in P^*$ and $\langle y, f\rangle \leqq -\varepsilon$. But $\langle y, f\rangle = \langle f, y^*\rangle \geqq 0$, and we have a contradiction.

Finally, for any $u \in X$,

$$(3.8) \qquad \begin{aligned} \langle u, M^* y^*\rangle &= \langle Mu, y^*\rangle, \quad \text{(definition of } M^*)\\ &= \langle Mu, \kappa y\rangle, \quad \text{(equation (3.6))},\\ &= \langle y, Mu\rangle, \quad \text{(definition of } \kappa)\\ &= \langle u, \tilde{M} y\rangle, \end{aligned}$$

where we define the linear operator $\tilde{M}: X \to X^* = Y$ by

$$(3.9) \qquad a(u, v) = \langle v, Mu\rangle = \langle u, \tilde{M} v\rangle.$$

Thus, $-M^* y^* + p \in P^*$ iff $-\tilde{M} y + p \in P^*$.

Summing up, we see that y^* satisfies LDF iff $y^* = \kappa y$ where y solves:

$$\text{(LD)} \qquad \underset{y \in X}{\text{Maximize}} \quad \langle -y, q\rangle \quad \text{subject to} \quad \begin{aligned} -\tilde{M} y + p &\geqq 0,\\ y &\geqq 0, \end{aligned}$$

and we will take this to be the dual of LP in our further work.

Since X is partially ordered, we may also consider the *least element problem* (LE): Find $x \in P$ such that $Mx + q \geqq 0$ and $x \leqq u$ for every $u \in P$ satisfying $Mu + q \geqq 0$. LE has at most one solution, for if x_1 and x_2 were two solutions we would have $x_1 \leqq x_2$ and $x_2 \leqq x_1$ which implies that $x_1 = x_2$.

In the special case $X = R^n$, there exists a very satisfactory theory for LP and LD, and Mangasarian (1976) used this as the starting point for his study of the relationship between LP and LC (the linear complementarity problem). LE has also been studied in the finite dimensional case (Cottle and Veinott (1972)).

The case when X is infinite dimensional is much more difficult. It is usually assumed, for example by Ekeland and Temam (1974, p. 66), that the Arrow–Hurwicz constraint qualification is satisfied, namely that there exists $u \in P$ such that $Mu + q$ is an interior point of P^*. An example of Craven (1977, p. 331) illustrates the difficulties which can arise when P^* does not have any interior points and when M is not an open map. Dempster (1975) develops a general framework for the analysis of LP and LD.

In the present paper we prove the existence of solutions to LP and LE by using the theory of variational inequalities. We do not prove the existence of a solution to LD, although in § 6 we give an example in which LD does have a solution.

4. The linear complementarity problem, the variational inequality, and the unilateral minimization problem. The *linear complementarity problem* (LC) is as follows: Find $x \in P$ such that

$$(4.1) \qquad \text{(LC)} \qquad Mx + q \geqq 0, \qquad \langle x, Mx + q\rangle = 0.$$

The *variational inequality* (VI) is: Find $x \in P$ such that

$$(4.2) \qquad \text{(VI)} \qquad a(x, v - x) + \langle v - x, q\rangle \geqq 0,$$

for all $v \in P$.

If a is symmetric then the *unilateral minimization problem* (or quadratic programming problem) (UM) is: Find $x \in P$ such that

$$\text{(4.3)} \qquad \text{(UM)} \qquad J(x) \leqq J(u) \quad \text{for all } u \in P$$

where

$$\text{(4.4)} \qquad J(u) = a(u, u) + 2\langle q, u\rangle.$$

The basic result on variational inequalities is due to Stampacchia (1964): if a is coercive then there exists a unique solution to VI.

The connection between VI and UM was also observed by Stampacchia (1964): if a is symmetric and coercive, then VI is equivalent to UM.

The relationship between VI and LC was noted independently by a number of workers including Lions and Stampacchia (1967, p. 172), Karamardian (1971), Moré (1971). The basic result is (Cottle (1976, Prop. 1, p. 181)):

THEOREM 4.1. *LC is equivalent to VI.*

5. The relationship between the linear program, the least element problem, and the linear complementarity problem.

THEOREM 5.1. *If a is coercive and satisfies Condition Z, then LE has a solution, namely the unique solution of VI.*

Proof. The proof is a modification of proof of Stampacchia (1965, p. 151) who implicitly used Condition Z in the special form: if $u, v \in P$ and $\inf(u, v) = 0$ then $a(u, v) = 0$.

Let u be the unique solution of VI so that $u \in P$ and

$$a(u, v-u) + \langle v-u, q\rangle \geqq 0$$

for all $v \in P$.

In particular, choosing $v = u + w$ for any $w \in P$ we conclude that $Mu + q \geqq 0$.

Now let w be any element such that $w \in P$ and $Mw + q \geqq 0$. We assert that $w \geqq u$. To see this, let $\zeta = \min(u, w) \in X$, so that $w - \zeta \geqq 0$ and $u - \zeta \geqq 0$. Furthermore, $\inf(w-\zeta, u-\zeta) = \inf(w, u) - \zeta = 0$.

Then

$$\begin{aligned} a(u-\zeta, u-\zeta) &= [a(\zeta, \zeta-u) + \langle \zeta-u, q\rangle] - [a(u, \zeta-u) + \langle \zeta-u, q\rangle], \\ &\leqq [a(\zeta, \zeta-u) + \langle \zeta-u, q\rangle], \end{aligned}$$

because u satisfies VI. But

$$\begin{aligned} a(\zeta, \zeta-u) + \langle \zeta-u, q\rangle &= a(w-\zeta, u-\zeta) + [a(w, \zeta-u) + \langle \zeta-u, q\rangle], \\ &\leqq 0, \end{aligned}$$

because the first term on the right is nonpositive by Condition Z and the second term is nonpositive since $Mw + q \geqq 0$ and $\zeta - u \leqq 0$.

Combining the above inequalities we see that $a(u-\zeta, u-\zeta) \leqq 0$. Remembering that a is coercive we conclude that $u = \zeta$. Thus, $w \geqq \zeta = u$ so that u is a solution of LE.

THEOREM 5.2.

(i) *If p is positive and x solves LE then x solves LP.*

(ii) *If a satisfies Condition S and p is strictly positive, then LP has at most one solution.*

(iii) *If a satisfies Condition S, p is strictly positive, and x solves LP then x solves LE.*

Proof. (i) is obvious. To prove (ii), let x_1 and x_2 be two solutions of LP. By Condition S, $\zeta = \inf(x_1, x_2) \in P$ satisfies $M\zeta + q \geqq 0$ and $\langle \zeta, p\rangle \leqq \langle x_1, p\rangle$. Since x is optimal, $\langle \zeta, p\rangle = \langle x_1, p\rangle$ and we conclude that $\zeta = x_1$. Similarly, $\zeta = x_2$, so that $x_1 = x_2$.

To prove (iii), let $u \in P$ satisfy $Mu + q \geqq 0$. Set $\zeta = \inf(u, x)$. Then $M\zeta + q \geqq 0$ and $\langle \zeta, p\rangle = \langle x, p\rangle$ so that $\zeta = x$. Hence, $u \geqq x$ and x solves LE. □

Remembering that if a is coercive and a satisfies Condition Z then a satisfies Condition S (Theorem 2.1) we find the following:

THEOREM 5.3. *If a is coercive and satisfies Condition Z, and if p is strictly positive, then LP, LE, VI, and LC all have the same unique solution.*

THEOREM 5.4. *Assume that x solves VI, that y solves LD, that $\langle x, p\rangle + \langle y, q\rangle = 0$, that a is coercive and satisfies Condition Z, and that $p + q \geqq 0$.*

Then $y \geqq x$.

Proof. Set $w = \inf(x, y)$. Then

$$\begin{aligned} a(x-w, x-w) &= a(x-y, x-w) + a(y-w, x-w), \\ &\leqq a(x-y, x-w), \end{aligned}$$

since $y - w \geqq 0$, $x - w \geqq 0$, and $\inf(y-w, x-w) = 0$. But,

$$\begin{aligned} a(x-y, x-w) &= a(x, x-w) - a(y, x-w) \\ &= a(x, x-w) - a(x-w, y) \\ &= a(x, x-w) + \langle x-w, -\tilde{M}y\rangle \\ &= [a(x, x-w) + \langle x-w, q\rangle] - \langle x-w, p+q\rangle + \langle x-w, -\tilde{M}y + p\rangle. \end{aligned}$$

The first term on the right is negative because x solves VI. The second term is negative because $p + q \in P^*$ and $x - w \in P$. The third term is zero because the equality

$$0 = \langle x, p\rangle + \langle y, q\rangle = \langle x, -\tilde{M}y + p\rangle + \langle y, Mx + q\rangle$$

implies that $\langle x, -\tilde{M}y + p\rangle = 0$ and hence, since $0 \leqq w \leqq x$, that $\langle w, -\tilde{M}y + p\rangle = 0$.

Combining the above, we conclude that $a(x-w, x-w) \leqq 0$ so that $x = w$. Then $y \geqq w = \inf(x, y) = x$. □

It may be observed that if x solves LP, y solves LD, $\langle x, p\rangle + \langle y, q\rangle = 0$, and $y \geqq x$, then we have that

$$0 \leqq \langle x, Mx + q\rangle \leqq \langle y, Mx + q\rangle = 0;$$

that is, x solves LC.

6. A one-dimensional problem. We consider a special case of Example 2 (§ 2.2): $X = H_0^1(0, 2)$,

$$\begin{aligned} &\min \langle x, p\rangle = \int_0^2 1x(t)\,dt \quad \text{subject to} \quad x(t) \geqq 0 \text{ a.e., and} \\ &Mx + q \equiv -\ddot{x}(t) + (t-1) \geqq 0, \end{aligned} \tag{6.1}$$

with the corresponding dual problem

$$\begin{aligned} &\max \langle y, -q\rangle = -\int_0^2 (t-1)y(t)\,dt \quad \text{subject to} \quad y(t) \geqq 0 \text{ a.e.,} \\ &-\tilde{M}y + p \equiv \ddot{y}(t) + 1 \geqq 0. \end{aligned} \tag{6.2}$$

The inequality $-\ddot{x}+(t-1)\geqq 0$ is interpreted in the sense that

$$\langle\varphi, Mx+q\rangle=\int_0^2 [\dot{x}(t)\dot{\varphi}(t)+(t-1)\varphi(t)]\,dt\geqq 0, \tag{6.3}$$

for all nonnegative $\varphi\in H_0^1(0,2)$, and the inequality $\ddot{y}+1\geqq 0$ is interpreted in the same way.

This problem was chosen because it is a simple problem with the same general structure as the problem for a cavitating journal bearing which is discussed in the next section.

There is a straightforward procedure for obtaining possible solutions of such one-dimensional problems; these solutions can then be varied *a posteriori*. We assume that $x(t)>0$ for $0<t<\tau$ and $x(t)=0$ for $\tau\leqq t\leqq 2$, where τ is an unknown constant corresponding to the free boundary (the point $t=\tau$). If x also satisfies LC then $\langle x, -\ddot{x}+(t-1)\rangle=0$, so that $-\ddot{x}(t)+(t-1)=0$ for $0\leqq t\leqq\tau$. The general solution of the equation $-\ddot{x}+(t-1)=0$ is

$$x(t)=A+Bt+\tfrac{1}{6}(t-1)^3. \tag{6.4}$$

Using the conditions $x(0)=x(\tau)=0$ to determine the constants A and B we find

$$x(t)=t(t-\tau)[-3+t+\tau]/6. \tag{6.5}$$

To determine τ we note that the condition $-\ddot{x}+(t-1)\geqq 0$ implies that for all smooth nonnegative $\varphi\in H_0^1(0,2)$,

$$\begin{aligned}\langle\varphi, Mx+q\rangle&=\int_0^2[\dot{x}\dot{\varphi}+(t-1)\varphi]\,dt\\&=\int_0^\tau[\dot{x}\dot{\varphi}+(t-1)\varphi]\,dt+\int_\tau^2(t-1)\varphi\,dt\\&=\dot{x}\varphi]_0^\tau+\int_0^\tau[-\ddot{x}\varphi+(t-1)\varphi]\,dt+\int_\tau^2(t-1)\varphi\,dt\\&=\dot{x}(\tau-)\varphi(\tau)+\int_\tau^2(t-1)\varphi\,dt\\&\geqq 0.\end{aligned} \tag{6.6}$$

This is only possible if $\tau\geqq 1$ (so that $t-1\geqq 0$ for $t\in[\tau,2]$) and $\dot{x}(\tau-)\geqq 0$. But, $x(t)\geqq 0$ for $t\leqq\tau$ and $x(\tau)=0$ so $\dot{x}(\tau-)\leqq 0$. We conclude that $\dot{x}(\tau-)=\dot{x}(\tau+)=\dot{x}(\tau)=0$. The condition $\dot{x}(\tau)=0$ leads to an algebraic equation for τ, namely,

$$\dot{x}(\tau)=\tau[-3+2\tau]/6=0;$$

thus, $\tau=\frac{3}{2}$ and

$$x(t)=\begin{cases}t(t-\frac{3}{2})^2/6, & 0\leqq t\leqq\frac{3}{2},\\ 0, & \frac{3}{2}\leqq t\leqq 2,\end{cases} \tag{6.7}$$

is our trial solution.

Using (6.6) and (6.7) we see that x is such that $x\geqq 0$, $-\ddot{x}+(t-1)\geqq 0$, and $\langle x, Mx+q\rangle=0$, so that x is a solution of LC. Invoking Theorems 4.1 and 5.3, we conclude that x is the unique solution of LP.

We now assume that $0=\langle x, p\rangle+\langle y, q\rangle$. Since $\langle x, \ddot{y}+1\rangle=0$, it follows that $\ddot{y}(t)+1=0$ when $x(t)>0$, that is, when $0<t<\tau$. On the other hand, since $\langle y, -\ddot{x}+(t-1)\rangle=0$, it follows that $y(t)=0$ when $-\ddot{x}+(t-1)>0$, that is, when $\tau<t<2$. We conclude that $\ddot{y}(t)+1=0$ for $0\leqq t\leqq \frac{3}{2}$ and $y(t)=0$ for $\frac{3}{2}\leqq t\leqq 2$. Solving this boundary value problem we obtain

$$(6.8)\qquad y(t)=\begin{cases} t[-2t+3]/4, & 0\leqq t\leqq \frac{3}{2},\\ 0, & \frac{3}{2}\leqq t\leqq 2.\end{cases}$$

The condition $y\geqq 0$ is seen to be satisfied.

Direct computation yields

$$(6.9)\qquad \langle x, p\rangle=\int_0^2 x(t)\,dt=\tfrac{9}{128}=-\int_0^2 (t-1)y(t)\,dt=-\langle y, q\rangle.$$

The solutions $x(t)$ and $y(t)$ are plotted in Figure 6.1. We note that $y\geqq x$ as proved in Theorem 5.4.

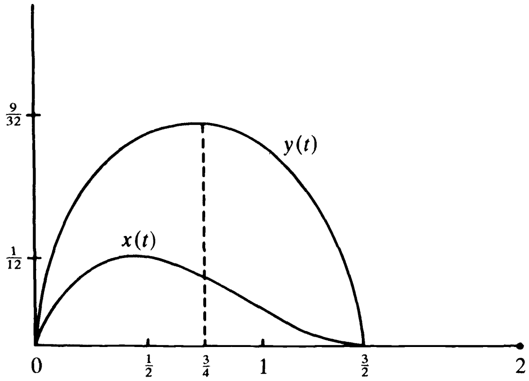


FIG. 6.1. *x(t) and y(t).*

It is possible to give two justifications for the free boundary condition $\dot{x}(\tau)=0$. Firstly, if $x\in H_0^2(0,1)$, as is often the case, then $\dot{x}(t)$ is continuous so that $\dot{x}(\tau)=\dot{x}(\tau+)=0$. Secondly, a reasonable interpretation of the condition $-\ddot{x}(\tau)+(\tau-1)\geqq 0$ is that

$$\lim_{\Delta t\to 0}-\frac{\dot{x}(\tau+\Delta t)-\dot{x}(\tau-\Delta t)}{2\Delta t}+(\tau-1)\geqq 0.$$

Since $\dot{x}(\tau+\Delta t)=0$ and $\dot{x}(\tau-0)\leqq 0$, it follows that $\dot{x}(\tau-0)=0$.

7. Lubrication cavitation of journal bearings. A large number of physical problems can be formulated as linear complementarity problems in which a differential equation (ordinary or partial) must be solved subject to the inequality constraint that the solution be nonnegative; roughly speaking, at any point the solution must either be zero or satisfy the differential equation (Cryer (1977), (1979), Duvaut and Lions (1972)). The reformulation of such linear complementarity problems as linear programs has two advantages: (i) it suggests alternative methods of solving the problems; and (ii) it sometimes provides a physically meaningful interpretation. As an example of such linear complementarity problems we consider here the problem of a cavitating journal bearing.

A journal bearing consists of a circular cylinder (the journal) which is rotating inside a support structure (the bearing). The narrow gap between the journal and the bearing is filled with a thin film of lubricating fluid. Various geometries are possible. In

Fig. 7.1 we show a partial journal bearing of finite length. The term 'partial' refers to the fact that the journal is not completely enclosed within the bearing, and is partially exposed to the atmosphere.

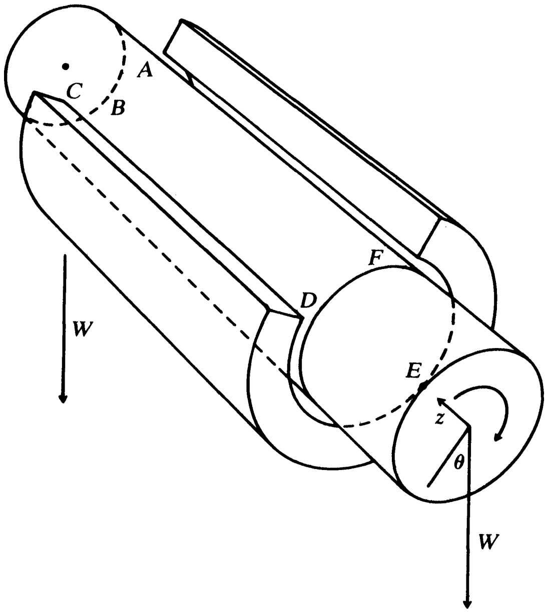


FIG. 7.1. *A partial journal bearing.*

It is required to determine the pressure x of the lubricant, and the load W borne by the bearing. Because the gap between the journal and the bearing is very narrow, the simplifications of lubrication theory can be applied. In particular, it is assumed that the pressure does not vary across the gap, so that the problem becomes a two-dimensional problem in the rectangular domain $\Omega = ABCDEF$ in the θz-plane (Fig. 7.2).

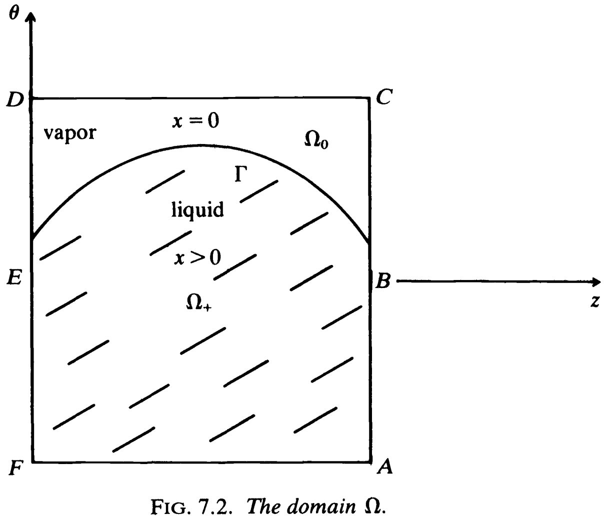


FIG. 7.2. *The domain* Ω.

The lubricant flows in from a reservoir along the entry edge AF and flows out through the ends ABC and DEF as well as through the exit edge CD. At all these points the lubricant is in contact with the atmosphere, and if the pressure is normalized so that atmospheric pressure is zero, then the boundary conditions are that $x=0$ on $\partial\Omega$. That is,

$$x \in X = H_0^1(\Omega). \tag{7.1}$$

The lubricant occurs in both liquid and gaseous phases. It is assumed that the lubricant vaporizes when the pressure is zero, so that the inequality $x \geqq 0$ must be satisfied everywhere. If the pressure is greater than zero then the lubricant is in the liquid phase and satisfies the simplified form of the Navier–Stokes equations known as Reynolds' equation. After introducing dimensionless variables, the equation takes the form (Pinkus and Sternlicht (1961)):

$$Mx + q \equiv -\frac{\partial}{\partial\theta}\left(h^3 \frac{\partial x}{\partial\theta}\right) - \alpha^2 \frac{\partial}{\partial z}\left(h^3 \frac{\partial x}{\partial z}\right) + \frac{dh}{d\theta} = 0, \tag{7.2}$$

where α is a positive constant, and where $h = h(\theta)$ is a given function which is proportional to the width of the gap.

On the free boundary Γ, the interface between the liquid and gaseous phases, the boundary conditions are

$$x = 0, \qquad \partial x/\partial n = 0, \quad \text{on } \Gamma, \tag{7.3}$$

where $\partial/\partial n$ denotes the normal derivative.

In the engineering literature (Pinkus and Sternlicht (1961)) the problem is formulated mathematically as a classical free boundary problem: Find x and Γ such that x satisfies (7.2) subject to the boundary conditions (7.1), and (7.3). However, in a large number of papers in the engineering literature, beginning with the work of Christopherson (1941), numerical approximations have been obtained in a completely different way: equation (7.2) is replaced by finite differences, and the resulting system of algebraic equations is solved as a finite-dimensional linear complementarity problem (Cryer (1971)) which may be considered as a discretization of the infinite-dimensional linear complementarity problem

$$x \geqq 0, \qquad Mx + q \geqq 0, \qquad \langle x, Mx + q\rangle = 0. \tag{7.4}$$

We may thus take (7.4) as the starting point for a mathematical analysis of the problem. The problem is a special case of Example 2 (§ 2.2), and it follows from Theorem 5.1 that there exists a unique solution $x \in H_0^1(\Omega)$ of LE, VI, and LC.

In the engineering literature, there has been some discussion of an appropriate variational principle for the problem (Christopherson (1957)). The formulation as a variational inequality leads to two useful variational principles:

(1) Since a is symmetric, the problem is equivalent to the unilateral minimization problem

$$\inf_{v \geqq 0} J(v) = a(v, v) + 2\langle v, q\rangle.$$

(2) For any strictly positive function $p(\theta, z)$, the problem is equivalent to the linear programming problem

$$\min \langle x, p\rangle \equiv \int_\Omega x(\theta, z) p(\theta, z)\, d\theta\, dz,$$

subject to $x \geqq 0$, $Mx+q \geqq 0$. In particular, if $-\pi/2 < \theta_F < \theta_D < \pi/2$ (see Figs. 7.1 and 7.2), then $p = \cos \theta > 0$ and $\langle x, p \rangle$ is the load W borne by the bearing in the vertical direction (Fig. 7.1). That is, the solution x minimizes the vertical load.

After completing this paper, we became aware of the work of McAllister and Rohde (1976) and Cimatti (1977) where the journal bearing problem is also considered using variational inequalities.

Acknowledgment. The authors wish to thank Professor O. L. Mangasarian with whom they had many helpful discussions. The authors also appreciate the helpful comments of the referee.

Note added in proof. Further references on variational inequalities include: H. BREZIS (1972), *Problèmes unilateraux*, J. Math. Pures Appl., 51, pp. 1–168. U. MOSCO (1969), *Convergence of convex sets and solutions of variational inequalities*, Advances in Math. 3, pp. 510–585.

REFERENCES

R. A. ADAMS (1975), *Sobolev Spaces*, Academic Press, New York.

J. C. BIERLEIN (1975), *The journal bearing*, Scientific American, July, pp. 50–64.

M. BRELOT (1945), *Minorantes sous-harmoniques, extrémales et capacités*, J. Math. Pures Appl., 24, pp. 1–32.

——— (1965), *Eléments de la Théorie Classique du Potential*, 3rd edition, Les Cours de Sorbonne, Centre de Documentation Universitaire, Paris.

D. G. CHRISTOPHERSON (1941), *A new mathematical method for the solution of film lubrication problems*, Proc. Inst. Mech. Engrs., 6, pp. 126–135.

——— (1957), *Boundary conditions in lubricating films*, The Engineer, 203, p. 100.

G. CIMATTI (1977), *On a problem of the theory of lubrication governed by a variational inequality*, Applied Math. Optimization, 3, pp. 227–242.

R. W. COTTLE (1976), *Complementarity and variational problems*, Symposia Mathematica, 19, pp. 177–208.

R. W. COTTLE, G. H. GOLUB AND R. S. SACHER (1978), *On the solution of large, structured linear complementarity problems: the block partitioned case*, Appl. Math. Optimization, 4, pp. 347–363.

R. W. COTTLE AND J. S. PANG (1976), *A least element theory of solving linear complementarity problems as linear programs*, Technical Summary Report #1702, Mathematics Research Center, University of Wisconsin, Madison.

——— (1978), *On solving linear complementary problems as linear programs*, Math. Programming Study 7, pp. 88–107.

R. W. COTTLE AND A. F. VEINOTT, JR. (1972), *Polyhedral sets having a least element*, Math. Programming, 3, pp. 238–249.

B. D. CRAVEN (1977), *Lagrangean conditions and quasiduality*, Bull. Australian Math. Soc., 16, pp. 325–339.

C. W. CRYER (1971), *The method of Christopherson for solving free boundary problems for infinite journal bearings by means of finite differences*, Math. Comput., 25, pp. 435–443.

——— (1977), *A bibliography of free boundary problems*, Technical Summary Report #1793, Mathematics Research Center, University of Wisconsin, Madison.

——— (1979), *A survey of variational inequalities*, Technical Summary Report, Mathematics Research Center, University of Wisconsin, Madison, in preparation.

M. A. H. DEMPSTER (1975), *Abstract optimization and its applications*, Lecture Notes, Dept. of Mathematics, University of Melbourne.

N. DUNFORD AND J. SCHWARTZ (1966), *Linear Operators*, vol. I, Wiley–Interscience, New York.

J. F. DURAND (1968), *Résolution numérique de problèmes aux limites sous-harmoniques*, thesis, Université de Montpellier.

G. DUVAUT AND J. L. LIONS (1972), Les Inéquations en Mécanique et en Physique, Dunod, Paris.

I. EKELAND AND R. TEMAM (1974), *Analyse Convexe et Problèmes Variationnels*, Dunod, Paris.

M. FIEDLER AND V. PTAK (1962), *On matrices with non-positive off-diagonal elements and positive principal minors*, Czech. Math. J., 12, pp. 382–400.

S. KARAMARDIAN (1971), *Generalized complementary problem*, J. Optimization Theory Appl., 8, pp. 161–168.

J. L. KELLEY AND I. NAMIOKA (1976), *Linear Topological Spaces*, second corrected printing, Springer-Verlag, New York.

H. LEWY AND G. STAMPACCHIA (1969), *On the regularity of the solution of a variational inequality*, Comm. Pure Appl. Math., 22, pp. 153–188.

J. L. LIONS AND E. MAGENES (1972), *Non-homogeneous Boundary Value Problems and Applications, I.* Springer-Verlag, Berlin.

J. L. LIONS AND G. STAMPACCHIA (1967), *Variational inequalities*, Comm. Pure Appl. Math., 20, pp. 493–519.

W. LITTMAN (1963), *Generalized subharmonic functions : monotonic approximations and an improved maximum principle*, Ann. Scuola Norm. Superiore Pisa Sci. Fis. Mat., (3) 17, pp. 207–222.

O. L. MANGASARIAN (1976), *Linear complementarity problems solvable by a single linear program*, Math. Programming, 10, pp. 263–270.

——— (1977), *Characterization of linear complementarity problems as linear programs*, Mathematical Programming Study 9, pp. 74–87.

——— (1979), *Simplified characterizations of linear complementarity problems solvable as linear programs*, Math. of Operations Res. 4, No. 3.

G. T. MCALLISTER AND S. M. ROHDE (1976), *An optimization problem in hydrodynamic lubrication theory*, Appl. Math. and Optimization, 2, pp. 223–235.

J. J. MORÉ (1971), *The application of variational inequalities to complementarity problems and existence theorems*, Tech. Rep. no. 71-110, Dept. of Computer Science, Cornell University, Ithaca, N.Y.

J. J. MOREAU (1971), *Majorantes sur-harmoniques minimales d'une fonction continue*, Ann. Inst. Fourier Grenoble, 21, pp. 129–156.

J. S. PANG (1976), *Least-element complementarity theory*, Ph.D. dissertation, Dept. of Operations Research, Stanford University, CA, Sept.

——— (1977), *A note on an open problem in linear complementarity*, Math. Programming 13, pp. 360–363.

——— (1978), *On cone orderings and the linear complementarity problem*, Linear Algebra and Appl., 22, pp. 267–281.

O. PINKUS AND B. STERNLICHT (1961), *Theory of Hydrodynamic Lubrication*, McGraw-Hill, New York.

R. J. PLEMMONS (1976), *A survey of M-matrix characterizations I: nonsingular M-matrices*, Technical Summary Report #1651, Mathematics Research Center, University of Wisconsin, Madison.

G. POOLE AND T. BOULLION (1974), *A survey of M-matrices*, SIAM Rev. 16, pp. 419–427.

T. RADO (1972), *Subharmonic Functions and the Problem of Plateau*, Chelsea, New York.

G. STAMPACCHIA (1964), *Formes bilinéaires coercitives sur les ensembles convexes.* C. R. Acad. Sci. Paris, 258, pp. 4413–4416.

——— (1965), *Le problème de Dirichlet pour les équations élliptiques du second ordre à coefficients discontinus*, Ann. Inst. Fourier Grenoble, 15, pp. 189–258.

A. TAMIR (1973), *The complementarity problem of mathematical programming*, Ph.D. thesis, Dept. of Operations Research, Case Western Reserve University.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 1, January 1980

0363-0129/80/1801-0006$01.00/0

NECESSARY CONDITIONS FOR OPTIMALITY OF ELLIPTIC SYSTEMS WITH POSITIVITY CONSTRAINTS ON THE STATE*

P. MICHEL†

Abstract. Recently, necessary conditions for optimality were established in the case of nonlinear elliptic multi-dimensional systems with integral constraints [P. Michel, *Condition nécessaire d'optimalité pour des systèmes d'équations elliptiques non linéaires*, Quatrièmes journées de contrôle (Metz, 18–21 Mai 1976), polycopié Université de Metz, 1976]. Here this problem is solved in the case of additional constraints of positivity of the state's components on measurable sets.

Introduction. The study of necessary conditions for optimality of partial differential systems was a long time limited to the case without constraints on the state [3], [7], [1]. Recently, these limitations were partially removed: in [2] there is a statement in the case of integral constraints (and it seems correct only for inequality constraints); in [5] and [6], there are necessary conditions, respectively for elliptic and parabolic systems, in the case of nonlinear equations, finite-dimensional state and integral constraints.

The present study sets the necessary conditions in the case of additional positivity constraints, like $y(x) \geqq 0$ on a given measurable set; such constraints are useful in the applications.

1. Problem statement. Let us consider the following optimal control problem.

Problem (P): minimize the cost functional

$$\int_\Omega g_0(y(x), x)\, dx$$

for $y \in V$ and u an admissible control such that:

(1) for $1 \leqq l \leqq m_0 : A_l(u) \cdot y(x) + f_l(y(x), u(x), x) = 0$ a.e. in Ω,
for $1 \leqq k \leqq m : y_k(x) = 0$ a.e. in $\partial\Omega$;

(2) $u(x) \in U(x)$ a.e. in Ω,

(3) for $1 \leqq i \leqq i_0 : \int_\Omega g_i(y(x), x)\, dx \leqq 0$,
for $i_0 + 1 \leqq i \leqq i_1 : \int_\Omega g_i(y(x), x)\, dx = 0$;

(4) for $1 \leqq k \leqq m : y_k(x) \geqq 0$ a.e. in Ω_k,

where Ω is an open bounded subset of $\mathbb{R}^n$; $\partial\Omega$ denotes the boundary of Ω; U is a topological space and for each $x \in \Omega$, $U(x)$ is a subset of U; for $1 \leqq k \leqq m, f_l(1 \leqq l \leqq m_0)$ and $g_i(0 \leqq i \leqq i_1)$ are real-valued functions respectively defined on $\mathbb{R}^m \times U \times \Omega$ and on $\mathbb{R}^m \times \Omega$, which are of class C^1 with respect to the first variable, i.e., such that for each $v \in U$ and almost every $x \in \Omega$, the functions

$$z \to f_l(z, v, x) \quad \text{and} \quad z \to g_i(z, x)$$

are of class C^1 in $\mathbb{R}^m$.

For $1 \leqq k \leqq m$, Ω_k is a measurable subset of Ω.

For $1 \leqq l \leqq m_0$, the elliptic operator $A_l(u)$ is defined by:

$$(5) \qquad A_l(u) \cdot y(x) = -\sum_{k=1}^{m} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial}{\partial x_i}\left(a_{ijkl}(x, u(x)) \frac{\partial y_k}{\partial x_j}(x) \right) + \sum_{k=1}^{m} b_{kl}(x, u(x)) y_k(x)$$

* Received by the editors November 17, 1978, and in revised form March 5, 1979.

† Université de Paris I, C.M.E. (Panthéon 219), 75231 Paris CEDEX 05, France.

with realvalued functions a_{ijkl} and b_{kl} defined on $\Omega \times U$. The state $y=(y_1, \cdots, y_m)$ belongs to the space $V=H_0^1(\Omega)^m$, $H_0^1(\Omega)$ denoting the Sobolev space

$$\left\{ z \in L^2(\Omega) \,\middle|\, \frac{\partial z}{\partial x_i} \in L^2(\Omega) \ \ (1 \leqq i \leqq n) \text{ and } z=0 \text{ on } \partial\Omega \right\}.$$

$y \in V$ is a *solution* of (1) iff it satisfies: for $1 \leqq l \leqq m_0$ and for $\varphi \in H_0^1(\Omega)$

$$\sum_{k,i,j} \int_\Omega a_{ijkl}(x, u(x)) \frac{\partial y_k}{\partial x_j} \frac{\partial \varphi}{\partial x_i} \, dx + \sum_k \int_\Omega b_{kl}(x, u(x)) y_k \varphi \, dx + \int_\Omega f_l(y(x), u(x), x) \varphi(x) \, dx = 0. \tag{6}$$

Let $\bar{y} \in V$ be fixed, and q, r and s be positive integers such that

$$q > 2; \quad \frac{1}{q} \geqq \frac{1}{2} - \frac{1}{n}; \quad r = \frac{q}{q-1}; \quad s = \frac{q}{q-2}. \tag{7}$$

An *admissible control* u is a measurable function from Ω into U which verifies (2). A *regular control* is an admissible control which verifies: there exist $C_u(x) \in L^s(\Omega)$ and $d_u \geqq 0$ such that

$$\begin{aligned} &u(x) \in U(x) \quad \text{a.e. in } \Omega, \\ &a_{ijkl}(x, u(x)) \in L^\infty(\Omega) \quad \forall i, j, k, l, \\ &b_{kl}(x, u(x)) \in L^s(\Omega) \quad \forall k, l, \\ &f_l(\bar{y}(x), u(x), x) \in L^r(\Omega) \quad \forall l, \\ &\left| \frac{\partial f_l}{\partial z_k}(z, u(x), x) \right| \leqq C_u(x) + d_u |z|^{q-2}, \quad \forall k, l, \forall z \in \mathbb{R}^m. \end{aligned} \tag{8}$$

S denotes the set of regular controls.

2. Assumptions. Let $\bar{u}$ be a regular control and $\bar{y}$ be a function in V which satisfy conditions (1), (2), (3), (4).

Assumption 1. There exist $C(x) \in L^r(\Omega)$ and $d \geqq 0$ such that, for $0 \leqq i \leqq i_1$, $z \in \mathbb{R}^m$, and $1 \leqq k \leqq m$

$$\left| \frac{\partial g_i}{\partial z_k}(z, x) \right| \leqq C(x) + d|z|^{q-1}. \tag{9}$$

Assumption 2. For every measurable function $(z(x), v(x))$ from Ω into $\mathbb{R}^m \times U$, all the functions:

$$\begin{aligned} &f_l(z(x), v(x), x) \quad \text{and} \quad \frac{\partial f_l}{\partial z_k}(z(x), v(x), x), \\ &g_i(z(x), x) \quad \text{and} \quad \frac{\partial g_i}{\partial z_k}(z(x), v(x), x), \\ &a_{ijkl}(x, v(x)) \quad \text{and} \quad b_{kl}(x, v(x)) \end{aligned}$$

are measurable.

Notation 1. For $x \in \Omega$, $v \in U(x)$ and $Y=(y_0, y_1, \cdots, y_n) \in (\mathbb{R}^m)^{n+1}$, let $h(x, v, Y)$

be the point $Z=(z_0,\cdots,z_n)\in(\mathbb{R}^{m_0})^{n+1}$ whose components are, for $1\leqq l\leqq m_0$

$$z_{0l}=f_l(y_0,v,x)+\sum_{k=1}^{m} b_{kl}(x,v)y_{0k},$$

$$z_{il}=\sum_{k=1}^{m}\sum_{j=1}^{n} a_{ijkl}(x,v)y_{jl}\quad \forall 1\leqq i\leqq n.$$

Assumption 3. For almost all $x\in\Omega$ and for every $Y\in(\mathbb{R}^m)^{n+1}$, the set $h(x,U(x),Y)$ is a closed convex set; for any measurable functions $Y(x)$ and $Z(x)$ verifying $Z(x)\in h(x,U(x),Y(x))$ a.e., there exists a measurable function $v(x)\in U(x)$ such that $Z(x)=h(x,v(x),Y(x))$ a.e.

Notation 2. V_0 denotes the space $H_0^1(\Omega)^{m_0}$.

Assumption 4. For each y' belonging to the dual space V_0' of V_0, there exists at least one solution $y\in V$ of the linear system:

$$\text{for } 1\leqq l\leqq m_0: A_l(\bar{u})\cdot y(x)+\sum_{k=1}^{m}\frac{\partial f_l}{\partial z_k}(\bar{y}(x),\bar{u}(x),x)y_k(x)-y_l'=0.$$

Notation 3. For $x\in\Omega$, $v\in U$ and $p=(p_1,\cdots,p_{m_0})\in V_0$, one sets

$$H(p,v,x)=\sum_l f_l(\bar{y}(x),v,x)p_l(x)+\sum_{k,l} b_{kl}(x,v)\bar{y}_k(x)p_l(x)$$
$$+\sum_{i,j,k,l} a_{ijkl}(x,v)\frac{\partial\bar{y}_k}{\partial x_j}(x)\frac{\partial p_l}{\partial x_i}(x). \tag{10}$$

Assumption 5. For each $p\in H_0^1(\Omega)^{m_0}$,

$$\inf_{v\in U(x)} H(p,v,x)=\inf_{u\in S} H(p,u(x),x)\text{ a.e. in }\Omega;$$

the infimum on S is taken according to the order on $L^1(\Omega)$.

Remarks. The usual assumptions (see for example [2], [7]) involve these assumptions. The convexity assumption 3 is the only nonstandard assumption: generally $U(x)$ is a fixed finite-dimensional convex set, all the functions are assumed to be differentiable with respect to the control, and the conclusion is the maximum of $(\partial H/\partial v)(p,\bar{u}(x),x)\cdot v$; to obtain the maximum of the Hamiltonian H, one needs then additional convexity assumptions which are quite stronger than Assumption 3. The existence of a solution of the linearized system of (1) (Assumption 4) is satisfied in the usual case $m_0=m$ with the coercevity conditions.

3. Necessary conditions of optimality.

Notation 4. The adjoint operators of $A(u)$ and $(\partial f/\partial z)(\bar{y}(x),u(x),x)$ are respectively defined by their components, for $1\leqq k\leqq m$:

$$(A(u)^*\cdot p)_k=\sum_{i,j,l}\frac{\partial}{\partial x_j}\left(a_{ijkl}(x,u(x))\frac{\partial p_l}{\partial x_i}\right)+\sum_l b_{kl}(x,u(x))p_l(x), \tag{11}$$

$$\left(\frac{\partial f}{\partial z}(\bar{y}(x),u(x),x)^*\cdot p\right)_k=\sum_l\frac{\partial f_l}{\partial z_k}(\bar{y}(x),u(x),x)p_l(x). \tag{12}$$

THEOREM. *Let $(\bar{y},\bar{u})$ be an optimal solution of problem (P) with $\bar{u}$ a regular control, and Assumptions* 1 *to* 5 *be satisfied. Then there exist real numbers α_i $(0\leqq i\leqq i_1)$ and*

$p \in V_0$ such that:

(1) *p and α_i $(0 \leqq i \leqq i_1)$ are not all zero;*

(2) *for $0 \leqq i \leqq i_0$, α_i is nonnegative;*

(3) *for $1 \leqq i \leqq i_0$, $\alpha_i \int_\Omega g_i(\bar{y}(x), x)\, dx = 0$;*

(4) *the function $\gamma \in V'$ whose components are*

$$(13) \qquad \gamma_k = (A(\bar{u})^* \cdot p)_k + \left(\frac{\partial f}{\partial z}(\bar{y}(x), \bar{u}(x), x)^* \cdot p\right)_k + \sum_{i=1}^{i_1} \alpha_i \frac{\partial g_i}{\partial z_k}(\bar{y}(x), x)$$

satisfies the properties for $1 \leqq k \leqq m$

$$(14) \qquad \gamma_k \cdot \bar{y}_k = 0,$$

$$(15) \qquad y \in H_0^1(\Omega) \text{ and } y(x) \geqq 0 \text{ on } \Omega_k \Rightarrow \gamma_k \cdot y \geqq 0.$$

(5) *the Hamiltonian $H(p, v, x)$ defined by* (10) *attains, for almost every x in Ω, its minimum on the set $U(x)$ at $\bar{u}(x)$.*

Remark. For $\Omega_k = \varnothing$ $(1 \leqq k \leqq m)$, one obtains the result of [5]: $\gamma_k = 0$ and p is a solution of the adjoint of the linearized system of (1). To illustrate the fourth conclusion, let us consider the regular case where the functions γ_k may be identified with elements of $L^2(\Omega)$ (that is if $p \in H_0^2(\Omega)^{m_0}$); then one gets

$$\gamma_k \cdot y = \int_\Omega \gamma_k(x) y(x)\, dx,$$

and with the denseness of $H_0^1(\Omega)$ in $L^2(\Omega)$, one obtains:

$$\begin{aligned} \gamma_k(x) &\geqq 0 \quad \text{a.e. in } \Omega, \\ \gamma_k(x) &= 0 \quad \text{a.e. in } \Omega - \Omega_k, \\ \gamma_k(x) &= 0 \quad \text{a.e. in } \{x \in \Omega_k;\ \bar{y}_k(x) > 0\}. \end{aligned}$$

The first part of the proof of the theorem. This is identical with the proof in the case without permanent constraints [5]. One sets, for a family $\lambda = (\lambda_u)_{u \in S}$ of scalar measurable functions $\lambda_u(x)$ in Ω,

$$\begin{aligned} \|\lambda\| = \sum_u \sum_{i,j,k,l} (&\|\lambda_u(x) a_{ijkl}(x, u(x))\|_{L^\infty} + \|\lambda_u(x) b_{kl}(x, u(x))\|_{L^s} \\ &+ \|\lambda_u(x) f(\bar{y}(x), u(x), x)\|_{L^r} + \|\lambda_u(x) C_u(x)\|_{L^s} + d_u \|\lambda_u(x)\|_{L^\infty}). \end{aligned}$$

In the space E of the families λ which satisfy: $\|\lambda\| < \infty$, $\|\cdot\|$ is a norm, and with this norm, E is a Banach space. M denotes the closed convex subset of the λ which verify: $\forall u \in S$, $\lambda_u(x) \geqq 0$ a.e., and $\sum_u \lambda_u(x) = 1$ a.e.; S becomes a subset of M by the correspondence which associates to $u_0 \in S$ the family ($\lambda_{u_0}(x) = 1$ a.e., and $\forall u \neq u_0$, $\lambda_u(x) = 0$ a.e.); $\bar{\lambda}$ denotes the family associated to $\bar{u}$.

The same arguments as in [6] show that $(\bar{y}, \bar{\lambda})$ is an optimal solution on $V \times M$ for the problem obtained by substituting in problem (p) the system {for $1 \leqq l \leqq m_0$, $F_l(y, \lambda) = 0$} to the system (1), where $F_l(y, \lambda)$ is defined by

$$(16) \qquad F_l(y, \lambda) = \sum_u \lambda_u(x)[A_l(u) \cdot y(x) + f_l(y(x), u(x), x)].$$

LEMMA 1. *The functions F_l are defined in $V \times E$, with value in the dual of $H_0^1(\Omega)$; they are strongly differentiable at $(\bar{y}, \bar{\lambda})$ with differential*

$$(17) \qquad F_l'(\bar{y}, \bar{\lambda}) \cdot (y, \lambda) = F_l(\bar{y}, \lambda) + A_l(\bar{u}) \cdot y + \sum_{k=1}^{m} \frac{\partial f_l}{\partial z_k}(\bar{y}(x), \bar{u}(x), x) y_k.$$

Proof. The proof is in [5, Lemma 3.4].

LEMMA 2. *For* $0 \leqq i \leqq i_1$, *the functions* $\int_\Omega g_i(y(x), x)\, dx$ *are strongly differentiable at* $\bar{y}$ *with differential*

$$l_i(y) = \int_\Omega \sum_{k=1}^{m} \frac{\partial g_i}{\partial z_k}(\bar{y}(x), x) y_k(x)\, dx. \tag{18}$$

Proof. The proof is in [5, Lemma 3.5].

Second part of the proof of the theorem. For $1 \leqq k \leqq m$, the function

$$h_k(y) = \begin{cases} -1, & \text{if } \Omega_k \text{ is negligible,} \\ \operatorname{ess\,sup}_{\Omega_k} (-y(x)), & \text{if not,} \end{cases} \tag{19}$$

is defined from $L^2(\Omega)$ into $\mathbb{R} \cup \{+\infty\}$; it is convex and lower semicontinuous: the set $\{(\alpha, y) \in \mathbb{R} \times L^2(\Omega);\ h_k(y) \leqq \alpha\}$ is a closed convex set. And the condition (4) of problem (P) is equivalent to

$$\text{for } 1 \leqq k \leqq m: \quad h_k(y_k) \leqq 0. \tag{20}$$

All the results of [4] remain valid in the case of functions which are sums of strongly differentiable functions and of lower semicontinuous convex functions, if some of the one-dimensional components of the convex functions are valued in $\mathbb{R} \cup \{+\infty\}$: there is no modification of the notations, the statements and the proofs, which are all independent of the possible infinite value.

The arguments of [6] show that $(\bar{\lambda}, \bar{y})$ is an optimal solution of the problem obtained by substitution in problem (p) of (1) by $\{\text{for } 1 \leqq l \leqq m_0 : F_l(y, \lambda) = 0\}$ and of (4) by (20).

It follows from Assumption 4 that the differential of $(F_1, \cdots, F_{m_0})$ at $(\bar{y}, \bar{\lambda})$ is a surjection from $V \times M$ onto V_0', and the regularity assumption (3.4 of [4]) is satisfied: the proof is the same as that in [6, Prop. 5.1]. According to [4, Theorem 3.5] there exist real numbers $\alpha_i (0 \leqq i \leqq i_1)$ and $\beta_k (1 \leqq k \leqq m)$, and p_l belonging to the bidual of $H_0^1(\Omega)(1 \leqq l \leqq m_0)$, which are not all zero, such that:

$$\text{for } 0 \leqq i \leqq i_0: \quad \alpha_i \geqq 0; \quad \text{and for } 1 \leqq k \leqq m: \quad \beta_k \geqq 0, \tag{21}$$

$$\text{for } 1 \leqq i \leqq i_0: \quad \alpha_i \int_\Omega g_i(\bar{y}(x), x)\, dx = 0; \text{and for } 1 \leqq k \leqq m: \quad \beta_k h_k(\bar{y}_k) = 0, \tag{22}$$

and for each $(y, \lambda) \in V \times M$ verifying $\|y - \bar{y}\| \leqq 1$ and $\|\lambda - \bar{\lambda}\| \leqq 1$:

$$\sum_i \alpha_i l_i(y - \bar{y}) + \sum_k \beta_k [h_k(y_k) - h_k(\bar{y}_k)] + \sum_l p_l \cdot F_l'(\bar{y}, \bar{\lambda}) \cdot (y - \bar{y}, \lambda - \bar{\lambda}) \geqq 0. \tag{23}$$

The p_l are identical to elements of $H_0^1(\Omega)$. Conclusions 2 and 3 of the theorem result from (21) and (22). The α_i and p_l are not all zero: otherwise the β_k would be also zero (from (22) and (23)), which is impossible. For $\lambda = \bar{\lambda}$, $\|y_k - \bar{y}_k\| \leqq 1$ and $y_r = \bar{y}_r (r \neq k)$, one obtains, with inequality (23) and notation (13).

$$\gamma_k \cdot (y_k - \bar{y}_k) + \beta_k [h_k(y_k) - h_k(\bar{y}_k)] \geqq 0. \tag{24}$$

If $h_k(\bar{y}_k) \neq 0$, then $\beta_k = 0$ (from (22)) and consequently $\gamma_k = 0$. If $h_k(\bar{y}_k) = 0$, relation (24) gives $\gamma_k \cdot \bar{y}_k \leqq 0$ and $\gamma_k \cdot \bar{y}_k \geqq 0$ for $y_k = t\bar{y}_k$, $t < 1$ and $t > 1 : \gamma_k$ verifies relation (14); and for $y_k(x) \geqq 0$ a.e. in Ω_k, $h_k(y_k) \leqq 0$ and $\gamma_k \cdot y_k \geqq 0$: this is obtained with $ty_k + (1 - t)\bar{y}_k$, $0 < t < 1$ such that $t\|y_k - \bar{y}_k\| \leqq 1$.

To obtain the last conclusion, let us consider $u \in S$ and any measurable subset P of Ω; for $y = \bar{y}$ and the family $\lambda \in M$:

$$\lambda_u(x) = t, \quad \text{if } x \in P, \quad \text{and} \quad \lambda_u(x) = 0 \quad \text{if } x \notin P,$$
$$\lambda_{\bar{u}}(x) = 1 - t \quad \text{if } x \in P, \quad \text{and} \quad \lambda_{\bar{u}}(x) = 1 \quad \text{if } x \notin P,$$
$$\lambda_v = 0 \quad \text{for } v \neq u \quad \text{and} \quad v \neq \bar{u}$$

inequality (23) is

$$t \int_P (H(p, u(x), x) - H(p, \bar{u}(x), x))\, dx \geqq 0;$$

this inequality holds for any measurable subset P of Ω and consequently

$$H(p, u(x), x) \geqq H(p, \bar{u}(x), x) \quad \text{a.e. in } \Omega.$$

The last conclusion results from Assumption 5. The proof of the theorem is complete.

Assumption 6. For each $y' \in V'$, there exists at least one solution $p \in V_0$ of the linear system

$$A(\bar{u})^* \cdot p + \frac{\partial f}{\partial z}(\bar{y}(x), \bar{u}(x), x)^* \cdot p = y'.$$

COROLLARY. *Let* $(\bar{y}, \bar{u})$ *be an optimal solution of problem* (p), *with* $\bar{u}$ *a regular control, and assumptions* 1 *to* 6 *be satisfied. Then there exist* $\alpha_i \in \mathbb{R}$ $(0 \leqq i \leqq i_1)$, $\tilde{p} \in V_0$ *and* $\tilde{q} \in V_0$ *such that*

(1) $\tilde{p}, \tilde{q}$ *and* $\alpha_i (0 \leqq i \leqq i_1)$ *are not all zero*;

(2) *for* $0 \leqq i \leqq i_0$, α_i *is nonnegative*;

(3) *for* $1 \leqq i \leqq i_0$, $\alpha_i \int_\Omega g_i(\bar{y}(x), x)\, dx = 0$;

(4) $\tilde{p}$ *is a solution of*

$$A(\bar{u})^* \cdot \tilde{p} + \frac{\partial f}{\partial z}(\bar{y}(x), \bar{u}(x), x)^* \cdot \tilde{p} + \sum_i \alpha_i \frac{\partial g_i}{\partial z}(\bar{y}(x), x) = 0;$$

(5) $\tilde{q}$ *satisfies the relations*

$$\left(A(\bar{u})^* \cdot \tilde{q} + \frac{\partial f}{\partial z}(\bar{y}(x), \bar{u}(x), x)^* \cdot \tilde{q}\right) \cdot \bar{y} = 0;$$

and for each $y \in V$ *such that* $y_k(x) \geqq 0$ *a.e. on* $\Omega_k (1 \leqq k \leqq m)$

$$\left(A(\bar{u})^* \cdot \tilde{q} + \frac{\partial f}{\partial z}(\bar{y}(x), \bar{u}(x), x)^* \cdot \tilde{q}\right) \cdot y \geqq 0.$$

(6) *The Hamiltonian* $H(\tilde{p} + \tilde{q}, v, x)$ *attains its minimum on* $U(x)$ *at* $\bar{u}(x)$, *a.e. in* Ω.

Proof. For $\gamma = (\gamma_1, \cdots, \gamma_m)$, there exists $\tilde{q} \in V_0$ such that

$$A(\bar{u})^* \cdot \tilde{q} + \frac{\partial f}{\partial z}(\bar{y}(x), \bar{u}(x), x)^* \cdot \tilde{q} = \gamma.$$

Setting $\tilde{p} = p - \tilde{q}$, we see that the corollary is the transcription of the theorem's results.

4. Extensions. The method which has been used, works in more general cases. With additional terms in the operator $A(u)$ like

$$\sum_{i,k} C_{ikl}(x, u(x)) \frac{\partial y_k}{\partial x_i} + \frac{\partial}{\partial x_i}(d_{ikl}(x, u(x)) y_k(x))$$

the single difference is the corresponding Assumptions 6 and 3. If the control appears in the integrals (cost function and constraints):

$$\int_\Omega g_i(y(x), u(x), x)\, dx$$

one need conditions

$$\left|\frac{\partial g_i}{\partial z_k}(z, u(x), x)\right| \leqq \alpha_u(x) + \beta_u |z|^{q-1}$$

with $\alpha_u \in L^r(\Omega)$, in the definition (8) of regular controls, and convexity conditions (Assumption 3) for the modified function $h(x, v, Y) = (Z, \xi)$, where Z is defined by Notation 1 and $\xi \in \mathbb{R}^{i_1+1}$ is defined by its components

$$\text{for } 0 \leqq i \leqq i_1: \quad \xi_i = g_i(y_0, v, x).$$

For parabolic systems and measurable subsets E_k of $]0, T[\times\Omega$, constraints like

$$y_k(t, x) \geqq 0 \quad \text{a.e. in } E_k$$

are equivalent to

$$\operatorname*{ess\,sup}_{E_k} (-y_k(t, x)) \leqq 0;$$

the same method applies and one obtains the corresponding modified results of [6].

One important and easy generalization is possible to the case of constraints like:

$$\varphi_i\left(x, y_1(x), \cdots, y_m(x), \frac{\partial y_1}{\partial x_1}(x), \cdots, \frac{\partial y_1}{\partial x_n}(x), \frac{\partial y_2}{\partial x_1}(x), \cdots, \frac{\partial y_m}{\partial x_n}(x)\right) \leqq 0 \quad \text{a.e. on } \Omega_i$$

where φ_i is measurable with respect to x and lower semi-continuous and convex with respect to the other variables.

REFERENCES

[1] N. U. AHMED AND K. L. TED, *Necessary conditions for optimality of Cauchy problems for parabolic partial differential systems*, this Journal, 13 (1975), pp. 981–993.

[2] I. K. GOGODZE, *Necessary conditions for optimal control of an elliptic problem with integral restrictions*, Bull. Acad. Sci. Georgian SSR, 81 (1976), pp. 17–20. (In Russian.)

[3] J. L. LIONS, *Contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles*, Dunod et Gauthiers-Villars, Paris, 1969.

[4] P. MICHEL, *Problème des inégalités. Applications à la programmation et au contrôle optimal*, Bull. Soc. Math. France, 101 (1973), pp. 413–439.

[5] ———, *Condition nécessaire d'optimalité pour des systèmes d'équations elliptiques non linéaires*, Quatrièmes journées de contrôle (Metz, 18–21 Mai 1976), polycopié Université de Metz, 1976.

[6] ———, *Condition nécessaire d'optimalité pour un système régi par des équations aux dérivées partielles non linéaires de type parabolique*, Bull. Soc. Math. France, 105 (1977), pp. 65–88.

[7] T. ZOLEZZI, *Necessary conditions for optimal controls of elliptic or parabolic problems*, this Journal, 10 (1972), pp. 594–607.

© 1980 Society for Industrial and Applied Mathematics
0363-0129/80/1801-0007$01.00/0

ADDENDUM: A NOTE ON THE LACK OF EXACT CONTROLLABILITY FOR MILD SOLUTIONS IN BANACH SPACES*

ROBERTO TRIGGIANI†

The main result—Theorem 1.2—of [1], concerning the lack of exact controllability in finite time for the system

$$S(t)x_0+\int_0^t S(t-\tau)Bu(\tau)\,d\tau,$$

holds true within the class of locally L_p-controls for all $p>1$, but not for $p=1$ as claimed in [1]. Actually, lack of exact controllability over the finite interval $[0, T]$ is guaranteed even within the class of controls which are locally L_p, $p>1$, just *near* T.

As pointed out in [2], a slip has occurred in the inequality at the bottom of page 408 of [1] and, for $p>1$, we remedy it by the Hölder inequality as follows (we use the notation of [1]):

$$\|Q\tilde{u}-Q_\varepsilon\tilde{u}\|=\left\|\int_{T-\varepsilon}^{T} S(T-t)Bu(t)\,dt\right\|\leqq M_\alpha\, e^{\alpha\varepsilon}\|B\|\int_0^T \chi_{[T-\varepsilon,T]}\|u(t)\|\,dt$$
$$\leqq M_\alpha\, e^{\alpha\varepsilon}\|B\|\varepsilon^{1/q}\|\tilde{u}\|_p\to 0\quad \text{as } \varepsilon\to 0$$

where χ is the characteristic function of $[T-\varepsilon, T]$ and

$$\|\tilde{u}\|_p=\left\{\int_0^T \|u(t)\|^p\,dt\right\}^{1/p},\qquad 1/p+1/q=1,\quad p>1.$$

Hence Q, being the uniform limit of compact operators, is indeed compact as an operator from $L_p[[0, T], U]$, $p>1$, into X.

However, for $p=1$ and B onto X, the operator Q is not compact. In fact, we shall show that, in this case, the image under Q of a finite sphere in $L_1[[0, T], U]$ is dense in the unit sphere of X. Let v be a unit vector in X. By strong continuity of $S(t)$, given $\delta>0$, there is $\varepsilon>0$ such that

$$\sup_{0\leqq t\leqq T-\varepsilon}\|S(T-t)v-v\|<\delta.$$

There exists u in U such that $Bu=v$. By a version of the open mapping theorem, [3, Lemma 9, p. 194] as v runs over the unit sphere of X, the corresponding vectors u can be taken in a finite sphere of U of radius, say, r. Define a control $u(t)$ by: $u(t)\equiv 0$ for $0\leqq t\leqq T-\varepsilon$ and $u(t)\equiv u/\varepsilon$ for $T-\varepsilon<t\leqq T$. Then

$$\|\tilde{u}\|_1=\int_0^T \|u(t)\|\,dt=\|u\|\leqq r$$

and

$$\left\|\int_0^T S(T-t)Bu(t)\,dt-v\right\|\leqq\frac{1}{\varepsilon}\int_{T-\varepsilon}^{T}\|S(T-t)v-v\|\,dt<\delta.\qquad \text{Q.E.D.}$$

We finally remark that, as pointed out in [4, top paragraph, p. 481], when B is onto X and $S(t)$ is a group, exact controllability on any $[0, T]$ does hold. In fact a controller

* This Journal, 15 (1977), pp. 407–411. Received by the editors March 16, 1979.

† Department of Mathematics, Iowa State University, Ames, Iowa 50011.

$u(t)$ defined by $Bu(t)=S(t-T)x_1/T$ (which is continuous in t) steers the origin to the desired final state x_1 over $[0, T]$.

REFERENCES

[1] R. Triggiani, *A note on the lack of exact controllability for mild solutions in Banach spaces*, this Journal, 15 (1977), pp. 407–411.

[2] J. E. Allahverdiev and A. V. Shapiro, *A remark on exact controllability in Banach spaces*, unpublished manuscript.

[3] H. L. Royden, *Real Analysis*, 2nd Ed., Macmillan, New York, 1968.

[4] R. Triggiani, *Controllability and observability in Banach space with bounded operators*, this Journal, 13 (1975), pp. 462–491.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 2, March 1980

THE "BANG-BANG" PRINCIPLE FOR THE TIME-OPTIMAL PROBLEM IN BOUNDARY CONTROL OF THE HEAT EQUATION*

E. J. P. GEORG SCHMIDT†

Abstract. Following previous work by H. O. Fattorini, J. Henry and the present author it is proved that the time-optimal controls associated with arbitrary reachable target temperature distributions in boundary control for the heat equation (with bounds on the admissible controls) are "bang-bang".

1. Introduction. In this paper the "bang-bang" property of time optimal controls is proved in the boundary control problem for the heat equation with arbitrary reachable target, under a mild assumption on the bound to which the controls are subjected (satisfied, for example, if that bound is not the smallest one under which the target is reachable). This property was first proved by Fattorini [3], for target functions satisfying a deep (and not in general easily verifiable) sufficient condition due, in its general form, to Russell [9]. Subsequently Schmidt in a paper [10] submitted to this journal in 1977, and currently under revision, proved it for stationary (or steady state) targets. Recently Henry [7], working in the context of distributed controls, presented an argument, containing essentially new ideas, to prove the general result; unfortunately the proof appears to contain a serious gap. This paper draws on ideas occurring in all the work cited above; in particular we follow Henry in defining a norm on the (invariant) reachable set to obtain a Banach space. Instead of applying the separation argument which yields the "bang-bang" principle in that space, as Henry tries to do, we work in a certain subspace, using an idea which we had applied previously to stationary targets. We could have used a subspace employed by Fattorini, but instead work with the subspace generated by states reachable from arbitrary initial states with 0 control, thus bypassing the use of deep results from the theory of moments.

2. The heat equation and its solution. Let Ω be a bounded domain in R^n, whose boundary $\partial\Omega$ is a C^∞ manifold. Let Δ denote the Laplacian in 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 B "=" $a(\partial/\partial\nu)+1$. We consider the following initial boundary value problem, which describes the evolution of the temperature $u(x, t)$ at point x and time t when the ambient temperature at the boundary is $f(x, t)$ and the initial temperature distribution is $u_0(x)$,

$$\begin{aligned} &\frac{\partial u}{\partial t}(x, t) = \Delta u(x, t) && \text{for } x \in \Omega,\ t > 0, \\ &Bu(x, t) = f(x, t) && \text{for } x \in \partial\Omega,\ t > 0, \\ &u(x, 0) = u_0(x) && \text{for } x \in \Omega. \end{aligned} \tag{1}$$

Let H denote the state space $L_2(\Omega)$ (with inner product $(\cdot\,,\cdot)$ and norm $\|\cdot\|$), and L_∞ denote the space of controls $L_\infty(\partial\Omega\, x(0, \infty))$ (with norm $\|\cdot\|_\infty$). In [3] Fattorini introduced a notion of weak solution for the case $a = 0$ which can easily be generalized to the case $a > 0$. Combining his methods with those of Glashoff and Weck [6] who studied the latter case, one can show that for each $u_0 \in H$ and $f \in L_\infty$, (1) has a unique

* Received by the editors September 13, 1978.

† Department of Mathematics, McGill University, Montreal, Quebec, Canada H3C 3G1. This work was supported by the National Research Council of Canada under Grant A7271.

weak solution with a lot of relevant properties which will be summarized in Theorem 1 below.

We recall first that the self-adjoint operator obtained when the Laplacian is defined on a suitable domain of functions satisfying the boundary condition $Bu(x)=0$ has a complete orthonormal system $\{\varphi_k\}_{k=1}^{\infty}$ of eigenfunctions corresponding to negative eigenvalues $\{-\lambda_k\}_{k=1}^{\infty}$ arranged in decreasing order:

$$\Delta\varphi_k = -\lambda_k\varphi_k \quad \text{in } \Omega, \qquad B\varphi_k = 0 \quad \text{on } \partial\Omega. \tag{2}$$

Moreover, asymptotically,

$$\lambda_k \sim Ck^{2/n} \tag{3}$$

where C is a certain positive constant, and the eigenfunctions belong to $C^{\infty}(\bar{\Omega})$, and satisfy estimates

$$\sup_{x\in\bar{\Omega}} |D^r\varphi_k(x)| \leqq C_r k^{m_r}, \tag{4}$$

where D^r is any partial derivative of order r and C_r and m_r are suitable positive constants. For more details and proofs see Agmon [1].

Finally, before stating the theorem, we define a family of translation operators J_s in L_∞ by

$$\begin{aligned} &[J_sf](t) = f(s+t) \quad \text{if } s \geqq 0,\\ &[J_sf](t) = \begin{cases} 0 & \text{for } 0 \leqq t \leqq |s|,\\ f(s+t) & \text{for } t > |s|, \end{cases} \quad \text{if } s < 0. \end{aligned} \tag{5}$$

THEOREM 1. *For each $u_0 \in H$ and $f \in L_\infty$ there exists a unique weak solution $u(x, t)$ to* (1); *this function belongs to $L_2(\Omega \times (0, T))$ for each finite $T > 0$, and also to $C^{\infty}(\Omega \times (0, \infty))$. For each $t > 0$, $u(\cdot, t) \in H$ and moreover can be represented as*

$$u(\cdot, t) = V_tu_0 + S_tf,$$

where

(a) *$\{V_t\}_{t\geqq 0}$ is a strongly continuous semi-group of linear contractions on H;*

(b) *$S_t\colon L_\infty \to H$ is continuous from the weak*-topology of L_∞ to the norm topology on H;*

(c) *for each $t_1, t_2 \geqq 0$ one has*

$$S_{t_1+t_2}f = V_{t_2}S_{t_1}f + S_{t_2}J_{t_1}f; \tag{6}$$

(d) *if $u_0(x)$ and $f(x, t)$ are essentially bounded below by m* (*or above by M*) *the same is true for $u(x, t)$*;

(e) $$V_tc + S_tc = c \tag{7}$$

(*where c is to be interpreted as the constant function*);

(f) $$\|S_tf\| \leqq \|V_t(\|f\|_\infty) - \|f\|_\infty\|, \tag{8}$$

where $\|f\|_\infty$ is to be interpreted as a constant function on Ω;

(g) $$V_tu_0 = \sum_{k=1}^{\infty} e^{-\lambda_k t}(u_0, \varphi_k)\varphi_k; \tag{9}$$

(*h*) (10) $$S_tf = \sum_{k=1}^{\infty} \left[\int_0^t \int_{\partial\Omega} e^{-\lambda_k(t-s)} \varphi_k^{\partial}(y) f(y, s)\, dS_y\, ds \right] \varphi_k,$$

where $\varphi_k^{\partial}(y)$ is equal to $\varphi_k(y)/a$ if $a > 0$, and $-\partial\varphi_k/\partial\nu$ if $a = 0$, and where dS_y denotes an element of area of $\partial\Omega$.

We comment only on the proof of (f). From the maximum principle (d) it follows from

$$-\|f\|_\infty \leqq f(x, t) \leqq \|f\|_\infty$$

that also

$$-[S_t\|f\|_\infty](x) \leqq [S_tf](x) \leqq [S_t\|f\|_\infty](x),$$

and hence using also the obvious property (e) (which like the maximum principle (d) depends on the fact that the constant a in B appears where it does) implies that

$$|[S_tf](x)| \leqq [S_t\|f\|_\infty](x) = \|f\|_\infty - [V_t\|f\|_\infty](x),$$

from which the estimate (8) follows at once.

3. Properties of the reachable set. We define, for each $u_0 \in H$, $t > 0$ and $L \subset L_\infty$

(11) $$R_t(u_0; L) = \{v \in H: \text{there exists } f \in L \text{ with } v = V_tu_0 + S_tf\}.$$

It is well known that $R_t(u_0, L_\infty)$ is always dense in H (a nice proof is given in MacCamy, Mizel and Seidman [8]). It is also known that the function 0 always belongs to $R_t(u_0, L_\infty)$. This property, known as null controllability, was proved for general domains by Russell in [9], but can also be deduced from null controllability for balls (established by more elementary means in Fattorini and Russell [5]) using an extension argument suggested by Seidman in [12]. An immediate consequence, in fact a reformulation of the property of null controllability, is

(12) $$V_t(H) \subset S_t(L_\infty), \quad \text{for each } t > 0.$$

From this fact one obtains a simple proof of the invariance of the reachable set which was proved by Henry in [7] using a previous result of Fattorini [2], and which is also proved in Seidman [13].

THEOREM 2. *$R_t(u_0; L_\infty)$ is the same set for all $u_0 \in H$ and $t > 0$.*

Proof. Note first that an immediate consequence of (12) is the fact that $R_t(u_0; L_\infty) = R_t(0; L_\infty)$, so we only have to prove that $R_s(0; L_\infty) = R_t(0; L_\infty)$ where $s < t$. Given $S_tf \in R_t(0; L_\infty)$ it follows from (6) and (12) that

$$\begin{aligned} S_tf &= V_sS_{t-s}f + S_s(J_{t-s}f) \\ &= S_s[f_1 + (J_{t-s}f)] \end{aligned}$$

for some $f_1 \in L_\infty$, so that $S_tf \in R_s(0; L_\infty)$ and $R_t(0; L_\infty) \subset R_s(0; L_\infty)$. Conversely, given $S_sf \in R_s(0; L_\infty)$ one can deduce from (6) that $S_sf = S_t(J_{s-t}f) \in R_t(0; L_\infty)$; hence also $R_s(0; L_\infty) \subset R_t(0; L_\infty)$.

Now let R denote the set $R_t(0; L_\infty)$. Since R is the range of S_t for each $t > 0$, one can define on R norms

(13) $$\|v\|_t = \inf \{\|f\|_\infty : f \in L_\infty, S_tf = v\}.$$

In this way one obtains a Banach space $(R, \|\cdot\|_t)$ isometrically isomorphic to the quotient space $L_\infty/N(S_t)$ (where $N(S_t)$ is the null space of S_t). These norms are in fact equivalent since, if $s < t$, the identity $S_sf = S_t(J_{s-t}f)$ implies that

$$\|v\|_t \leqq \|v\|_s, \tag{14}$$

so that the closed graph theorem implies the equivalence of the two norms. For the sake of definiteness we define $\|v\|_R = \|v\|_1$. The injection $I: R \to H$ is continuous; for each $f \in L_\infty$ such that $v = S_1 f$ one has

$$\|v\| = \|S_1 f\| \leqq \|S_1\|_{B(L_\infty, H)} \|f\|_\infty,$$

where $B(L_\infty, H)$ denotes the bounded, linear operators from L_∞ to H and $\|\cdot\|_{B(L_\infty,H)}$ is the operator norm. Hence $\|v\| \leqq \|S_1\|_{B(L_\infty,H)}\|v\|_R$. Since R is dense in H, the adjoint map $I^*: H \to R^*$ is 1–1.

4. The "bang-bang" property of time optimal controls. Let $L_M = \{f \in L_\infty : \|f\|_\infty \leqq M\}$. Given $u_0, u_1 \in H$ such that $u_1 \in R_t(u_0, L_M)$ for some $t > 0$, the time-optimal control problem is to find $f_* \in L_M$ such that

$$u_1 = V_{t_*}u_0 + S_{t_*}f_*, \quad \text{where } t_* = \inf\{t > 0 : u_1 \in R_t(u_0; L_M)\}.$$

The existence of such a control f_* is standard (and follows easily from the properties of V_t and S_t described in Theorem 1). The "bang-bang" property of f_* is that f_* is an extreme point of L_M in L_∞; in other words that $f_*(x, t) = \pm M$ almost everywhere on $\partial\Omega \times (0, t_*)$. The precise statement of the theorem, and its proof, involves the subspace X of R which is obtained as the closure in R of $\bigcup_{t>0} V_t(H)$.

THEOREM 3. *Given $u_0 \in H$ and $u_1 \in R$. Suppose $u_1 \in R_t(u_0, L_M)$ for some $t > 0$, where*

$$M > \operatorname{dist}_R (u_1, X) = \inf\{\|u_1 - v\|_R : v \in X\}. \tag{15}$$

Then there exists a nontrivial solution $w(x, t)$ to the adjoint heat equation

$$\begin{aligned} &\frac{\partial w}{\partial t}(x, t) + \Delta w(x, t) = 0 \quad &&\text{for } x \in \Omega,\ t \in (0, t_*), \\ &Bw(x, t) = 0 &&\text{for } x \in \partial\Omega,\ t \in (0, t_*) \end{aligned} \tag{16}$$

such that the function $w^\partial(x, t)$, defined on the boundary by $w^\partial(x, t) = (w(x, t))/a$ if $a > 0$ and by $w^\partial(x, t) = -(\partial w/\partial \nu)(x, t)$ if $a = 0$, does not vanish on any set of positive measure, and such that the time optimal control f_* is given by

$$f_*(x, t) = M \operatorname{sgn} w^\partial(x, t) \quad \text{for } x \in \Omega,\ t \in (0, t_*). \tag{17}$$

We precede the proof of this theorem by three lemmas.

LEMMA 1. *Let $u_1 \in R$, and suppose $M > \operatorname{dist}_R (u_1, x)$. Then for each $t > 0$ there exists $v_1 \in x$ and $g_1 \in L_\infty$ such that $u_1 = v_1 + S_t g_1$ and $\|g\|_1 < M$.*

Proof. From the hypothesis and the definition of X it follows that there exists $v \in X$ with $\|u_1 - v\|_R < M$. By the definition of $\|\cdot\|_R = \|\cdot\|_1$, there exists $g \in L_\infty$ such that $u_1 - v = S_1 g$ with $\|g\|_\infty < M$. Now, if $t \geqq 1$, $S_1 g = S_t(J_{1-t}g)$ and so $u_1 = v + S_t(J_{1-t}g)$; while, if $t < 1$, $S_1 g = V_t(S_{1-t}g) + S_t(J_{1-t}g)$ and so in this case $u_1 = [v + V_t(S_{1-t}g)] + S_t(J_{1-t}g)$. In both cases the assertion holds.

LEMMA 2. *Let $V_{t,X}: H \to X$ denote the operator V_t regarded as a map of H into X.*

(a) *$V_{t,X}$ is a bounded, linear operator.*

(b) *Suppose $l \in X^*$ and $l \neq 0$; there exists $\varepsilon^0 > 0$ such that*

$$V_{\varepsilon,X}{}^* l \neq 0 \quad \text{for each } \varepsilon < \varepsilon^0.$$

Proof. The boundedness of $V_{t,X}$ follows from the closed graph theorem, using the continuity of the operator $V_t\colon H \to H$ and of the immersion $I\colon R \to H$.

Suppose now that (b) does not hold. Then there exists a sequence $\varepsilon_n \downarrow 0$ such that $V_{\varepsilon_n,X}{}^*l = 0$. For each $t > 0$ one then has, with $\varepsilon_n < t$, $V_{t,X} = V_{\varepsilon_n,X}V_{t-\varepsilon_n}$ and hence taking adjoints, $V_{t,X}{}^*l = V_{t-\varepsilon_n}{}^*V_{\varepsilon_n,X}{}^*l = 0$ so that $V_{t,X}{}^*l = 0$ for each $t > 0$. Thus for any $u \in H$

$$(V_{t,X}{}^*l, u) = \langle l, V_{t,X}u\rangle = 0,$$

and by definition of X it follows that $l = 0$, a contradiction. (Note: here $\langle l, v\rangle$ denotes the action of $l \in X^*$ on $v \in X$.)

The final lemma is of interest in itself.

LEMMA 3. *Suppose $u_1 \in R$ and that there exists $f \in L_\infty$ with $\|f\|_\infty < M$ such that $u_1 = V_tu_0 + S_tf$. Then for $\varepsilon > 0$ sufficiently small one can find $f_\varepsilon \in L_\infty$ with $\|f\|_\infty < M$ such that $u_1 = V_{t-\varepsilon}u_0 + S_{t-\varepsilon}f_\varepsilon$.*

Proof. Using the semi-group property of $\{V_t\}_{t\geqq 0}$ and (6) one has

$$u_1 = V_{t-\varepsilon}u_0 + S_{t-\varepsilon}(J_\varepsilon f) + V_{t-\varepsilon}(V_\varepsilon u_0 - u_0) + V_{t-\varepsilon}(S_\varepsilon f).$$

Let $0 < \delta < \frac{1}{2}(M - \|f\|_\infty)$. We show that for ε sufficiently small, one can find $f_{1,\varepsilon}$ and $f_{2,\varepsilon}$ in L_δ such that

$$V_{t-\varepsilon}(V_\varepsilon u_0 - u_0) = S_{t-\varepsilon}f_{1,\varepsilon},$$

and

$$V_{t-\varepsilon}(S_\varepsilon f) = S_{t-\varepsilon}f_{2,\varepsilon}.$$

Then, letting $f_\varepsilon = J_\varepsilon f + f_{1,\varepsilon} + f_{2,\varepsilon}$ one has the desired result. Note first that, setting $w_\varepsilon = V_\varepsilon u_0 - u_0$ or $S_\varepsilon f$ it follows from the strong continuity of the semi group $\{V_t\}_{t\geqq 0}$ and from (8) that $\|w_\varepsilon\| \to 0$ as $\varepsilon \to 0$. Now since $V_{t,X} = V_{s,X}V_{t-s}$ when $s < t$ one has

$$\|V_{t,X}\|_{B(H,X)} \leqq \|V_{s,X}\|_{B(H,X)}\|V_{t-s}\|_{B(H,H)} \leqq \|V_{s,X}\|_{B(H,X)}.$$

Hence by the equivalence of the norms $\|\cdot\|_t$ and in particular by (14) it follows that for $\varepsilon < t/2$

$$\begin{aligned}\|V_{t-\varepsilon,X}w_\varepsilon\|_{t-\varepsilon} &\leqq \|V_{t-\varepsilon,X}w_\varepsilon\|_{t/2}\\ &\leqq C\|V_{t-\varepsilon,X}w_\varepsilon\|_R\\ &\leqq C\|V_{t-\varepsilon,X}\|_{B(H,X)}\|w_\varepsilon\| \leqq C\|V_{t/2,X}\|_{B(H,X)}\|w_\varepsilon\|;\end{aligned}$$

thus for ε sufficiently small $\|V_{t-\varepsilon,X}w_\varepsilon\|_{t-\varepsilon} < \delta$. Applying this estimate to the two alternatives for w_ε, and using the definition of $\|\cdot\|_{t-\varepsilon}$ one obtains $f_{1,\varepsilon}$ and $f_{2,\varepsilon}$.

We now come to the

Proof of Theorem 3. By Lemma 1 we can write $u_1 = v_1 + S_{t_*}g_1$ with $\|g_1\|_\infty < M$ and $v_1 \in X$. Then

$$C = \{v \in X\colon \text{there exists } f \in L_M \text{ with } v = S_{t_*}(f - g_1)\} \tag{18}$$

is a closed convex set in X. Since $0 = S_{t_*}(g_1 - g_1)$ and $g_1 \in L_M$ one has $0 \in C$. Moreover 0 lies in the interior of C; for the norm equivalence $\|v\|_{t_*} \leqq C\|v\|_R$ implies that if $\|v\|_R < C^{-1}[M - \|g_1\|_\infty] = \delta$ one can find $h \in L_\delta$ such that $v = S_{t_*}h$, in which case $v = S_{t_*}(f - g_1)$ with $f = g_1 + h \in L_M$, so that $v \in C$.

Since $u_1 = V_{t_*}u_0 + S_{t_*}f_* = v_1 + S_{t_*}g_1$ one has that $v_1 - V_{t_*}u_0 = S_{t_*}(f_* - g_1)$; since $f_* \in L_M$ and $v_1 - V_{t_*}u_0 \in X$ it follows that $v_1 - V_{t_*}u_0 \in C$. For the separation argument which proves the "bang-bang" property of f_*, it remains to prove that $v_1 - V_{t_*}u_0$ is a boundary point of C. Suppose $v_1 - V_{t_*}u_0$ is not a boundary point; then, since 0 lies in the interior of C, there exists r with $0<r<1$ such that $r^{-1}(v_1 - V_{t_*}u_0) \in C$. Thus, by the definition of C, there exists f in L_M with $v_1 - V_{t_*}u_0 = rS_{t_*}(f-g_1)$. Letting $f_1 = rf + (1-r)g_1$ one has $v_1 - V_{t_*}u_0 = S_{t_*}(f_1 - g_1)$ where now, since $\|g_1\|_\infty < M$ also $\|f_1\|_\infty < M$. Thus $u_1 = v_1 + S_{t_*}g_1 = V_{t_*}u_0 + S_{t_*}f_1$ with $\|f_1\|_\infty < M$. By Lemma 3 this implies that $t_* \neq \inf\{t\colon u_1 \in R_t(u_0; L_M)\}$, a contradiction which implies that $v_1 - V_{t_*}u_0$ is indeed a boundary point.

Since C is a convex set with nonempty interior in X, and since $v_1 - V_{t_*}u_0 = S_{t_*}(f_* - g_1)$ is a boundary point there exists $l \in X^*$ with $l \neq 0$ such that for each $v \in C$

$$\langle l, S_{t_*}(f_* - g_1) - v\rangle \geqq 0.$$

We choose elements v of C having a particular form. Let χ_ε denote the characteristic function of $\partial\Omega \times (0, t_* - \varepsilon)$, with $\varepsilon > 0$. Then for each $f \in L_M$, $f_\varepsilon = (1-\chi_\varepsilon)f_* + \chi_\varepsilon f \in L_M$ and

$$S_{t_*}(f_\varepsilon - g_1) = S_{t_*}(f_* - g_1) + S_{t_*}[\chi_\varepsilon(f - f_*)]$$

belongs to X since both $S_{t_*}(f_* - g_1)$ and $S_{t_*}[\chi_\varepsilon(f-f_*)] = V_\varepsilon S_{t_*-\varepsilon}[\chi_\varepsilon(f-f_*)]$ do. Hence $v = S_{t_*}(f_\varepsilon - g_1) \in C$, and thus

(19) $$\langle l, S_{t_*}[\chi_\varepsilon(f_* - f)]\rangle = \langle l, S_{t_*}(f_* - g_1) - S_{t_*}(f_\varepsilon - g_1)\rangle \geqq 0,$$

for each $f \in L_M$ and $\varepsilon > 0$. We need to transform this inequality. Fix $\varepsilon > 0$ and consider $\langle l, S_{t_*}h\rangle$ for any $h \in L_\infty$ with essential support in $\partial\Omega \times (0, t_* - \varepsilon)$. Since $S_{t_*}h = V_{\varepsilon/2,X}S_{t_*-\varepsilon/2}h$ one has

$$\langle l, S_{t_*}h\rangle = \langle S_{t_*-\varepsilon/2}{}^* V_{\varepsilon,X/2}{}^* l, h\rangle.$$

If $\varepsilon/2 < \varepsilon_0$, the critical constant corresponding to l in Lemma 2, $v_\varepsilon = V_{\varepsilon/2,X}{}^* l \neq 0$. It is then easy to verify, using the representation (10), and the estimates (3) and (4) which justify the application of Fubini's theorem, that

$$\langle l, S_{t_*}h\rangle = \int_0^{t_*-\varepsilon} \int_{\partial\Omega} w_\varepsilon^\partial(y, s)h(y, s)\, dS_y\, ds,$$

where

$$w_\varepsilon(x, s) = \sum_{k=1}^\infty e^{-\lambda_k(t_* - \varepsilon/2 - s)}(v_\varepsilon, \varphi_k)\varphi_k(x)$$

is a nonvanishing solution of the adjoint heat equation on $\Omega \times (0, t_* - \varepsilon)$ and the boundary function $w_\varepsilon^\partial(x, s)$ is obtained from $w_\varepsilon(x, s)$ as in the statement of Theorem 3. Easy estimates also guarantee that $w_\varepsilon^\partial(y, s) \in L_1(\partial\Omega \times (0, t_* - \varepsilon))$. Using the arbitrariness of h it follows that, if $\varepsilon_1 < \varepsilon_2$, $w_{\varepsilon_1}(y, s) = w_{\varepsilon_2}(y, s)$ for $s < t_* - \varepsilon_2$. Thus finally, we obtain a nonzero solution of (16) such that, for any $h \in L_\infty$ with support in $\partial\Omega \times (0, t_* - \varepsilon)$ for some $\varepsilon > 0$,

$$\langle l, S_{t_*}h\rangle = \int_0^{t_*} \int_{\partial\Omega} w^\partial(y, s)h(y, s)\, dS_y\, ds.$$

Hence (18) yields

(20) $$\int_0^{t_*} \int_{\partial\Omega} w^\partial(y, s)\chi_\varepsilon(y, s)[f_*(y, s) - f(y, s)]\, dS_y\, ds \geqq 0,$$

for each $f \in L_M$ and for each $\varepsilon > 0$. It was proved by Fattorini [3] in the case that $\partial\Omega$ is analytic and by Schmidt and Weck [11] in the case that $\partial\Omega$ is C^∞, that when w is a nontrivial solution of (16) the boundary function $w^\partial(y, s)$ cannot vanish on a set of positive measure. Thus (17) is an immediate consequence of (20). This completes the proof of Theorem 3.

Remarks. 1. The condition $M > \text{dist}_R\ (u_1, X)$ is difficult to verify in general. It is however, automatically satisfied if $M > \inf\{M \in (0, \infty) : u_1 \in R_t(u_0, L_M) \text{ for some } t > 0\}$.

2. The time optimal problem can also be posed for controls restricted to $L_{m,M} = \{f \in L_\infty : m \leqq f(x, t) \leqq M \text{ a.e.}\}$. In this case one can use identity (7) to replace u_1, u_0 and f_* by $u_1-(M+m)/2$, $u_0-(M+m)/2$, $f_*-(M+m)/2$ respectively, thus concluding that $f_*(x, t) = (M+m)/2 + \text{sgn}\,[w^\partial(x, t)](M-m)/2$.

3. Henry [7] considered the case of distributed controls (in which the control is by means of an inhomogeneity in the equation rather than in the boundary condition); the argument used in this paper carries over with only slight modifications.

REFERENCES

[1] S. Agmon, *Lectures on Elliptic Boundary Value Problems*, Van Nostrand, New York, 1965.

[2] H. O. Fattorini, *Control in finite time of differential equations in Banach spaces*, Comm. Pure Appl. Math., 19 (1966), pp. 17–34.

[3] ———, *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.

[4] H. O. Fattorini and D. L. Russell, *Exact controllability theorems for linear parabolic equations in one space dimension*, Arch. Rational Mech. Anal., 4 (1971), pp. 272–292.

[5] ———, *Uniform bounds on biorthogonal functions for real exponentials with an application to the control theory of parabolic equations*, Quart. of Appl. Math., 32 (1971), pp. 45–69.

[6] K. Glashoff and N. Weck, *Boundary control of parabolic differential equations in arbitrary dimensions: supremum norm problems*, this Journal, 14 (1976), pp. 662–681.

[7] J. Henry, *Contrôle en temps optimal pour les systèmes gouvernés par une équation de type parabolique*, preprint.

[8] 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.

[9] D. L. Russell, *A unified boundary controllability theory for hyperbolic and parabolic partial differential equations*, Studies in Appl. Math., 52 (1973), pp. 189–211.

[10] G. Schmidt, *Boundary control for the heat equations with stationary targets*, this Journal, submitted.

[11] G. Schmidt and N. Weck, *On the boundary behaviour of solutions to elliptic and parabolic equations—with applications to boundary control for parabolic equations*, this Journal, 16 (1978), pp. 593–598.

[12] T. I. Seidman, *A well-posed problem for the heat equation*, Bull. Amer. Math. Soc., 80 (1974), pp. 901–902.

[13] ———, *Time-invariance of the reachable set for linear control problems*, J. Math. Anal. Appl., to appear.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 2, March 1980

0363-0129/80/1802-0002 $01.00/0

FEEDBACK STABILIZATION OF DISTRIBUTED PARAMETER SYSTEMS BY A FUNCTIONAL OBSERVER*

NOBUO FUJII†

Abstract. Feedback stabilization of unstable parabolic equations is of great interest. The fact that it is not necessarily possible to stabilize the equations by means of static feedback schemes when both observation and control can be realized only through the boundary is illustratively shown by a simple example. In view of this, a functional observer of Luenberger type is derived and then utilized in order to stabilize unstable parabolic equations for which observation of the state and control can be carried out only through the boundary.

1. Introduction. The investigation of feedback stabilization of distributed parameter systems has received attention in these years. For parabolic equations there are investigations by Y. Sakawa and T. Matsushita [1], [2] and T. Nambu [3]. For hyperbolic equations Y. Sakawa [4] and M. Slemrod [5] considered feedback stabilization using the invariance principle of J. K. Hale [6] and J. P. LaSalle [7].

As for parabolic equations, stabilization of the systems by means of interior output-interior input scheme are treated in [1], of interior output-boundary input scheme in [2]; T. Nambu, instead, considered stabilization by boundary output-interior input. Apparently, stabilization by boundary output-boundary input can be treated in the same manner as that in [1] if the eigenfunctions of the eigenvalue problem associated with the parabolic equation form an orthogonal system in the space of functions square integrable over the boundary. But unfortunately this is in general not the case.

To clarify the situation, now examine a simple example.

Example 1. Consider one dimensional heat equation:

$$\frac{\partial u}{\partial t}=\frac{\partial^2 u}{\partial x^2}+au, \qquad 0<x<1, \quad t>0, \tag{1.1}$$

$$\begin{aligned}&\frac{\partial u}{\partial n}=g(x)f(t), \qquad x\in\{0,1\}, \quad t>0,\\ &g(0)=-1, \qquad g(1)=0,\end{aligned} \tag{1.2}$$

$$u(x,0)=u_0(x), \qquad 0<x<1, \tag{1.3}$$

where a is some constant and $\partial/\partial n$ denotes outer normal differentiation on the boundary. Let the boundary observation law be

$$h(t)=(u(x,t),w(x))_S=\eta_1 u(0,t)+\eta_2 u(1,t), \qquad (\eta_1,\eta_2\text{: real}).$$

A closed loop system is composed by setting

$$f(t)=h(t)=\eta_1 u(0,t)+\eta_2 u(1,t). \tag{1.4}$$

* Received by the editors November 3, 1978, and in revised form February 28, 1979.

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

Eigenvalues and the corresponding eigenfunctions of the eigenvalue problem

$$\lambda\phi = \frac{\partial^2\phi}{\partial x^2} + a\phi, \qquad 0<x<1, \tag{1.5}$$
$$\frac{\partial\phi}{\partial n} = 0, \qquad x \in \{0, 1\}$$

are given by

$$\lambda_m = a - (m-1)^2\pi^2, \tag{1.6}$$
$$\phi_m = \cos(m-1)\pi x, \qquad m = 1, 2, \cdots.$$

The multiplicity of each eigenvalue is clearly equal to one. Open loop system (1.1)–(1.3) is unstable provided $a > 0$.

It is obvious that the eigenfunctions do not form an orthogonal system over S; clearly, for any nontrivial $w(x)$ (i.e., for nontrivial pair of η_1, η_2), there cannot exist an integer N such that

$$(w(x), \phi_m(x))_S = \eta_1 + (-1)^{m-1}\eta_2 = 0, \qquad m \geqq N. \tag{1.7}$$

Since the stabilization method of [1] tacitly requires that (1.7) hold for some integer N, it does not work for the system. In addition to this fact, it is possible to prove: *if* $a > 4\pi^2$, *the closed loop system defined by* (1.1)–(1.4) *is unstable whatever* η_1 *and* η_2 *may be.* The proof is obtained with an easy but tedious examination of the roots of a transcendental equation with the help of diagrams; hence, it is omitted.

Thus alternative methods are required to be developed in order to stabilize parabolic equations for which only boundary observations and boundary controls are available.

In this paper, we shall present the feedback stabilization scheme with a functional observer of Luenberger type for parabolic systems. In § 2, the functional observer will be constructed. In § 3, the feedback stabilization problem will be solved

Throughout this paper, Δ denotes, as usual, Laplacian operator in Euclidean n-space and $\partial/\partial n$ stands for outer normal differentiation on the boundary. Also we shall often designate different constants by the same letter K if we are not interested in their magnitudes. If there is no confusion column or row vector $(g_1, \cdots, g_k)$ will often be abbreviated as g without any suffix etc. Similarly, the scalar product of two vectors, say g and f, will simply be denoted by gf if its meaning is obvious from the context.

2. Functional observer. In actual systems, it is often the case that information of the system can be obtained only through the boundary. Hence, it is required to estimate system's behavior, based on the information, by means of an appropriate machine.

In this section, we shall construct a functional observer for parabolic equations making use of the information obtained through sensors on the boundary.

Let D be a bounded domain in n-dimensional Euclidean space and S be its sufficiently smooth boundary. Consider a parabolic initial boundary value problem:

$$\frac{\partial u}{\partial t} = \Delta u + q(x)u, \qquad x \in D, \quad t>0, \tag{2.1}$$

$$\frac{\partial u}{\partial n} + \sigma(x)u = \sum_{i=1}^{m} g_i(x)f_i(t) \quad (\text{abbreviated as} = g(x)f(t)), \quad x \in S, \quad t>0, \tag{2.2}$$

$$u(x,0) = u_0(x), \qquad x \in D. \tag{2.3}$$

Here u is called a state variable and $f_i(t)$ are control inputs; thus, inputs are exerted on the system through the boundary. Suppose that the real valued $q(x)$ is Hölder continuous with exponent α in $\bar{D}$ $(=D \cup S)$, real valued $\sigma(x)$, $g(x)$ are continuous on S and $f(t)$ is continuous for $t \geqq 0$. Assume, furthermore, that the function $u_0(x)$ is defined and continuously differentiable in an open set which contains $\bar{D}$.

Let observation laws be defined by

$$h_k(t) = (u(x, t), w_k(x))_S, \qquad k = 1, 2, \cdots, l, \tag{2.4}$$

where $w_k(x)$ are continuous on S and $(\cdot, \cdot)_S$ denotes, as usual, the inner product in $L_2(S)$ the space of functions square integrable over S. Using observations $h_k(t)$, let us construct a functional observer whose outputs asymptotically approach to the values of functionals defined by

$$y_k(t) = (u(x, t), \rho_k(x)), \qquad k = 1, \cdots, r, \tag{2.5}$$

where $\rho_k(x)$ belong to $L_2(D)$ and $(\cdot, \cdot)$ denotes the inner product in $L_2(D)$. In view of the linearity of (2.1)–(2.3), we can decompose a solution of them as

$$u(x, t) = u_1(x, t) + u_2(x, t). \tag{2.6}$$

Here $u_1(x, t)$ stands for the solution for $f(t) \equiv 0$ and $u_2(x, t)$ for $u_0(x) \equiv 0$. As is well known, u_1 can be expressed by

$$u_1(x, t) = \sum_{i=1}^{\infty} \sum_{j=1}^{m_i} a_{ij} e^{\lambda_i t} \phi_{ij}. \tag{2.7}$$

Here λ_i are eigenvalues of the eigenvalue problem:

$$\begin{aligned} &\lambda \phi = \Delta \phi + q\phi, && x \in D, \\ &\frac{\partial \phi}{\partial n} + \sigma(x)\phi = 0, && x \in S, \end{aligned} \tag{2.8}$$

ϕ_{ij} are corresponding eigenfunctions, m_i, the multiplicity of λ_i and a_{ij} are defined by

$$a_{ij} = (u_0, \phi_{ij}), \qquad i = 1, 2, \cdots, \quad j = 1, \cdots, m_i. \tag{2.9}$$

Hence, outputs $h_k(t)$ can be written as

$$\begin{aligned} h_k(t) &= (u_1(x, t), w_k(x))_S + (u_2(x, t), w_k(x))_S \\ &= \sum_{i=1}^{M} \sum_{j=1}^{m_i} a_{ij} e^{\lambda_i t} (\phi_{ij}, w_k)_S + \sum_{i=M+1}^{\infty} \sum_{j}^{m_i} a_{ij} e^{\lambda_i t} (\phi_{ij}, w_k)_S \\ &\qquad + (u_2(x, t), w_k(x))_S, \quad k = 1, \cdots l. \end{aligned} \tag{2.10}$$

Let us consider the second term of the right hand side of (2.10) which will be denoted by $d_k(t)$. Using Schwarz inequality we can easily obtain

$$\begin{aligned} |d_k(t)| \leqq e^{\lambda_{M+1} t} &\left(\sum_{i=M+1}^{\infty} \sum_{j} |\lambda_i|^{2\nu} |a_{ij} e^{(\lambda_i - \lambda_{M+1})t}|^2 \right)^{1/2} \\ &\cdot \left(\sum_{i=M+1}^{\infty} \sum_{j} \frac{|(\phi_{ij}, w_k)_S|^2}{|\lambda_i|^{2\nu}} \right)^{1/2}, \quad k = 1, \cdots, l, \end{aligned}$$

where $\nu (>0)$ is some constaint. If $\nu > 1/2$ and $t \geqq t_0 > 0$, the series expansion of the right hand side converges and bounded, where t_0 is some fixed time. Thus, it follows that

$$|d_k(t)| \leqq K e^{\lambda_{M+1} t}, \quad \text{i.e.,} \quad \|d(t)\| \leqq K e^{\lambda_{M+1} t}. \tag{2.11}$$

As for the third term of the right hand side of (2.10), we can obtain the following:

LEMMA. *On the above assumptions, there exists functions $T_k(t)$, which are continuous in $t>0$, such that*

$$(u_2(x, t), w_k(x))_S = \int_0^t T_k(t-s)f(s)\,ds, \qquad k=1, \cdots, l, \tag{2.12}$$

hold.

In view of this lemma, $h_k(t)$ can be represented as

$$h_k(t) = \sum_{i=1}^{M} \sum_{j=1}^{m} a_{ij}\, e^{\lambda_i t}(\phi_{ij}, w_k)_S + d_k(t) + \int_0^t T_k(t-s)f(s)\,ds. \tag{2.13}$$

Now let us introduce N-dimensional vector X by

$$X = \mathrm{col}\,(X_{11}, \cdots, X_{1m_1}, \cdots, X_{Mm_M}),$$

$$X_{ij} = (u(x, t), \phi_{ij}), \qquad i=1, \cdots, M, \quad j=1, \cdots, m_i,$$

where $N=\sum_{i=1}^{M} m_i$. From (2.1)–(2.3) and with the help of Green's formula it follows that

$$\begin{aligned} \frac{dX_{ij}}{dt} &= \frac{d}{dt}(u(x, t), \phi_{ij}) = (\Delta u + qu, \phi_{ij}) \\ &= \lambda_i(u, \phi_{ij}) + \sum_k (g_k, \phi_{ij})_S f_k(t), \\ X_{ij}(0) &= a_{ij}, \qquad i=1, \cdots, M, \quad j=1, \cdots, m_i, \end{aligned} \tag{2.14}$$

or

$$\frac{dX}{dt} = AX + Gf(t), \qquad X(0) = \mathrm{col}\,(a_{11}, \cdots, a_{Mm_M}), \tag{2.15}$$

where $N\times N$matrix A and $N\times m$ matrix G are defined by

$$A = \mathrm{diag}\,(\lambda_1, \cdots, \lambda_1, \cdots, \underbrace{\lambda_i, \cdots, \lambda_i}_{m_i}, \cdots, \lambda_M),$$

$$G = (G_1, \cdots, G_k, \cdots, G_m)_{N\times m}$$

$$G_k = \mathrm{col}\,((g_k, \phi_{ij})_S, \cdots, (g_k, \phi_{Mm_M})_S), \qquad k=1, \cdots, m.$$

From (2.14) we can obtain

$$a_{ij}\, e^{\lambda_i t} = X_{ij}(t) - \int_0^t e^{\lambda_i(t-s)} \sum_k (g_k, \phi_{ij})_S f_k(s)\,ds. \tag{2.16}$$

Substituting (2.16) into (2.13) we can obtain

$$\begin{aligned} h_k(t) = \sum_{i=1}^{M} \sum_{j=1}^{m_i} &\left[(\phi_{ij}, w_k)_S X_{ij} - (\phi_{ij}, w_k)_S \int_0^t e^{\lambda_i(t-s)} \sum_h (g_h, \phi_{ij})_S f_h(s)\,ds \right] \\ &+ d_k(t) + \int_0^t T_k(t-s)f(s)\,ds. \end{aligned}$$

If we define $l \times N$ matrix W by

$$W = \begin{pmatrix} W_1 \\ \vdots \\ W_l \end{pmatrix}, \qquad W_k = \text{row}\,((\phi_{11}, w_k)_S, \cdots, (\phi_{Mm_M}, w_k)_S), \qquad k = 1, \cdots, l,$$

and $l \times m$ matrix $H(t)$ by

$$H(t) = \begin{bmatrix} H_{11} & \cdots & H_{1m} \\ & \vdots & \\ H_{l1} & \cdots & H_{lm} \end{bmatrix},$$

$$H_{kh}(t) = T_{kh}(t) - \sum_{i=1}^{M} \sum_{j=1}^{m_i} (\phi_{ij}, w_k)_S\, e^{\lambda_i t}(g_h, \phi_{ij})_S,$$

Then, we obtain

$$\begin{aligned} h(t) &= WX + d(t) + \int_0^t H(t-s)f(s)\,ds \\ &= WX + d(t) + H * f. \end{aligned} \tag{2.17}$$

Here, of course, $d(t) = \text{col}\,(d_1(t), \cdots, d_l(t))$ and $\cdot * \cdot$ denotes the convolution.

Now we are in place to construct a functional observer according to D. G. Luenberger [10]. Assume that $\rho_h(x)$ in (2.5) are expressed by

$$\rho_h(x) = \sum_{i=1}^{p} \sum_{j=1}^{m_i} (\rho_h, \phi_{ij})\phi_{ij}, \qquad h = 1, \cdots, r, \tag{2.18}$$

with some integer $p(>0)$. For any given $\mu > 0$ choose M such that $M \geqq p$ and $\lambda_{M+1} < -\mu$ hold. Define $l \times m_i$ matrices $\tilde{W}_i$ by

$$\tilde{W}_i = \begin{bmatrix} (\phi_{i1}, w_1)_S & \cdots & (\phi_{im_i}, w_1)_S \\ & \vdots & \\ (\phi_{i1}, w_l)_S & \cdots & (\phi_{im_i}, w_l)_S \end{bmatrix}, \qquad i = 1, \cdots, M.$$

Furthermore, consider an N-dimensional lumped parameter system defined by

$$\begin{aligned} &\frac{dz}{dt} = Fz(t) + Bh(t) + Cf(t) + DH * f(t),\ z(0) = z_0, \\ &\tilde{y}(t) = Pz, \end{aligned} \tag{2.19}$$

where F, B, C, D and P are constant matrices of appropriate sizes. Our aim is to choose F, B, C, D and P in order that output $\tilde{y}(t)$ asymptotically approaches $y(t)$ given by (2.5). In this connection, we can easily obtain the following proposition.

PROPOSITION 1. *Assume that the conditions* $l \geqq \max_{1 \leqq i \leqq M} m_i$ and

$$\text{rank}\ \tilde{W}_i = m_i, \qquad i = 1, \cdots, M \tag{2.20}$$

hold. Then, we can find matrices F, B, C, D and P such that the output $\tilde{y}(t)$ *of* (2.19) *satisfies*

$$\|\tilde{y}(t) - y(t)\| \leqq K e^{-\mu t} \tag{2.21}$$

with some constant K, which may depend on $X(0)$ *and* z_0, *and the system* (2.19) *is asymptotically stable.*

Proof. From (2.15), (2.17) and (2.19), it follows readily

$$\frac{d}{dt}(X-z)=(A-BW)X-Fz+(G-C)f(t)-(B+D)H*f-Bd(t). \tag{2.22}$$

In view of (2.20), (W, A) is an observable pair [1]; hence, we can design matrix B such that matrix $A-BW$ has eigenvalues whose real parts are smaller than $-\mu$. Fix such B. Define F, C and D by

$$F=A-BW, \qquad C=G, \qquad D=-B, \tag{2.23}$$

then we can reduce (2.22) to

$$\frac{d}{dt}(X-z)=F(X-z)+Bd(t).$$

From (2.11) and the choice of F, it follows readily

$$\|X-z\|\leqq Ke^{-\mu t}, \tag{2.24}$$

where constant K may depend on $X(0)$ and z_0. Define $r\times N$ matrix P by

$$P=\begin{bmatrix}(\rho_1, \phi_{11}) & \cdots & (\rho_1, \phi_{Mm_M})\\ & \vdots & \\ (\rho_r, \phi_{11}) & \cdots & (\rho_r, \phi_{Mm_M})\end{bmatrix};$$

then, in view of (2.18), $y(t)$ can be rewritten as

$$y(t)=PX.$$

From this and (2.24) it follows readily that (2.21) holds. This completes the proof.

Remark 1. The functional observer constructed contains the convolutional operation; hence, strictly speaking, it is not a finite dimensional system. The author believes that it is impossible to construct a purely finite dimensional observer if the inputs to the system are present and are exerted through the boundary. Our observer is, however, not so meaningless in that the convolution term can be calculated with the aid of an analog or a digital computer. Thus, in practice, the observation of functionals will result in discrete time manner.

Note that the observer reduces to a very finite dimensional system if the inputs are absent.

Now the proof of the lemma is left.

Proof of Lemma. As was shown in [11], for arbitrary $T(>0)$, $u_2(x, t)$ in (2.6) can be represented as

$$u_2=\int_0^t\int_S U(x, y; t-s)\psi(y, s)\,dS_y\,ds, \qquad 0<t<T, \tag{2.25}$$

where $U(x, y; t-s)$ is a fundamental solution of the parabolic operator and

$$\psi(x, t)=-2g(x)f(t)-2\sum_{k=1}^{\infty}\int_0^t M_k(x, y; t-s)g(y)f(s)\,dS_y\,ds, \tag{2.26}$$

$$\begin{aligned} M_1(x, y; t-s)&=2\frac{\partial}{\partial n}U(x, y; t-s),\\ M_{k+1}(x, y; t-s)&=\int_s^t\int_S M_1(x, \xi; t-\sigma)M_k(\xi, y; \sigma-s)\,dS_\xi\,d\sigma. \end{aligned} \tag{2.27}$$

For the moment fix T. Let ν be a constant which satisfies $\frac{1}{2}<\nu<1$ and $1-\alpha/2<\nu<1$. Then we can obtain the following estimate [11]:

$$|U(x, y; t-s)| \leqq K\frac{1}{(t-s)^{\nu}}\frac{1}{|x-y|^{n-2\nu}}, \qquad 0\leqq s<t\leqq T, \quad x, y \in \bar{D},$$

$$|M_{k+1}(x, y; t-s)| \leqq K\frac{1}{(t-s)^{\nu+k(\nu-1)}}\frac{1}{|x-y|^{n+1-2\nu-\alpha+k(2-2\nu-\alpha)}},$$

$$k=0, 1, \cdots, k_0-2,$$

$$|M_{k_0}(x, y; t-s)| \leqq K,$$

$$|M_{k_0+k}(x, y; t-s)| \leqq K_0\frac{|K_1(t-s)^{1-\nu}|^k}{\Gamma((1-\nu)k+1)}, \qquad k=1, 2, \cdots,$$

$$0\leqq s<t\leqq T, \qquad x, y \in S,$$

where K, K_0 and K_1 are some constants and k_0 is the smallest integer which satisfies

$$n+1-2\nu-\alpha+k_0(2-2\nu-\alpha)\leqq 0, \qquad \nu+k_0(\nu-1)\leqq 0.$$

Using these estimates we can easily obtain the following:

$$(2.28) \quad \int_0^t \int_S |M_k(x, y; t-s)g(y)f(t)|\, dS_y\, ds \leqq \begin{cases} KT^{1-\nu-k(\nu-1)}<\infty, & k<k_0-1, \\ KT<\infty, & k=k_0, \end{cases}$$

$$(2.29) \quad \int_0^t \int_S |M_{k_0+k}(x, y; t-s)g(y)f(t)|\, dS_y\, ds \leqq KK_0\frac{1}{K_1}\frac{(K_1T)^{(1-\nu)k+1}}{\Gamma((1-\nu)k+2)}<+\infty,$$

$$k=1, 2, \cdots.$$

Here we used [11, Lemma 1, § 2, Chap. 5]. Thus the series expansion in (2.26) is absolutely convergent. Furthermore, in view of [11, Lemma 1, § 3, Chap. 1], it follows that the each term in the expansion is continuous with respect to $x\in S$ and $t\in[0, T]$; hence, $\psi(x, t)$ is continuous on $S\times[0, T]$.

Next let us consider functionals $(u_2(x, t), w_k(x))_S$ which can be expressed by

$$(2.30) \qquad (u_2(x, t), w_k(x))_S = \int_S \bar{w}_k(x)\, dS_x \int_0^t \int_S U(x, y; t-s)\psi(y, s)\, dS_y\, ds,$$

where $\bar{w}$ denotes the complex conjugate of w. Estimate (2.28) and expression (2.26) enable us to rewrite (2.30) into the form:

$$(2.31) \qquad \begin{aligned} (u_2(x, t), w_k(x))_S &= -2\int_0^t ds \int_S \int_S \bar{w}_k(x)U(x, y; t-s)g(y)f(s)\, dS_y\, dS_x \\ &\quad -2\int_0^t ds \int_S \int_S \mathrm{d}S_y\, dS_x \bar{w}_k(x)U(x, y; t-s) \\ &\quad \cdot\left(\sum_{j=1}^{\infty}\int_0^s \int_S M_j(y, \xi; s-\tau)g(\xi)f(\tau)\, dS_\xi\, d\tau\right). \end{aligned}$$

The first term on the right hand side can be rewritten as

$$\int_0^t T_{k,1}(t-s)f(s)\, ds,$$

where

$$T_{k,1}(t-s)=-2\int_S\int_S \bar{w}_k(x)U(x,y;t-s)g(y)\,dS_y\,dS_x, \qquad k=1,\cdots,l, \tag{2.32}$$

are continuous with respect to t, s, $0\leqq s<t\leqq T$. As for second term, we can obtain the following estimates for $k=1,2,\cdots,l$:

$$\begin{aligned}&\int_0^t ds\int_S dS_x\int_S dS_y\int_0^s d\tau\int_S dS_\xi|\bar{w}_k(x)U(x,y;t-s)M_j(y,\xi;s-\tau)g(\xi)f(\tau)|\\ &\leqq\begin{cases}KT, & j\leqq k_0,\\ KK_0K_1^{h+1}\dfrac{1}{\Gamma((1-\nu)(h+1)+1)}\dfrac{T^{(1-\nu)(h+1)+1}}{(1-\nu)(h+1)+1}, & j=k_0+h,\ h=1,2,\cdots.\end{cases}\end{aligned}\tag{2.33}$$

Here we used estimates (2.28), (2.29) and [11, Lemma 1, § 2, Chap. 5]. The series expansion in (2.31), hence, is absolutely convergent and the second term of the right hand side of (2.31) can be expressed by

$$\int_0^t T_{k,2}(t-s)f(s)\,ds,$$

with

$$T_{k,2}(t-s)=\sum_{j=1}^{\infty}\int_0^t d\tau\int_S\int_S\int_S dS_x\,dS_y\,dS_\xi\,\bar{w}_k(x)U(x,y;t-\tau)M_j(y,\xi;\tau-s)g(\xi), \tag{2.34}$$
$$k=1,2,\cdots,l.$$

Again, $T_{k,2}(t-s)$ are concluded to be continuous with respect to t, s, $0\leqq s<t\leqq T$ with the help of the same argument as that for $\psi(x,t)$.

Since T is arbitrary in the above discussion we can, finally, obtain for $t\geqq 0$

$$(u_2(x,t),w_k(x))_S=\int_0^t T_k(t-s)f(s)\,ds, \qquad k=1,2,\cdots,l, \tag{2.35}$$

where

$$T_k(t)=T_{k,1}(t)+T_{k,2}(t) \tag{2.36}$$

are continuous with respect to $t>0$. The proof is thereby completed.

3. Stabilization. In this section we shall consider a stabilization problem of parabolic equations for which only boundary input and boundary output can be utilized. To solve the problem, use is made of the functional observer derived in the previous section. Finally, we shall apply the theory to the system considered in Example 1.

Consider again the parabolic initial boundary value problem described by (2.1)–(2.3) and the observer given by (2.19), where input $h(t)$ of it is given by (2.4). In order to construct a closed loop system, let the boundary input $f(t)$, which is the same as that in (2.19), be the output $\tilde{y}(t)$ of the observer, i.e.,

$$f(t)=\tilde{y}(t). \tag{3.1}$$

Thus we obtain a closed loop system described by

$$\begin{aligned}&\frac{\partial u}{\partial t}=\Delta u+q(x)u, \qquad x\in D,\quad t>0,\\ &u(x,0)=u_0(x), \qquad x\in D,\end{aligned}\tag{3.2}$$

$$\frac{\partial u}{\partial n}+\sigma(x)u=g(x)\tilde{y}(t), \qquad x\in S, \quad t>0, \tag{3.3}$$

$$h_k(t)=(u(x,t), w_k(x))_S, \qquad k=1,\cdots,l, \tag{3.4}$$

$$\frac{dz}{dt}=Fz(t)+Bh(t)+C\tilde{y}(t)+DH^*\tilde{y}(t), \qquad z(0)=z_0, \tag{3.5}$$

$$\tilde{y}(t)=Pz. \tag{3.6}$$

Let us seek a set of sufficient conditions which ensures the exponentially asymptotic stability of the parabolic equation and, in this context, determine matrices F, B, C, D and P.

Let μ be an arbitrary positive constant and, in what follows, fix it. Define matrices $\tilde{G}_i$ by

$$\tilde{G}_i=\begin{pmatrix} (g_1,\phi_{i1})_S & \cdots & (g_m,\phi_{i1})_S \\ \vdots & & \\ (g_1,\phi_{im_i})_S & \cdots & (g_m,\phi_{im_i})_S \end{pmatrix}, \qquad i=1,2,\cdots.$$

Now, for controllers $g_i(x)$ and observers $w_k(x)$, let the following assumptions hold.

Assumption 1. The conditions

$$\operatorname{rank}\tilde{G}_i=m_i, \qquad i=1,\cdots,M, \tag{3.7}$$

hold where M is an integer such that $\lambda_{M+1}<-\mu$.

Assumption 2. The conditions

$$\operatorname{rank}\tilde{W}_i=m_i, \qquad i=1,\cdots,M, \tag{3.8}$$

hold where, of course, $\tilde{W}_i$ are defined in the previous section.

Based on Assumption 1, it is possible to choose functions $\rho_1,\cdots,\rho_m\in L_2(D)$ such that all the eigenvalue of the eigenvalue problem

$$\lambda u=\Delta u+q(x)u, \qquad x\in D,$$

$$\frac{\partial u}{\partial n}+\sigma(x)u=\sum_{i=1}^{m} g_i(x)(u(x),\rho_i(x)), \qquad x\in S,$$

satisfy Re $\lambda<-\mu$[1], [4]. Further $\rho_i(x)$ can be chosen in the form

$$\rho_i(x)=\sum_{k=1}^{M}\sum_{j=1}^{m_k}(\rho_i,\phi_{kj})\phi_{kj}, \qquad i=1,\cdots,m. \tag{3.9}$$

For such a set of functions, define $m\times n$ matrix P by

$$P=\begin{bmatrix} (\rho_1,\phi_{11}) & \cdots & (\rho_1,\phi_{Mm_M}) \\ \vdots & & \vdots \\ (\rho_m,\phi_{11}) & \cdots & (\rho_m,\phi_{Mm_M}) \end{bmatrix}. \tag{3.10}$$

On the other hand, in view of Assumption 2 and the discussion in the previous section, we can find matrix B such that the real parts of the eigenvalues s_i of matrix F, which is defined by

$$F=A-BW, \tag{3.11}$$

satisfy

$$\operatorname{Re} s_i<-\mu, \qquad i=1,\cdots,N. \tag{3.12}$$

Let matrices C and D be such that

$$C=G, \qquad D=-B; \tag{3.13}$$

thus, matrices F, B, C, D and P are all determined. From the choice of the matrices, it is obvious that equation (3.5) is exponentially stable and the system defined by (3.5) and (3.6) is a functional observer for functionals $y_i(t)$ given by

$$y_i(t)=(u(x,t),\rho_i(x)), \qquad i=1,\cdots,m. \tag{3.14}$$

Let $\delta_i(t)$ denote the difference of $y_i(t)$ and $\tilde{y}_i(t)$, i.e.,

$$\tilde{y}_i(t)=y_i(t)+\delta_i(t), \qquad i=1,\cdots,m. \tag{3.15}$$

Then it follows, in view of the proof of Proposition 1, that

$$|\delta_i(t)|\leqq Ke^{-\mu t}, \qquad \left|\frac{d}{dt}\delta_i(t)\right|\leqq K e^{-\mu t}, \qquad i=1,\cdots,m, \tag{3.16}$$

for $t>0$. From the above consideration, finally, it follows that the closed loop system described by (3.2)–(3.6) is equivalent to the parabolic initial boundary value problem

$$\frac{\partial u}{\partial t}=\Delta u+q(x)u, \qquad x\in D, \quad t>0, \tag{3.17}$$

$$\frac{\partial u}{\partial n}+\sigma(x)=\sum_{i=1}^{m} g_i(x)(u(x,t),\rho_i(x))+\sum_{i=1}^{m} g_i(x)\delta_i(t), \qquad x\in S, \quad t>0, \tag{3.18}$$

$$u(x,0)=u_0(x), \qquad x\in D, \tag{3.19}$$

and the system (3.5), (3.6).

Now we can easily establish the following proposition.

PROPOSITION 2. *Suppose that $g_i(x)(i=1,\cdots,m)$ and $\sigma(x)$ are twice continuously differentiable on S. Then, with Assumptions* 1 *and* 2, *the solution of the initial boundary value problem* (3.17)–(3.19) *satisfies*

$$\|u(\cdot,t)\|\leqq K e^{-\mu t} \tag{3.20}$$

for some constant K.

Proof. The proof is straightforward. Let $\tilde{g}_k(x)(k=1,\cdots,m)$ be arbitrary twice continuously differentiable functions on S. Introduce m functions ψ_k by the relation

$$\psi_k=\sum_{i=1}^{m} g_i(\tilde{g}_k,\rho_i)_S+g_k-\tilde{g}_k, \tag{3.21}$$

which, obviously, are twice continuously differentiable on S. Then, as is well known, there exist m functions $\phi_k(x)$, which are defined and twice continuously differentiable on $\bar{D}$, such that

$$\frac{\partial\phi_k}{\partial n}=\psi_k, \qquad \phi_k=\tilde{g}_k, \qquad x\in S, \quad k=1,\cdots,m. \tag{3.22}$$

Consider function $\Phi(x,t)$ defined by

$$\Phi(x,t)=\sum_{k=1}^{m}\phi_k(x)\delta_k(t).$$

In view of this and (3.22), it follows readily

$$\frac{\partial\Phi}{\partial n}+\sigma(x)\Phi=\sum g_i(\Phi(x,t),\rho_i(x))_S+\sum g_i(x)\delta_i(t),\qquad x\in S. \tag{3.23}$$

Let us seek the solution of (3.17)–(3.19) in the form

$$u(x,t)=v(x,t)+\Phi(x,t). \tag{3.24}$$

Using (3.23), we can reduce the initial boundary value problem to

$$\frac{\partial v}{\partial t}=\Delta v+q(x)v-\frac{\partial\Phi}{\partial t}+\Delta\Phi+q\Phi,\qquad x\in D,\quad t>0, \tag{3.25}$$

$$\frac{\partial v}{\partial n}+\sigma(x)v=\sum g_i(x)(v,\rho_i)_S,\qquad x\in S,\quad t>0, \tag{3.26}$$

$$v(x,0)=u_0(x)-\Phi(x,0),\qquad x\in D. \tag{3.27}$$

The forcing term on the right-hand side of (3.25) clearly satisfies

$$\left\|-\frac{\partial\Phi}{\partial t}+\Delta\Phi+q\Phi\right\|\leqq Ke^{-\mu t}$$

because of the definition of Φ and (3.16). From these and the choice of ρ_i the assertion of the proposition follows. This completes the proof.

Let us apply the theory developed to the system given in Example 1.

Example 2. Consider again the control system (1.1)–(1.3) with $a=5\pi^2$. Thus unstable eigenvalues $5\pi^2$, $4\pi^2$ and π^2 appear. Let observation $h(t)$ be given by

$$h(t)=(u(x,t),w(x))_S=u(0,t).$$

For the sake of simplicity, let μ in Proposition 2 be $3\pi^2$; hence, number M above can be taken to be three ($\lambda_4=-4\pi^2<-\mu$). From the choice of the control and observation laws and the fact that $m_i=1(i=1,2,\cdots)$, it follows that $\tilde{G}_i$ and $\tilde{W}_i$ above are scalars and given by

$$\tilde{G}_i=-1,\qquad \tilde{W}_i=1,\qquad i=1,2,\cdots.$$

Thus Assumptions 1 and 2 are automatically satisfied; moreover, the system is controllable and observable [8], [9]. Note also that

$$A=\operatorname{diag}(5\pi^2,4\pi^2,\pi^2),\qquad W=\operatorname{row}(1,1,1),$$
$$G=\operatorname{col}(-1,-1,-1).$$

The functional observer can easily be constructed as follows:

(i) Determine matrix B such that the eigenvalues of $A-BW$ are $\{-4\pi^2,-5\pi^2,-6\pi^2\}$. This is accomplished by setting

$$B=\operatorname{col}\left(\frac{495}{2}\pi^2,-240\pi^2,\frac{35}{2}\pi^2\right).$$

(ii) According to the procedure given in [1], we can design an interior sensor $\rho(x)$ such that the eigenvalues λ_i of closed loop system (1.1), (1.2) together with $f(t)=(u,\rho)$ satisfy $\operatorname{Re}\lambda_i\leqq-4\pi^2$ $(i=1,2,\cdots)$. As such a $\rho(x)$, we can take, for example,

$$\rho(x)=\frac{495}{2}\pi^2-240\pi^2\cos(\pi x)+\frac{35}{2}\pi^2\cos(2\pi x).$$

(iii) The solution $u_2(t, x)$ for (1.1)–(1.3) with $u_0(x) = 0$ is expressed by

$$u_2(x, t) = -\int_0^t \frac{\exp(5\pi^2(t-s))}{\sqrt{\pi(t-s)}}\left(\sum_{n=-\infty}^{+\infty} \exp\left(-\frac{(x+2n)^2}{4(t-s)}\right)\right) f(s)\, ds;$$

i.e., $T(t)$ in (2.12) is given by

$$T(t) = -\frac{\exp(5\pi^2 t)}{\sqrt{\pi t}} \sum_{n=-\infty}^{+\infty} \exp\left(-\frac{n^2}{t}\right).$$

Thus $H(t)$ (a scalar in this case) is derived as

$$H(t) = -\frac{\exp(5\pi^2 t)}{\sqrt{\pi t}} \sum_{n=-\infty}^{+\infty} \exp\left(-\frac{n^2}{t}\right) + \exp(5\pi^2 t) + \exp(4\pi^2 t) + \exp(\pi^2 t).$$

(iv) Finally, set

$$F = A - BW, \quad C = G, \quad D = -B,$$
$$P = \text{row}\left(\tfrac{495}{2}\pi^2, -240\pi^2, \tfrac{35}{2}\pi^2\right).$$

Thus the functional observer is constructed and the resulting closed loop system is represented by

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + au, \qquad 0 < x < 1, \quad t > 0,$$

$$u(x, 0) = u_0(x), \qquad 0 < x < 1,$$

$$\frac{\partial u}{\partial n} = g(x)y(t), \qquad x \in \{0, 1\}, \quad t > 0,$$

$$\frac{dz}{dt} = Fz(t) + Bh(t) + Cy(t) + DH * y(t), \qquad z(0) = z_0,$$

$$h(t) = u(0, t),$$

$$y(t) = Pz(t).$$

In view of Proposition 2, the solution $u(x, t)$ of this system satisfies (3.20); i.e., the stabilization of the heat equation is accomplished.

4. Concluding remarks. The example given in § 1 shows, when the observation and the control are possible only through the boundary, that the controllability or the observability of the parabolic equation does not necessarily enable us to design a static feedback scheme for stabilization in contrast with the case where either the observation or the control can be carried out in the interior. To overcome this difficulty, a functional observer of Luenberger type is constructed (§ 2) and utilized (§ 3) in order to stabilize the parabolic equation. Our observer, however, contains convolutional operations; hence, our stabilization scheme is not purely finite dimensional.

The possibility of the stabilization by means of a dynamic compensator in the feedback path is left for future investigations.

Acknowledgment. The author wishes to express his sincere thanks to Professor Y. Sakawa and Dr. T. Nambu for their encouragements and helpful discussions. The author thanks the reviewer for his critical reading of the manuscript and helpful comments.

REFERENCES

[1] Y. SAKAWA AND T. MATSUSHITA, *Feedback stabilization of a class of distributed systems and construction of a state estimator*, IEEE Trans. Automatic Control, AC-20 (1975), pp. 748–753.

[2] ———, *Stabilization of distributed parameter systems of parabolic type and construction of an observer*, SICE Trans., 11 (1975), pp. 168–174. (In Japanese.)

[3] T. NAMBU, private communications.

[4] Y. SAKAWA, *Observability and stabilization of distributed parameter systems of hyperbolic type*, SICE Trans, 12 (1976), pp. 251–256. (In Japanese.)

[5] M. SLEMROD, *Stabilization of boundary control systems*, J. Differential Equations, 22 (1976), pp. 402–415.

[6] J. K. HALE, *Dynamic systems and stability*, J. Math. Anal. Appl., 26 (1969), pp. 39–59.

[7] J. P. LASALLE, *Stability theory for ordinary differential equations*, J. Differential Equations, 4 (1968), pp. 57–65.

[8] Y. SAKAWA, *Controllability for partial differential equations of parabolic type*, this Journal, 12 (1974), pp. 389–400.

[9] ———, *Observability and related problems for partial differential equations of parabolic type*, this Journal, 13 (1975), pp. 14–27.

[10] D. G. LUENBERGER, *Observers for multivariable systems*, IEEE Trans. Automatic Control, AC-11 (1966), pp. 190–197.

[11] A. FRIEDMAN, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 2, March 1980

ON THE ESTIMATION OF THE PARAMETER OF AN OPTIMAL INTERPOLATOR WHEN THE CLASS OF INTERPOLATORS IS RESTRICTED*

PAUL V. KABAILA† AND GRAHAM C. GOODWIN‡

Abstract. It is usual in time series analysis and control theory to assume that there is a close connection between the structure of the system and the structure of the class of interpolators or control laws under study. Moreover, in practice, restrictive assumptions are often made about the system since this leads ab initio to a simple structure for the optimal interpolator or optimal control law. This paper is concerned with an alternative viewpoint in which attention is focused on the determination of optimal interpolators and control laws from a restricted class when broad assumptions are made about the system. In particular, consistency and asymptotic normality results are developed for estimates of the parameter of an optimal interpolator when the class of interpolators is restricted. Results relevant to the choice of interpolator structure are also established.

1. Introduction. In this paper we will be concerned with the problem of estimating certain specific properties of a system by analyzing data collected from that system. We shall be particularly concerned with sets of properties that we term "noncomprehensive". By this we mean that exact knowledge of these properties is, in general, insufficient to specify all those properties that are usually considered to be of possible interest. For example, consider a weakly stationary stochastic process $\{x_t\}$. The properties that are usually considered to be of possible interest are the covariances $\{\cdots \gamma_{-1}, \gamma_0, \gamma_1, \cdots\}$. A specific property is the value of β (denoted β^*) which minimizes the mean-square prediction error for predictors of the form $\hat{x}_t = \beta x_{t-1}$. In fact $\beta^* = -\gamma_1/\gamma_0$. Note, however, that this property is "noncomprehensive" since it gives insufficient information to specify all those properties of the system that are usually considered to be of possible interest i.e. the covariances $\{\cdots \gamma_{-1}, \gamma_0, \gamma_1, \cdots\}$. A principal advantage in considering non-comprehensive properties is that statistics relating to such properties may usually be analyzed under very weak assumptions on the system generating the data.

Clearly, the philosophy of considering noncomprehensive properties can be applied to such problems as the estimation of the parameters of an optimal control law. The problem then considered is the choice of the best control law from within a restricted class of control laws. However, we confine our considerations to the examination of the limiting properties of certain estimators of the parameters of an optimal interpolator from within a restricted class. Our reasons for considering interpolation are twofold. First, prediction, which is a special form of interpolation, is germane to stochastic control. For example, consider an economic system when the effect of control actions of an individual are negligible. Then an individual's choice of the best control law from a restricted class of control laws can be based on the choice of the best predictor from an appropriate restricted class of predictors. When the control actions affect the properties of the system then the situation becomes more complicated and recursive analysis is required. Second, we shall avoid the additional problems associated with recursive algorithms by concentrating on interpolation. Furthermore, the considerations of linear, time-invariant interpolators allows us to apply the powerful techniques of harmonic analysis.

* Received by the editors November 22, 1977, and in revised form January 21, 1979. This work was supported in part by the Australian Research Grants Committee.

† Division of Mathematics and Statistics, CSIRO, Melbourne, Australia.

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

The basic philosophy of analyzing estimation of the parameters of an optimal predictor from within a restricted class has appeared in a number of recent reports e.g. [1] to [6]. This work should be contrasted with the work of Mann and Wald [7], Whittle [9] to [11], Walker [12] and Hannan [13] on "finite parameter" models for purely nondeterministic time-series in which the spectral density is considered to be the system property of interest. These latter papers analyze the performances of certain estimates of the spectral density under a variety of assumptions. Our concern is to determine some of the limiting properties of estimators of optimal interpolators from a restricted class under weaker assumptions on the system generating the data.

2. A class of interpolators and predictors. Interpolation is the estimation of the value of some sequence $\{x_t\}$ at $t=n$, i.e. x_n, by a function of $\{x_t|t\neq n\}$. Prediction is a special case of interpolation in which x_n is estimated by a function of $\{x_t|t<n\}$.

Here we consider a class of linear time-invariant interpolators of the form

$$\hat{x}_n = -\sum_{\substack{u\\ u\neq 0}} h_u(\theta)x_{n-u} \tag{2.1}$$

where θ denotes a parameter vector belonging to some compact subset Θ of a metric space $\mathcal{M}$ with distance measure $d(\cdot,\cdot)$. The sequence $\{h_u(\theta)\}$ has the value 1 at $u=0$ for all $\theta\in\Theta$. The interpolation error is $W_n(\theta)=x_n-\hat{x}_n=\sum_u h_u(\theta)x_{n-u}$.

We consider three types of processes:

Process Type 1. Consider the sequence $\{x_n(\omega)\in\mathbb{R}\colon n\in\mathbb{Z}, \omega\in \text{some set }\Omega\}$ and let $x_t(\omega)=0$ for $t<0$ independently of ω. Assume that $\gamma_m=\lim_{n\to\infty}(1/n)\sum_{t=1}^n x_t(\omega)x_{t+m}(\omega)$ exists for each $m\in\mathbb{Z}$ and for each $\omega\in\Omega'\subset\Omega$ where $\Omega-\Omega'$ is considered to be an unimportant set. It can be readily shown that

$$\gamma_m=\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^n x_{t+z}(\omega)x_{t+z+m}(\omega)\quad \text{for } z\in\mathbb{Z},\ w\in\Omega;$$

For process type 1, the mean square interpolation error is defined to be:

$$\sigma^2(\theta)=\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^n W_t(\theta)^2.$$

Process Type 2. Suppose $\{x_t(\omega)\in\mathbb{R}\colon t\in\mathbb{Z},\ \omega\in\Omega\}$ is a stochastic process defined on a probability space (Ω, S, P) for which $E\{x_t^2(\omega)\}<\infty$ for each $t\in\mathbb{Z}$ and that $\gamma_m=\lim_{n\to\infty}(1/n)\sum_{t=1}^n E\{x_tx_{t+m}\}$ exists for each $m\in\mathbb{Z}$.

It is also supposed that

$$\frac{1}{n}\sum_{t=1}^n x_tx_{t+m}\overset{\text{a.s}}{\to}\gamma_m$$

i.e. for all $\omega\in\Omega'$ where $P(\Omega-\Omega')=0$.

It can be readily shown that

$$\gamma_m=\lim_{n}\frac{1}{n}\sum_{t=1}^n E\{x_{t+z}x_{t+z+m}\}\quad\text{for } z\in\mathbb{Z}.$$

For process type 2, the mean square interpolation error is defined to be:

$$\sigma^2(\theta)=\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^n E\{w_t^2(\theta)\}.$$

Process Type 3. Suppose $\{x_t(\omega)\in\mathbb{R}\colon t\in\mathbb{Z},\ \omega\in\Omega\}$ is a weakly stationary stochastic process defined on a probability space (Ω, S, P).

Let

$$\gamma_m = E\{x_t x_{t+m}\} \quad \text{for each } m \in \mathbb{Z}.$$

Further suppose that

$$\frac{1}{n}\sum_{t=1}^{n} x_t x_{t+m} \overset{\text{a.s}}{\to} \gamma_m$$

i.e. for all $\omega \in \Omega'$ where $P(\Omega - \Omega') = 0$.

For process Type 3, the mean square interpolation error is defined to be

$$\sigma^2(\theta) = E\{w_t^2(\theta)\}.$$

Note, by Herglotz Theorem [22, p. 281], that for each of the above three processes, there exists a bounded nondecreasing function $F(\lambda)$ such that

$$\gamma_m = \int_{-\pi}^{\pi} e^{im\lambda}\, dF(\lambda). \tag{2.2}$$

A frequency domain expression for $\sigma^2(\theta)$ based on $F(\lambda)$ is given below:

LEMMA 2.1.

$$\sigma^2(\theta) = \int_{-\pi}^{\pi} h(\lambda, \theta)\, dF(\lambda) \tag{2.3}$$

where

$$h(\lambda, \theta) = \left|\sum_u h_u(\theta)\, e^{i\lambda u}\right|^2.$$

Provided

(a) *for process* 1, 2, 3

$$h_u(\theta) = 0 \textit{ for all } u \textit{ such that } |u| > M \in \mathbb{Z};$$

or

(b) *for process* 1

$$\sum_u |h_u(\theta)| < \infty, \qquad h_u(\theta) = 0 \quad \textit{for } u < M \in \mathbb{Z}$$

and

$$\left|\frac{1}{n}\sum_{t=1}^{n} x_{t-u}(\omega) x_{t-v}(\omega)\right| < k \quad \textit{independently of } u, v \textit{ for } \omega \in \Omega',\ n \in \mathbb{N};$$

or

(c) *for process* 3

$$\int_{-\pi}^{\pi} h(\lambda, \theta)\, dF(\lambda) < \infty \quad \textit{or} \quad \sum_u |h_u(\theta)| < \infty.$$

Proof. (a) The proof is by calculation.

(b) $$\frac{1}{n}\sum_{t=1}^{n}\sum_u |h_u x_{t-u}| \sum_v |h_v x_{t-v}| \leqq \frac{1}{n}\sum_{t=1}^{n}\sum_u h_u^2\left(\sum_{v=1}^{n-M} x_v^2\right) < \infty.$$

Hence for each fixed n, the series $(1/n)\sum_{t=1}^{n}\sum_u h_u x_{t-u}\sum_v h_v x_{t-v}$ is absolutely convergent. Hence the order of summation may be rearranged to $\sum_u h_u \sum_v h_v(1/n)\sum_{t=1}^{n} x_{t-u}x_{t-v}$. Next define $f_v^n = h_v(1/n)\sum_{t=1}^{n} x_{t-u}x_{t-v}$. Now $|f_v^n| < k|h_v|$ and $k\sum_v |h_v| < \infty$ by hypothesis. Hence by the dominated convergence theorem for a counting measure [17, p. 273] $\lim_n \sum_u f_u^n = \sum_u h_u\gamma_{u-v}$. Consequently

$$\begin{aligned}\sigma^2(\theta) &= \lim_n \frac{1}{n}\sum_{t=1}^{n}\left(\sum_u h_u x_{t-u}\right)^2\\ &= \sum_u h_u \sum_u h_v \gamma_{u-v}\\ &= \sum_u h_u \sum_v h_u \int_{-\pi}^{\pi} e^{i(u-v)\lambda}\, dF(\lambda)\\ &= \int_{-\pi}^{\pi} h(\lambda, \theta)\, dF(\lambda) \qquad [16, \text{p. } 64].\end{aligned}$$

(c) For the proof see Doob [14, p. 500]. □

An alternative expression for $\sigma^2(\theta)$ is given below:

LEMMA 2.2. *Suppose* $\sigma^2(\theta) = \int_{-\pi}^{\pi} h(\lambda, \theta)\, dF(\lambda) < \infty$. *Then*

$$\sigma^2(\theta) = \sum_s \alpha_s(\theta)\gamma_s \tag{2.4}$$

where

$$\alpha_s(\theta) = \frac{1}{2\pi}\int_{-\pi}^{\pi} h(\lambda, \theta)\, e^{i\lambda s}\, d\lambda \tag{2.5}$$

provided

(a) $\sum_s |\alpha_s(\theta)| < \infty$;

or

(b) $F(\lambda)$ *is absolutely continuous with* $f(\lambda)$ *the Radon–Nikodym derivative of* $F(\lambda)$ *and* $f(\lambda)$, $h(\lambda, \theta) \in L^2$.

Proof. (a) For the proof see [15, Prop. 4.3.7].

(b) The proof follows from Parseval's Theorem for L_2 functions. □

THEOREM 2.1. (a) *If* Θ *is a compact subset of the metric space* $\mathcal{M}$ *with distance measure* $d(\cdot, \cdot)$ *and if* $h(\lambda, \theta)$ *is continuous in* $(\lambda, \theta) \in Z = [-\pi, \pi] \times \Theta$, *then* $\sigma^2(\theta)$ *is continuous in* $\theta \in \Theta$.

(b) $\sum_s |\alpha_s(\theta)| < \infty$ *is a sufficient condition for* $h(\lambda, \theta)$ *to be continuous in* $\lambda \in [-\pi, \pi]$.

Proof. (a) Since $h(\lambda, \theta)$ is continuous in $(\lambda, \theta) \in Z$ and since Z is compact, $h(\lambda, \theta)$ is uniformly continuous on Z. In other words, for any $\varepsilon > 0$ there exists an $\eta > 0$ such that $d(\theta, \theta') < \eta$ implies that $|h(\lambda, \theta) - h(\lambda, \theta')| < \varepsilon$. Hence

$$\begin{aligned}|\sigma^2(\theta') - \sigma^2(\theta)| &\leqq \int_{-\pi}^{\pi} |h(\lambda, \theta') - h(\lambda, \theta)|\, dF(\lambda)\\ &\leqq \varepsilon \int_{-\pi}^{\pi} dF(\lambda) \quad \text{for } d(\theta, \theta') < \eta\\ &= \varepsilon\gamma_0.\end{aligned}$$

Consequently, $\sigma^2(\theta)$ is continuous in θ.

(b) The proof follows from the Weierstrass M test and uniform convergence. □

Remark 2.1. To emphasize the fact that we are dealing with interpolators from within a restricted class we discuss one point of departure from the "classical" case in which the property of prime interest is the spectral density. It is well known [8, p. 33] that in the case of systems described by pth order autoregressive models, that the best k-step-ahead linear predictor of order p can be obtained by concatenating k of the best 1-step-ahead linear predictors of order p. This is not, in general, true in the situation we consider. An example illustrating the distinction is given in Appendix A.

3. An estimator for $\sigma^2(\theta)$. Suppose that we wish to estimate $\sigma^2(\theta)$ on the basis of a finite sample $x_1, \cdots x_n$. A possible estimate is

$$S_n(\theta)=\int_{-\pi}^{\pi} h(\lambda,\theta)I_n(\lambda)\,d\lambda \tag{3.1}$$

where $I_n(\lambda)$ is the periodogram and is given by

$$\begin{aligned} I_n(\lambda)&\triangleq\frac{1}{2\pi n}\left|\sum_{t=1}^{n} x_t e^{i\lambda t}\right|^2 \\ &=\frac{1}{2\pi}\sum_{s=-n+1}^{n-1} C_s e^{i\lambda s}. \end{aligned} \tag{3.2}$$

C_s is the sample covariance, i.e.

$$C_s\triangleq\frac{1}{n}\sum_{t=1}^{n-s} x_t x_{t+s}, \qquad 0\leqq s<n;$$

$$C_{-s}\triangleq C_s, \qquad 0\leqq s<n.$$

Substituting (3.2) into (3.1) we obtain

$$S_n(\theta)=\sum_{s=-n+1}^{n-1} \alpha_s(\theta)C_s. \tag{3.3}$$

A comparison of (3.3) and (2.4) motivates $S_n(\theta)$ as an estimator of $\sigma^2(\theta)$ for it is supposed that C_s converges to γ_s and that $\alpha_s(\theta)\to 0$ as $|s|\to\infty$.

We denote by $\hat{\theta}_n$ any value of θ minimizing $\sigma^2(\theta)$. [Note we do not, for the moment, necessarily assume that $S_n(\theta)$ is minimized at a single value].

Remark 3.1. The theory described in this paper goes through, essentially unaltered, for other estimators of $\sigma^2(\theta)$. For example, we could consider

$$\tilde{S}_n(\theta)\triangleq\frac{2\pi}{n}\sum_j h(\lambda_j,\theta)I_n(\lambda_j)$$

where $\lambda_j=-\pi+(2\pi/n)j$, $j=0,\cdots(n-1)$ or

$$\tilde{\tilde{S}}_n(\theta)\triangleq\frac{\pi}{n}\sum_j h(\lambda_j,\theta)I_n(\lambda_j)$$

where $\lambda_j=-\pi+(\pi/n)j$, $j=0,\cdots(2n-1)$.

Also, it is possible to broaden the class of processes somewhat, e.g. to include cases where γ_m depends upon the realization or where the sample covariances are replaced by sample correlations as in [24].

For simplicity of exposition we will not consider these extensions. Details are given in [19].

4. The limiting behaviour of $\hat{\theta}_n$. We consider the class of interpolators and the processes introduced in § 2. The key assumptions are:

(S1) $\theta \in \Theta$ a nonempty compact subset of a metric space;

(S2) $h(\lambda, \theta)$ is continuous in $(\lambda, \theta) \in [-\pi, \pi] \times \Theta = Z$;

(S3) Equations (2.3) and (2.4) are valid, i.e.

$$\sigma^2(\theta) = \int_{-\pi}^{\pi} h(\lambda, \theta)\, dF(\lambda) = \sum_s \alpha_s(\theta)\gamma_s < \infty.$$

LEMMA 4.1. *For processes* 1, 2, 3 *and subject to assumptions* (S1), (S2), (S3), $\lim_n S_n(\theta) = \sigma^2(\theta)$ *uniformly in* $\theta \in \Theta$ *for each* $\omega \in \Omega'$. *Thus, for example if* $Z_n \triangleq \sup_{\theta\in\Theta} |S_n(\theta) - \sigma^2(\theta)|$ *then* $\lim_n Z_n(\omega) = 0$ *for each* $\omega \in \Omega'$.

Proof. Fix $\omega \in \Omega'$. Let $q_m(\lambda, \theta)$ be the Cesaro sum of the Fourier series of $h(\lambda, \theta)$ taken to M terms i.e.

$$q_m(\lambda, \theta) \triangleq \sum_{n=-M}^{M} \alpha_n(\theta)\left(1 - \frac{|n|}{M}\right) e^{in\lambda}.$$

Also let

$$J_1(\theta) = \left|S_n(\theta) - \int_{-\pi}^{\pi} q_M(\lambda, \theta) I_n(\lambda)\, d\lambda\right|,$$

$$J_2(\theta) = \left|\int_{-\pi}^{\pi} q_M(\lambda, \theta) I_n(\lambda)\, d\lambda - \int_{-\pi}^{\pi} q_M(\lambda, \theta)\, dF(\lambda)\right|,$$

$$J_3(\theta) = \left|\int_{-\pi}^{\pi} q_M(\lambda, \theta)\, dF(\lambda) - \sigma^2(\theta)\right|.$$

Our interest in these quantities is motivated by the fact that

$$|S_n(\theta) - \sigma^2(\theta)| \leqq J_1(\theta) + J_2(\theta) + J_3(\theta).$$

Now given $\varepsilon_1 > 0$, we may fix M so large that $|h(\lambda, \theta) - q_M(\lambda, \theta)| < \varepsilon_1$ uniformly in $(\lambda, \theta) \in Z$ since the Cesaro sum converges uniformly in $(\lambda, \theta) \in Z$.

Now

$$\begin{aligned} J_1(\theta) &\leqq \int_{\pi}^{\pi} |h(\lambda, \theta) - q_M(\lambda, \theta)| I_n(\lambda)\, d\lambda \\ &\leqq \varepsilon_1 \int_{\pi}^{\pi} I_n(\lambda)\, d\lambda = \varepsilon_1 C_0. \end{aligned}$$

Similarly

$$J_3(\theta) \leqq \varepsilon_1 \gamma_0.$$

Also

$$J_2(\theta) = \left|\sum_{s=-M}^{M} (C_s - \gamma_s)\alpha_s(\theta)\left(1 - \frac{|s|}{M}\right)\right|.$$

Since Z is a compact set and $h(\lambda, \theta)$ is continuous on that set, we may define $K = \max_{(\lambda,\theta)\in Z} |h(\lambda, \theta)| < \infty$. This implies that independently of $\theta \in \Theta$, $|\alpha_s(\theta)| \leqq K$. Hence $J_2(\theta) \leqq K \sum_{s=-M}^{M} |C_s - \gamma_s|$.

Now suppose that we are given $\varepsilon > 0$. Fix M so large that $\varepsilon_1\gamma_0 < \varepsilon/6$. Hence for $n_0 \in N_+$ sufficiently large $J_1(\theta) + J_3(\theta) < 3\varepsilon_1\gamma_0 < \varepsilon/2$ for all $n \geqq n_0$ independently of $\theta \in \Theta$. Hence $J_1(\theta) + J_2(\theta) + J_3(\theta) < \varepsilon$ for $n \geqq \max(n_0, n_1)$ uniformly in $\theta \in \Theta$. □

THEOREM 4.1. *For processes type* 1, 2, 3 *and subject to assumptions* (S1), (S2), (S3), *if* Θ_0 *denotes the set of* θ*'s minimizing* $\sigma^2(\theta)$ (*the minimum is denoted* $\sigma^2(\theta_0)$) *and* $\hat{\theta}_n$ *is any minimizing value of* $S_n(\theta)$, *then*

$$\lim_n d(\hat{\theta}_n, \Theta_0) = 0 \quad \textit{for } \omega \in \Omega',$$

$$\lim S_n(\hat{\theta}_n) = \sigma^2(\theta_0) \quad \textit{for } \omega \in \Omega'.$$

Proof. Fix $\omega \in \Omega'$. First note that Θ_0 is nonempty, since Θ is compact and $\sigma^2(\theta)$ is continuous (Theorem 2.1).

Then from Lemma 4.1, given $\varepsilon > 0$ there exists an n_0 such that for all $n \geqq n_0$ $|S_n(\theta) - \sigma^2(\theta)| < \varepsilon$ for all $\theta \in \Theta$. Hence

$$\sigma^2(\hat{\theta}_n) - \varepsilon \leqq S_n(\hat{\theta}_n) \leqq S_n(\theta_0) < \sigma^2(\theta_0) + \varepsilon$$

which implies that $\sigma^2(\hat{\theta}_n) < \sigma^2(\theta_0) + 2\varepsilon$. Consequently $\lim_n \sigma^2(\hat{\theta}_n) = \sigma^2(\theta_0)$. Now

$$|S_n(\hat{\theta}_n) - \sigma^2(\theta_0)| \leqq |S_n(\hat{\theta}_n) - \sigma^2(\hat{\theta}_n)| + |\sigma^2(\hat{\theta}_n) - \sigma^2(\theta_0)|.$$

The first term tends to zero by Lemma 4.1 and we have just shown the second tends to zero also. Hence $\lim_n S_n(\hat{\theta}_n) = \sigma^2(\theta_0)$ and the second part of the theorem is proved.

For convenience denote $d(\hat{\theta}_n, \theta_0)$ by y_n. We now claim $\lim_n y_n = 0$. This is proved by contradiction.

Suppose y_n does not converge to zero, then there exists an $\varepsilon > 0$ such that there is a subsequence of $\{n\}$ denoted $\{n_i\}$ for which $y_{n_i} > \varepsilon$ for all $i \in \mathbb{N}_+$. Consider the closed set $C \subset \Theta$ of θ's for which $d(\theta, \Theta_0) \geqq \varepsilon$. Clearly C is compact. Hence $\{n_i\}$ has a subsequence denoted $\{n_j\}$ for which $\{\hat{\theta}_{n_j}\}$ converges to $\theta^* \in C$. Obviously, $\theta^* \notin \Theta_0$. Hence $\lim_j \sigma^2(\hat{\theta}_{n_j}) = \sigma^2(\theta^*) > \sigma^2(\theta_0)$. But $\lim_j \sigma^2(\hat{\theta}_{n_j}) = \sigma^2(\theta_0)$. This contradiction establishes the first part of the theorem. □

5. Limiting behaviour of $\hat{\theta}_n$ under weaker assumptions. For processes 2, 3 it has been assumed that $(1/n)\sum_{t=1}^n x_t x_{t+n} \xrightarrow{\text{a.s.}} \gamma_m$ leading to the almost sure "consistency" results obtained in § 4. Here we consider,

Process Type 2′, 3′: As for processes type 2, 3 respectively excepting that

$$\frac{1}{n}\sum_{t=1}^{n} x_t x_{t+n} \xrightarrow{\text{Prob}} \gamma_m.$$

The following theorem allows us to convert the almost sure convergence results obtained in § 4 to in probability convergence results for processes type 2′, 3′.

THEOREM 5.1. *Suppose that under a certain set of conditions* (*call these condition* C).

$$a_n^i \overset{\text{a.s.}}{\to} a^i \quad \textit{for } i = 1, 2, \cdots \quad \textit{implies} \quad b_n \overset{\text{a.s.}}{\to} b.$$

Then under conditions C

$$a_n^i \xrightarrow{\text{Prob}} a^i \quad \textit{for } i = 1, 2 \cdots \quad \textit{implies} \quad b_n \xrightarrow{\text{Prob}} b.$$

Proof [19]. The theorem follows from the well known fact that $x_n \xrightarrow{\text{Prob}} x$ if and only if every subsequence of the x_j's contains a further subsequence which converges to x almost surely. □

Theorems 4.1 and 5.1 can be combined to yield:

THEOREM 5.2. *For processes type* $2'$, $3'$ *and subject to assumptions* (S1), (S2), (S3), *if* Θ_0 *denotes the set of* θ*'s minimizing* $\sigma^2(\theta)$ *and* $\hat{\theta}_n$ *denotes any minimizing value of* $S_n(\theta)$, *then*

$$\min_{\theta\in\Theta_0} d(\hat{\theta}_n, \theta) \xrightarrow{\text{Prob}} 0, \qquad S_n(\hat{\theta}_n) \xrightarrow{\text{Prob}} \sigma^2(\theta_0).$$

6. Limiting distributions relevant to $\hat{\theta}_n$. To obtain more detailed information about the way in which $\hat{\theta}_n$ converges we specialise to a consideration of a process $\{x_t\}$ which satisfies the following assumptions.

A1: $x_t = \sum_u l_u \varepsilon_{t-u}$ where $l_0 = 1$ and $l_u = 0$ for $u < 0$.

A2: $E\{\varepsilon_n | \mathscr{F}_{n-1}\} = 0$ a.s. for all n where $\mathscr{F}_n$ is the σ-field generated by $\{\varepsilon_n, \varepsilon_{n-1}, \cdots\}$.

A3: $E\{\varepsilon_n^2 | \mathscr{F}_{n-1}\} = \sigma^2 > 0$ a.s.

A4: Suppose there exists a random variable X with $E\{X^4\} < \infty$ such that $P\{|\varepsilon_n| > u\} \leqq cP\{|X| > u\}$ for some $0 < c < \infty$ and all n, all $u \geqq 0$.

B1: $\sum_u l_u^2 < \infty$;

B2: $\sum_u u l_u^2 < \infty$.

It is obvious that Assumptions A2 and A3 imply that $E\{\varepsilon_n \varepsilon_m\} = \sigma^2 \delta_{nm}$. Since B2 $\Rightarrow$ B1, Assumptions A1, A2, A3 and B2 imply that $\{x_t\}$ has an absolutely continuous spectral distribution (see, for example, [14, p. 499]). We denote the spectral density by $f(\lambda)$. Note that $\gamma_n = \sigma^2 \sum_u l_u l_{u+n}$. Under Assumptions A1, A2, A3 and B1, Hannan and Heyde [18] have proved that $(1/n)\sum_{t=1}^n x_t x_{t+m} \xrightarrow{\text{Prob}} \gamma_m$. Consequently a process satisfying A1, A2, A3 and B2 is necessarily a Type $3'$ process.

Consider the class of interpolators introduced in § 2. The mean-square interpolation error is given by $\sigma^2(\theta) = \int_{-\pi}^{\pi} h(\lambda, \theta) f(\lambda)\, d\lambda$. Let us now introduce the following assumptions on the class of interpolators:

C1: Suppose $\mathcal{M}$ is $\mathscr{R}^p$ and Θ is a compact subset of $\mathscr{R}^p$. Hence $\theta = [\theta_1, \cdots, \theta_p]^T$. It is also supposed that the Θ_0 defined in § 4 consists of a single element $\theta_0 \in \Theta - \text{bd}\,\Theta$. Here bd Θ denotes the boundary of the set Θ.

C2: $h(\lambda, \theta)$ is continuous in $(\lambda, \theta) \in Z$. There exists a neighborhood of θ_0 denoted N_0 in which $h(\lambda, \theta)$ is a twice differentiable function of the θ_i whose second derivatives w.r.t. θ_i are continuous in (λ, θ) for $\theta \in N_0$.

Let $h^{(i)}(\lambda, \theta)$ denote $\partial h(\lambda, \theta)/\partial\theta_i$ and $h^{(ij)}(\lambda, \theta)$ denote $\partial^2 h(\lambda, \theta)/\partial\theta_i\, \partial\theta_j$. Clearly assumptions C1 and C2 imply that $\sigma^2(\theta)$ may be differentiated under the integral sign so that

$$\int_{-\pi}^{\pi} h^{(i)}(\lambda, \theta_0) f(\lambda)\, d\lambda = 0. \tag{6.1}$$

Now Theorem 5.2 implies that under Assumptions A1, A2, A3, B1, C1 and C2, $d(\hat{\theta}_n, \theta_0) \xrightarrow{\text{Prob}} 0$, $\hat{S}_n(\hat{\theta}) \xrightarrow{\text{Prob}} \sigma^2(\theta_0)$.

THEOREM 6.1. *Suppose Assumptions* A1, A2, A3, A4, B2, C1 *and* C2 *hold. Let* $\boldsymbol{\psi} = (\psi_{ij})$, $\Phi = (\phi_{ij})$ *where*

$$\psi_{ij} \triangleq 4\pi \int_{-\pi}^{\pi} h^{(i)}(\lambda, \theta_0) h^{(j)}(\lambda, \theta_0) f^2(\lambda)\, d\lambda,$$

$$\phi_{ij} \triangleq \int_{-\pi}^{\pi} h^{(ij)}(\lambda, \theta_0) f(\lambda)\, d\lambda.$$

Now provided Φ *is nonsingular* $n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{\text{dist}} N(0, \Phi^{-1}\psi\Phi^{-1})$.

Proof. Since $\hat{\theta}_n \xrightarrow{\text{Prob}} \theta_0$, the limiting distribution can be obtained on the assumption that $\hat{\theta}_n \in N_0$. Now by the mean value theorem

$$0 = S_n^{(j)}(\theta_0) + \sum_{i=1}^{p} (\hat{\theta}_{n,i} - \theta_{0,i}) S_n^{(ij)}(\theta_n^*)$$

where $\theta_n^* = \lambda\hat{\theta}_n + (1-\lambda)\theta_0$ and $0<\lambda<1$. Hence

$$\sum_{i=1}^{p} \{-S_n^{(ij)}(\theta_n^*)\}\{n^{1/2}(\hat{\theta}_{n,i} - \theta_{0,i})\} = n^{1/2} S_n^{(j)}(\theta_0). \tag{6.2}$$

The proof is divided into two parts:

Part a. The proof that $S_n^{(ij)}\ (\theta_n^*) \xrightarrow{\text{Prob}} \phi_{ij}$ is given in Appendix B.

Part b. We prove that the limiting distribution of $n^{1/2}S_n^{(i)}(\theta_0)$ $(1 \leqq i \leqq p)$ is $N(0, \psi)$ as follows

$$n^{1/2}S_n^{(j)}(\theta_0) = n^{1/2} \int_{-\pi}^{\pi} h^{(j)}(\lambda, \theta_0) I_n(\lambda)\, d\lambda. \tag{6.3}$$

Now let us define $I(\lambda, \varepsilon) \triangleq (1/2\pi n)|\sum_{t=1}^{n} \varepsilon_t e^{i\lambda t}|^2$. As proved by Hannan [13] we may replace consideration of the expression (6.3) by

$$n^{1/2} \int_{-\pi}^{\pi} I(\lambda, \varepsilon) f(\lambda) h^{(j)}(\lambda, \theta_0)\, d\lambda. \tag{6.4}$$

But (6.1) implies that expression (6.4) may be replaced by

$$n^{1/2} \int_{-\pi}^{\pi} \left\{ I(\lambda, \varepsilon) - \frac{1}{2\pi n} \sum_{t=1}^{n} \varepsilon^2 \right\} f(\lambda) h^{(j)}(\lambda, \theta_0)\, d\lambda.$$

As shown in [13] this type of expression can be reduced to the consideration of an expression which is asymptotically normal by a result of Hannan and Heyde [18]. Finally, $n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{\text{dist}} N(0, \Phi^{-1}\psi\Phi^{-1})$ from (6.2) and [16, pp. 254–255]. □

COROLLARY 6.1. *Under the conditions that make Theorem* 6.1 *valid* $n[\sigma^2(\hat{\theta}_n) - \sigma^2(\theta_0)]$ *is in the limit distributed as* $\frac{1}{2}y^T\Phi y$ *where y is distributed as* $N(0, \Phi^{-1}\psi\Phi^{-1})$. *Let it be remarked that* $E\{\frac{1}{2}y^T\Phi y\} = \frac{1}{2}\text{tr}\{\psi\Phi^{-1}\}$.

Proof. By the mean value theorem

$$\sigma^2(\hat{\theta}_n) = \sigma^2(\theta_0) + \tfrac{1}{2} \sum_{i,j=1}^{p} (\hat{\theta}_{n,i} - \theta_{0,i})(\hat{\theta}_{n,j} - \theta_{0,j}) \int_{-\pi}^{\pi} h^{(ij)}(\lambda, \theta_n^*) f(\lambda)\, d\lambda$$

where $\theta_n^* = \lambda\theta_0 + (1-\lambda)\hat{\theta}_n$, $0<\lambda<1$. Hence

$$\begin{aligned} &n[\sigma^2(\hat{\theta}_n) - \sigma^2(\theta_0)] \\ &= \tfrac{1}{2} \sum_{i,j=1}^{p} n^{1/2}(\hat{\theta}_{n,i} - \theta_{0,i}) n^{1/2}(\hat{\theta}_{n,j} - \theta_{0,j}) \int_{-\pi}^{\pi} h^{(ij)}(\lambda, \theta_n^*) f(\lambda)\, d\lambda. \end{aligned} \tag{6.5}$$

The result now follows from Theorem 6.1. □

COROLLARY 6.2. *Under the conditions for the validity of Theorem* 6.1 $n[S_n(\theta_0) - S_n(\hat{\theta}_n)]$ *is in the limit distributed as* $\frac{1}{2}y^T\Phi y$ *where y is distributed as* $N(0, \Phi^{-1}\psi\Phi^{-1})$. *Let it be remarked that* $E\{\frac{1}{2}y^T\Phi y\} = \frac{1}{2}\text{tr}\,\{\psi\Phi^{-1}\}$.

Proof. By the mean value theorem

$$S_n(\theta_0)=S_n(\hat{\theta}_n)+\tfrac{1}{2}\sum_{i,j=1}^{p}(\theta_{0,i}-\hat{\theta}_{n,i})(\theta_{0,j}-\hat{\theta}_{n,j})S_n^{(ij)}(\theta_n^*)$$

where $\theta_n^*=\lambda\theta_0+(1-\lambda)\hat{\theta}_n$, $0<\lambda<1$. Consequently

$$n[S_n(\theta_0)-S_n(\hat{\theta}_n)]=\sum_{i,j=1}^{p}n^{1/2}(\theta_{0,i}-\hat{\theta}_{n,i})n^{1/2}(\theta_{0,j}-\hat{\theta}_{n,j})S_n^{(ij)}(\theta_n^*).$$

The result now follows from Theorem 6.1 and the fact that $S_n^{(ij)}(\theta_n^*)\xrightarrow{\text{Prob}}\phi_{ij}$. □

7. Additional limiting distributions relevant to $\hat{\theta}_n$. In § 6 the limiting distributions relevant to $\hat{\theta}_n$ were developed under fairly weak assumptions on $\{x_t\}$. To obtain more results assumptions additional to those of § 6 need to be introduced. Here we assume A1, A2, A3 as in the last section plus we introduce

A5: $E\{\varepsilon_n^3\varepsilon_m\}=E\{\varepsilon_n^3\}E\{\varepsilon_m\}$ for $n>m$.

A6: $E\{\varepsilon_n^4\}=\mu_4<\infty$. Define $\kappa_4=\mu_4-3\sigma^4$.

B3: $\sum_u u|l_u|<\infty$ (note that B3 $\Rightarrow$ B2, see Theorem C.1 of Appendix C).

Let it be remarked firstly that the above set of assumptions on the process is more stringent than the set of assumptions A1, A2, A3, A4 and B2 of the previous section.

Clearly, Assumptions A1, A2, A3, A5 and A6 are satisfied by $\{\varepsilon_n\}$ a sequence of independent random variables with $E\{\varepsilon_n\}=0$, $E\{\varepsilon_n^2\}=\sigma^2>0$ and $E\{\varepsilon_n^4\}=\mu_4<\infty$. However, Assumptions A1, A2, A3, A5 and A6 are weaker.

Assumptions A2 and A3 have been chosen because they imply that powerful convergence and central limit results hold. Assumptions A2, A3, A5 and A6, together, have been chosen because expressions of the form $E\{\varepsilon_i\}$, $E\{\varepsilon_i\varepsilon_j\}$ and $E\{\varepsilon_i\varepsilon_j\varepsilon_k\varepsilon_l\}$ have the values that would have been ascribed to them had the ε_t's been a sequence of independent random variables with $E\{\varepsilon_t\}=0$, $E\{\varepsilon_t^2\}=\sigma^2>0$ and $E\{\varepsilon_t^4\}=\mu_4<\infty$. (See Lemma C.1 of Appendix C.)

Assumption B3 implies that $\sum|t\gamma_t|<\infty$, see Theorem C.1 of Appendix C. The condition $\sum|t\gamma_t|<\infty$ implies that not only is $f(\cdot)$ continuous but it is also differentiable everywhere with a bounded derivative for $\lambda\in[-\pi,\pi]$. (See, again, Theorem C.1 of Appendix C.)

In addition to Assumptions C1 and C2 on the class of interpolators introduced in the previous section we also require:

C3: $h^{(i)}(\lambda,\theta)$ is such that $\sum_s|\alpha_s^{(i)}(\theta)|<\infty$ for each $\theta\in N_0$.

In Appendix C are proved a number of results pertaining to quantities related to $S_n(\theta)$.

It will be supposed throughout this section that Assumptions A1, A2, A3, A5, A6, B3, C1, C2 and C3 hold. The results developed in Appendix C will now be used to prove several theorems.

THEOREM 7.1. *For Φ nonsingular $n^{1/2}(S_n(\hat{\theta}_n)-\sigma^2(\theta_0))\xrightarrow{\text{dist}}N(0,\delta)$ where*

$$\delta\triangleq 4\pi\int_{\pi}^{\pi}[h(\lambda,\theta_0)f(\lambda)]^2\,d\lambda+\frac{\kappa_4}{\sigma^4}(\sigma^2(\theta_0))^2.$$

Also $\delta\geqq(2+\kappa_4/\sigma^4)(\sigma^2(\theta_0))^2$.

Proof. By the mean value theorem

$$S_n(\theta_0)=S_n(\hat{\theta}_n)+\tfrac{1}{2}\sum_{i,j=1}^{p}(\theta_{0,i}-\hat{\theta}_{n,i})(\theta_{0,j}-\hat{\theta}_{n,j})S_n^{(ij)}(\theta_n^*)$$

where $\theta_n^* = \lambda\theta_0 + (1-\lambda)\hat{\theta}_n$, $0 < \lambda < 1$.

Consequently

$$n^{1/2}[S_n(\hat{\theta}_n) - S_n(\theta_0)] = -\frac{n^{-1/2}}{2} \sum_{i,j=1}^{p} n^{1/2}(\theta_{0,i} - \hat{\theta}_{n,i}) n^{1/2}(\theta_{0,j} - \hat{\theta}_{n,j}) S_n^{(ij)}(\theta_n^*).$$

By Theorem 6.1 $n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{\text{dist}} N(0, \Phi^{-1}\psi\Phi^{-1})$. Also, the proof of Theorem 6.1 includes a proof of the fact that $S_n^{(ij)}(\theta_n^*) \xrightarrow{\text{Prob}} \phi_{ij}$.

Hence $n^{1/2}S_n(\hat{\theta}_n) = n^{1/2}S_n(\theta_0) + \kappa_n$ where $\kappa_n \xrightarrow{\text{Prob}} 0$. Also by Result C3.

$$n^{1/2}E\{S_n(\theta_0)\} = n^{1/2}\sigma^2(\theta_0) + h_0 \quad \text{where } \lim_n h_n = 0.$$

Hence

$$n^{1/2}[S_n(\hat{\theta}_n) - \sigma^2(\theta_0)] = n^{1/2}[S_n(\theta_0) - E\{S_n(\theta_0)\}] + m_n \quad \text{where } m_n \xrightarrow{\text{Prob}} 0.$$

Hence by a result of Cramer [16, § 20.6] the limiting distribution of $n^{1/2}[S_n(\hat{\theta}_n) - \sigma^2(\theta_0)]$ is the same as that of $n^{1/2}[S_n(\theta_0) - E\{S_n(\theta_0)\}]$. Now by Theorem C.4 of Appendix C $n^{1/2}[S_n(\theta_0) - E\{S_n(\theta_0)\}] \xrightarrow{\text{dist}} N(0, \delta)$.

If $\sigma^2(\theta_0)$ is constrained to equal $\int_{-\pi}^{\pi} h(\lambda, \theta_0) f(\lambda)\, d\lambda$ then a calculus of variations argument proves that the constrained minimum value of δ is $(2 + \kappa_4/\sigma^4)(\sigma^2(\theta_0))^2$. □

COROLLARY 7.1. $n^{1/2}[S_n(\hat{\theta}_n) - \sigma^2(\hat{\theta}_n)] \xrightarrow{\text{dist}} N(0, \delta)$ *where* δ *is defined as in the statement of Theorem* 7.1.

Proof. $n^{1/2}(S_n(\hat{\theta}_n) - \sigma^2(\hat{\theta}_n)) - n^{1/2}(S_n(\hat{\theta}_n) - \sigma^2(\theta_0)) = n^{1/2}(\sigma^2(\theta_0) - \sigma^2(\hat{\theta}_n)) \xrightarrow{\text{Prob}} 0$ by Corollary 6.1. The result now follows from Theorem 7.1. □

THEOREM 7.2. *For* Φ *nonsingular*

$$n^{1/2}\begin{bmatrix} S_n(\hat{\theta}_n) - \sigma^2(\theta_0) \\ \hat{\theta}_n - \theta_0 \end{bmatrix} \xrightarrow{\text{dist}} N(0, B)$$

where

$$B \triangleq \begin{bmatrix} \delta & \nu^T\Phi^{-1} \\ \Phi^{-1}\nu & \Phi^{-1}\psi\Phi^{-1} \end{bmatrix}$$

where $\nu = (\nu_j)$ *and* $\nu_j = 4\pi \int_{-\pi}^{\pi} h(\lambda, \theta_0) h^{(j)}(\lambda, \theta_0) f^2(\lambda)\, d\lambda$.

Proof. As in the proof of Theorem 7.1

$$n^{1/2}[S_n(\hat{\theta}_n) - \sigma^2(\theta_0)] = n^{1/2}[S_n(\theta_0) - E\{S_n(\theta_0)\}] + m_n$$

where $m_n \xrightarrow{\text{Prob}} 0$. Thus, when finding the limiting distribution we need only consider $n^{1/2}[S_n(\theta_0) - E\{S_n(\theta_0)\}]$.

Also, by Result C.3, $n^{1/2}E\{S_n^{(j)}(\theta_0)\} = k_n$ where $\lim_n k_n = 0$. Hence, when finding the limiting distribution of $n^{1/2}S_n^{(j)}(\theta_0)$ we may instead consider $n^{1/2}[S_n^{(j)}(\theta_0) - E\{S_n^{(j)}(\theta_0)\}]$.

As a preliminary we will consider the limiting distribution of

$$n^{1/2}\begin{bmatrix} S_n(\theta_0)-E\{S_n(\theta_0)\} \\ S_n^{(1)}(\theta_0)-E\{S_n^{(1)}(\theta_0)\} \\ \vdots \\ S_n^{(p)}(\theta_0)-E\{S_n^{(p)}(\theta_0)\} \end{bmatrix}. \tag{7.1}$$

By Theorem C.4 the vector (7.1) has a limiting distribution $N(0, A)$ where

$$A \triangleq \begin{bmatrix} \delta & \nu^T \\ \nu & \psi \end{bmatrix}.$$

Now recall equation (6.2) viz.

$$\sum_{i=1}^{p} \{-S_n^{(ij)}(\theta_n^*)\}\{n^{1/2}(\hat{\theta}_{n,i}-\theta_{0,i})\} = n^{1/2} S_n^{(j)}(\theta_0).$$

Hence the limiting distribution of

$$n^{1/2}\begin{bmatrix} S_n(\theta_0)-E\{S_n(\theta_0)\} \\ \hat{\theta}_{n,1}-\theta_{0,1} \\ \vdots \\ \hat{\theta}_{n,p}-\theta_{0,p} \end{bmatrix}$$

is $N(0, B)$ where

$$B \triangleq \begin{bmatrix} \delta & \nu^T\Phi^{-1} \\ \Phi^{-1}\nu & \Phi^{-1}\psi\Phi^{-1} \end{bmatrix}. \qquad \square$$

8. Limit results for two different interpolator classes on the basis of the same data. The methodology of §§ 6 and 7 can be used to examine the following kind of situation which has no analogue in the classical case (i.e. when the property of interest is the spectral density.)

From the outset we restrict our attention to processes of the type 3′. Suppose we are concerned with two classes of interpolators:

(1) $\hat{x}_t = -\sum_u' h_u(\theta)x_{t-u}$ where $\{h_u(\theta)\}$ is a sequence of reals for which $h_0(\theta)=1$ for all $\theta \in \Theta$ a subset of $\mathscr{R}^p$.

(2) $\hat{x}_t = -\sum_u' m_u(\phi)x_{t-u}$ where $\{m_u(\phi)\}$; is a sequence of reals for which $m_0(\phi)=1$ for all $\phi \in \Phi$ a subset of $\mathscr{R}^q$. In other words, $\phi = [\phi_1, \cdots, \phi_q]^T$.

The mean-square interpolation error for the first class of interpolators has already been defined by

$$\sigma^2(\theta) \triangleq \int_{-\pi}^{\pi} h(\lambda, \theta) f(\lambda)\, d\lambda.$$

The mean-square interpolation error for the second class of interpolators is denoted $\delta^2(\phi)$ and is defined by

$$\delta^2(\phi) \triangleq \int_{-\pi}^{\pi} m(\lambda, \phi) f(\lambda)\, d\lambda; \qquad m(\lambda, \phi) \triangleq \left|\sum_u m_u e^{iu\lambda}\right|^2.$$

The estimator of $\sigma^2(\theta)$ we will consider is $S_n(\theta)$ which has already been defined. The estimator of $\delta^2(\phi)$ we will consider is $V_n(\phi)$ defined by

$$V_n(\phi) \triangleq \int_{-\pi}^{\pi} I_n(\lambda) m(\lambda, \phi)\, d\lambda = \sum_{r=-n+1}^{n-1} \beta_r(\phi) c_r$$

where

$$\beta_r(\phi) \triangleq \frac{1}{2\pi}\int_{\pi}^{\pi} m(\lambda, \phi)\, e^{i\lambda r}\, d\lambda.$$

Also, let $\hat{\phi}_n$ denote a value of ϕ which minimizes $V_n(\phi)$. The following assumptions concerning $m(\lambda, \phi)$ will be referred to in this section. They are clearly analogous to Assumptions C1, C2 and C3 for $h(\lambda, \theta)$.

C1′: Φ is a compact subset of $\mathscr{R}^q$. It is also supposed that $\delta^2(\phi)$ is minimized at a single value of ϕ, denoted ϕ_0, and that $\phi_0 \in \Phi - \text{bd}\,\Phi$.

C2′: $m(\lambda, \phi)$ is continuous in $(\lambda, \phi) \in [-\pi, \pi] \times \Phi$. There exists a neighborhood of ϕ_0 denoted M in which $m(\lambda, \phi)$ is a twice differentiable function of the ϕ_j whose second derivatives w.r.t. ϕ_j are continuous in (λ, ϕ) for $\phi \in M_0$.

Let $m^{(i)}(\lambda, \phi)$ denote $\partial m(\lambda, \phi)/\partial\phi_i$ and $m^{(ij)}(\lambda, \phi)$ denote $\partial^2 m(\lambda, \phi)/\partial\phi_i\, \partial\phi_j$.

C3′: $m^{(i)}(\lambda, \phi)$ is such that $\sum|\beta_s^{(i)}(\phi)| < \infty$, $\phi \in M_0$.

Clearly, Assumptions C1′ and C2′ imply that $\delta^2(\phi)$ may be differentiated under the integral sign at ϕ_0 i.e. $\int_{-\pi}^{\pi} m^{(i)}(\lambda, \phi_0) f(\lambda)\, d\lambda = 0$.

Our motivation for considering two classes of interpolators is that we wish to develop results pertaining to measurement of the relative performance of two different interpolation classes based on the one set of data. Of course, $S_n(\theta)$ gives a measure of the relative performance of different interpolators from the one class for different values of θ. The theorems in previous sections are therefore concerned with the question "What can be said about the best member of a particular class of interpolators?" Here we are concerned with the comparison of members of two (or more) interpolator classes and our theory relates to the question "What can be said about the best interpolator from several classes of interpolators?".

THEOREM 8.1. *Under Assumptions* A1, A2, A3, A4, B2, C1, C2, C1′ *and* C2′, *and supposing* Φ, Ξ, *introduced below, are nonsingular then the limiting distribution of*

$$n^{1/2}\begin{bmatrix} \hat{\theta}_n - \theta_0 \\ \hat{\phi}_n - \phi_0 \end{bmatrix} \quad \textit{is} \quad N\left(0, \begin{bmatrix} A & B \\ B^T & C \end{bmatrix}\right)$$

where $A \triangleq \Phi^{-1}\psi\Phi^{-1}$, $B \triangleq \Phi^{-1}\Omega\Xi^{-1}$ *and* $C \triangleq \Xi^{-1}\Gamma\Xi^{-1}$ *and* $\Phi = (\phi_{ij})$, $\boldsymbol{\psi} = (\psi_{ij})$, $\Xi = (\Xi_{ij})$, $\Gamma = (\Gamma_{ij})$ *and* $\Omega = (\Omega_{ij})$ *where* $\phi_{ij} \triangleq \int_{-\pi}^{\pi} h^{(ij)}(\lambda, \theta_0) f(\lambda)\, d\lambda$ *as before*

$$\Xi_{ij} \triangleq \int_{-\pi}^{\pi} m^{(ij)}(\lambda, \theta_0) f(\lambda)\, d\lambda,$$

$$\psi_{ij} \triangleq 4\pi \int_{-\pi}^{\pi} h^{(i)}(\lambda, \theta_0) h^{(j)}(\lambda, \theta_0) f^2(\lambda)\, d\lambda,$$

$$\Gamma_{ij} \triangleq 4\pi \int_{-\pi}^{\pi} m^{(i)}(\lambda, \theta_0) m^{(j)}(\lambda, \theta_0) f^2(\lambda)\, d\lambda,$$

$$\Omega_{ij} \triangleq 4\pi \int_{-\pi}^{\pi} h^{(i)}(\lambda, \theta_0) m^{(j)}(\lambda, \theta_0) f^2(\lambda)\, d\lambda.$$

Proof. By Theorem 5.2 $\hat{\theta}_n \xrightarrow{\text{Prob}} \theta_0$, $\hat{\phi}_n \xrightarrow{\text{Prob}} \phi_0$. Hence we are able to obtain the limiting distribution on the assumption that $\hat{\theta}_n \in N_0$, $\hat{\phi}_n \in M_0$.

Now by the mean value theorem:

$$\sum_{i=1}^{p} \{-S_n^{(ij)}(\theta'_n)\}\{n^{1/2}(\hat{\theta}_{n,i}-\theta_{0,i})\} = n^{1/2} S_n^{(j)}(\theta_0),$$

$$\sum_{i=1}^{q} \{-V_n^{(ij)}(\phi'_n)\}\{n^{1/2}(\hat{\phi}_{n,i}-\phi_{0,i})\} = n^{1/2} V_n^{(j)}(\phi_0)$$

where

$$\theta'_n = \lambda_1\theta_0 + (1-\lambda_1)\hat{\theta}_n, \qquad 0<\lambda_1<1,$$

$$\phi'_n = \lambda_2\phi_0 + (1-\lambda_2)\hat{\phi}_n, \qquad 0<\lambda_2<1.$$

We now proceed in a manner similar to the proof of Theorem 6.1 to prove this result. □

THEOREM 8.2. *Under Assumptions* A1, A2, A3, A5, A6, B3, C1, C2, C3, C1′, C2′, C3′ *the limiting distribution of*

$$n[(S_n(\hat{\theta}_n) - V_n(\hat{\phi}_n)) - (\sigma^2(\theta_0) - \delta^2(\phi_0))]$$

is $N(0, \eta)$ *where*

$$\eta \triangleq 4\pi \int_{-\pi}^{\pi} [h(\lambda, \theta_0) - m(\lambda, \phi_0)]^2 f^2(\lambda)\, d\lambda$$

$$+\frac{\kappa_4}{\sigma^4}\left[\int_{-\pi}^{\pi} [h(\lambda, \theta_0) - m(\lambda, \phi_0)] f(\lambda)\, d\lambda\right]^2.$$

Also, $\eta \geqq (2+\kappa_4/\sigma^4)\Delta^2$ *where* $\Delta \triangleq \sigma^2(\theta_0) - \delta^2(\phi_0)$.

Proof. By arguments similar to those presented in the proof of Theorem 7.1 we may consider $n^{1/2}[(S_n(\theta_0) - V_n(\phi_0)) - E\{S_n(\theta_0) - V_n(\phi_0)\}]$. Theorem C.4 now implies the first part of the theorem.

For the second part we note that when Δ is constrained to equal $\int_{-\pi}^{\pi} (h(\lambda, \theta_0) - m(\lambda, \phi_0)) f(\lambda)\, d\lambda$ a calculus of variations argument shows that the smallest possible value of η is $(2+\kappa_4/\sigma^4)\Delta^2$. □

Remark 8.1. Note that even when $\sigma^2(\theta_0) = \sigma^2(\phi_0)$ (i.e., $\Delta = 0$) it may still be true that $\eta > 0$.

We next investigate the special case $h(\lambda, \theta_0) = m(\lambda, \phi)$ for all $\lambda \in [-\pi, \pi]$. This may be considered to be a suitable form of null hypothesis in structure choice problems (see [19] for a detailed discussion).

THEOREM 8.3. *Suppose Assumptions* A1, A2, A3, A4, B2, C1, C2, C1′, C2′ *hold and suppose that* Φ, Ξ *are nonsingular. Under the hypothesis that* $h(\lambda, \theta_0) - m(\lambda, \phi_0)$ *for all* $\lambda \in [-\pi, \pi]$ *the limiting distribution of* $n[(S_n(\hat{\theta}_n) - V_n(\hat{\phi}_n)) - (\sigma^2(\theta_0) - \delta^2(\phi_0))]$ *is the same as the distribution of* $-\frac{1}{2}x^T\Phi x + \frac{1}{2}y^T\Xi y$ *where* $[x^T y^T]^T$ *is distributed as*

$$N\left(0, \begin{bmatrix} A & B \\ B^T & C \end{bmatrix}\right)$$

where A, B *and* C *are defined in the statement of Theorem* 8.1. *The variance of the limiting distribution is*

$$\frac{1}{4}\sum_{i,j=1}^{p}\sum_{k,l=1}^{q} (C_{ij}C_{kl} + C_{ik}C_{jl} + C_{il}C_{jk})\Xi_{ij}\Xi_{kl}$$

$$-\frac{1}{2}\sum_{i,j=1}^{p}\sum_{k,l=1}^{q} (A_{ij}C_{kl} + B_{ij}B_{jl} + B_{il}B_{jk})\Phi_{ij}\Xi_{kl}$$

$$+\frac{1}{4}\sum_{i,j=1}\sum_{k,l=1} (A_{ij}A_{kl} + A_{ik}A_{jl} + A_{il}A_{jk})\Phi_{ij}\Phi_{kl}.$$

Proof. By the mean value theorem

$$S_n(\theta_0) = S_n(\hat{\theta}_n) + \tfrac{1}{2} \sum_{i,j=1}^{p} (\theta_{0,i} - \hat{\theta}_{n,i})(\theta_{0,j} - \hat{\theta}_{n,j}) S_n^{(ij)}(\theta_n^*)$$

where $\theta_n^* = \lambda_1 \theta_0 + (1-\lambda_1)\hat{\theta}_n$, $0 < \lambda_1 < 1$.

Similarly,

$$V_n(\phi_0) = V_n(\hat{\phi}_n) + \tfrac{1}{2} \sum_{i,j=1}^{q} (\phi_{0,i} - \hat{\phi}_{n,i})(\phi_{0,j} - \hat{\phi}_{n,j}) V_n^{(ij)}(\phi_n^*)$$

where $\phi_n^* = \lambda^2 \phi_0 + (1-\lambda_2)\hat{\phi}_n$, $0 < \lambda_2 < 1$.

The hypothesis that $h(\lambda, \theta_0) = m(\lambda, \theta_0)$ for all $\lambda \in [-\pi, \pi]$ immediately implies that $S_n(\theta_0) = V_n(\phi_0)$ and $\sigma^2(\theta_0) = \delta^2(\theta_0)$, thus

$$\begin{aligned} &n[(S_n(\hat{\theta}_n) - V_n(\hat{\phi}_n)) - (\sigma^2(\theta_0) - \delta^2(\phi_0))] \\ (8.1) \qquad &= \frac{n}{2} \sum_{k,l=1}^{q} (\phi_{0,k} - \hat{\phi}_{n,k})(\phi_{0,l} - \hat{\phi}_{n,l}) V_n^{(kl)}(\phi_n^*) \\ &\quad - \frac{n}{2} \sum_{i,j=1}^{p} (\theta_{0,i} - \hat{\theta}_{n,i})(\theta_{0,j} - \hat{\theta}_{n,j}) S_n^{(ij)}(\theta_n^*). \end{aligned}$$

But Theorem 8.1 now implies that the limiting distribution of the l.h.s. of equation (8.1) is as stated in the theorem. □

Let it be remarked that the theorem obviously applies in the case that $h(\cdot, \theta_0) = m(\cdot, \phi_0) = f(\cdot)^{-1}$, this being a classical null hypothesis. It is to be stressed, however, that it has not been assumed that $h(\cdot, \theta_0) = m(\cdot, \phi_0) = f(\cdot)^{-1}$. It is easily seen that it is possible to have $h(\cdot, \theta_0) = m(\cdot, \phi_0)$ without having $h(\cdot, \phi_0) = m(\cdot, \phi_0) = f(\cdot)^{-1}$.

9. Rapprochment with classical results. By considering assumptions additional to those made in §§ 2–9 we are able to recover many of the results of the classical theory of finite parameter models for purely nondeterministic time-series.

Consider any one of the Process Type 1, 2, 3, 2′ and 3′.

Additional Assumption 1: $\{x_t\}$ is weakly stationary and has an absolutely continuous spectral distribution. Again, the spectral density is denoted by $f(\cdot)$.

Additional Assumption 2: $\int_{-\pi}^{\pi} \log f(\lambda)\, d\lambda > -\infty$.

Introduce now the set $\{f(\cdot, \theta) | \theta \in \Theta\}$ which satisfies the following additional assumptions.

Additional Assumption 3: $f(\cdot) = f(\cdot, \theta_0)$ for a unique $\theta_0 \in \Theta$ (here two functions are considered equal if they differ at most on the set of λ-measure zero).

Additional Assumption 4: $f(\lambda, \theta) \geqq 0$ for all $(\lambda, \theta) \in Z$, $\int_{-\pi}^{\pi} f(\lambda, \theta)\, d\lambda < \infty$ and $\int_{-\pi}^{\pi} \log f(\lambda, \theta)\, d\lambda > -\infty$ for each $\theta \in \Theta$. Now define

$$m(\cdot, \theta) \triangleq \frac{\psi(\theta)}{2f(\cdot, \theta)} \quad \text{for each } \theta \in \Theta$$

where

$$\psi(\theta) \triangleq \exp \left\{ \frac{1}{2\pi} \int_{-\pi}^{\pi} \log 2\pi f(\lambda, \theta)\, d\lambda \right\}.$$

Also, let σ^2 denote $\psi(\theta_0)$. The next result if well known

THEOREM 9.1. *$\int_{-\pi}^{\pi} m(\lambda, \theta) f(\lambda)\, d\lambda$ is minimized at the single value of $\theta_0 \in \Theta$. (θ_0 is sometimes referred to as the "true value" and yields $f(\cdot, \theta_0) = f(\cdot)$ a.s.)*

Proof. For the proof see [12]. □

Introduce now the following additional Assumption.

Additional Assumption 5: *$m(\lambda, \theta)$ is continuous in $(\lambda, \theta) \in Z$ and Θ is a compact subset of a metric space $\mathcal{M}$ with distance $d(\cdot\,, \cdot)$.*

THEOREM 9.2. *Under Additional Assumptions* 1–5, *for Process Types* 1–3:

$$\lim_n d(\hat{\theta}_n, \theta_0) = 0$$

$$\lim_n S_n(\hat{\theta}_n) = \sigma^2(\theta_0), \quad \textit{for } \omega \in \Omega'.$$

For Process Types 2′ *and* 3′:

$$d(\hat{\theta}_n, \theta_0) \xrightarrow{\text{Prob}} 0,$$

$$S_n(\hat{\theta}_n) \xrightarrow{\text{Prob}} \sigma^2(\theta_0).$$

Proof. The proof is an immediate consequence of Theorems 4.1 and 5.2. □

In the next result we specialize to processes of the type 3′.

THEOREM 9.3. *Suppose Additional Assumptions* 1–5 *and Assumptions* A1, A2, A3, A4 *and* B2 *hold and that Assumptions* C1 *and* C2 *hold for $m(\lambda, \theta)$ replacing $h(\lambda, \theta)$. Let $W = (W_{ij})$ where*

$$W_{ij} \triangleq \frac{1}{4\pi} \int_{-\pi}^{\pi} \frac{m^{(i)}(\lambda, \theta_0) m^{(j)}(\lambda, \theta_0)}{m^2(\lambda, \theta_0)} \, d\lambda.$$

Now provided W is nonsingular $n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{\text{dist}} N(0, W^{-1})$.

Proof. The result follows from Theorem 6.1 and the following observations. Since $\int_{-\pi}^{\pi} \log m(\lambda, \theta)\, d\lambda = 0$ for $\theta \in \Theta$ and since for $\theta \in N_0$ this can be differentiated under the integral sign w.r.t. θ_j

$$\int_{-\pi}^{\pi} \frac{m^{(j)}(\lambda, \theta)}{m(\lambda, \theta)} \, d\lambda = 0$$

and

$$\int_{-\pi}^{\pi} \frac{m^{(ij)}(\lambda, \theta)}{m(\lambda, \theta)} \, d = \int_{-\pi}^{\pi} \frac{m^{(i)}(\lambda, \theta) m^{(j)}(\lambda, \theta)}{m^2(\lambda, \theta)} \, d\lambda \quad \text{for } \theta \in N_0$$

hence $\psi = 2\sigma^2 \Phi$ and the result now follows from Theorem 6.1. □

THEOREM 9.4. *Under the conditions required for the validity of Theorem* 9.3, *$n[\sigma^2(\hat{\theta}_n) - \sigma^2(\theta_0)]$ is in the limit distributed as $\frac{1}{2} x^T \Phi x$ where x is distributed as $N(0, W^{-1})$. Let it be remarked that $E\{\frac{1}{2} x^T \Phi x\} = \sigma^2 p$.*

Proof. The proof is an immediate consequence of Corollary 6.1 and Theorem 9.3.

THEOREM 9.5. *Under the conditions required for the validity of Theorem* 9.3 *$n[S_n(\theta_0) - S_n(\hat{\theta}_n)]$ is in the limit distributed as $\frac{1}{2} y^T \Phi y$ where y is distributed as $N(0, W^{-1})$. Let it be remarked that $E\{\frac{1}{2} y^T \Phi y\} = \sigma^2 p$.*

Proof. Immediate consequence of Corollary 6.2 and Theorem 9.3. □

THEOREM 9.6. *Suppose Additional Assumptions* 1–5 *and Assumptions* A1, A2, A3, A5, A6 *and* B3 *hold and that Assumptions* C1, C2 *and* C3 *hold for $m(\lambda, \theta)$ replacing $h(\lambda, \theta)$. Then if W is nonsingular*

$$n^{1/2}[S_n(\hat{\theta}_n) - \sigma^2(\theta_0)] \xrightarrow{\text{dist}} N(0, \delta)$$

where

$$\delta \triangleq \left(2+\frac{\kappa_4}{\sigma^4}\right)\sigma^4.$$

Proof. The proof is a consequence of Theorem 7.1. □

THEOREM 9.7. *Under the conditions required for the validity of Theorem* 9.6

$$n^{1/2}\begin{bmatrix} S_n(\hat{\theta}_n)-\sigma^2(\theta_0) \\ \hat{\theta}_n-\theta_0 \end{bmatrix} \xrightarrow{\text{dist}} N(0, B)$$

where

$$B=\begin{bmatrix} \delta & 0 \\ 0 & W^{-1} \end{bmatrix}.$$

Proof. The result follows from Theorem 7.2 and the following observations. As in the proof of Theorem 9.3 we note that

$$\int_{-\pi}^{\pi} \frac{m^{(j)}(\lambda, \theta_0)}{m(\lambda, \theta_0)}\, d\lambda = 0.$$

Hence $\nu_j=0$ for $j=1,\cdots,p$ since by assumption $f(\lambda)$ is proportional to $1/m(\lambda, \theta_0)$. During the proof of Theorem 9.3 it was also shown that $\psi=2\sigma^2\Phi$ and the result now follows by Theorem 7.2. □

10. Conclusions. This paper has developed results pertaining to the estimation of the parameters in optimal interpolators when the class of interpolators is restricted. The practical impact of the results is that they allow us to establish asymptotically valid confidence regions for parameters under weak assumptions on the system. Results pertaining to the case where two interpolators of different structure are fitted to the one piece of data have also been presented. These results form a theoretical basis for interpolator structure choice.

Appendix A. Consider the following type 3 process:

$$x_t=\varepsilon_t+\varepsilon_{t-1}$$

where $\{\varepsilon_t\}$ is an i.i.d. sequence, $E(\varepsilon_t)=0$, $E(\varepsilon_t^2)=1$ and $E(\varepsilon_t^4)=\mu_4<\infty$.

The best one-step ahead autoregressive predictor of order 1 is

$$\hat{x}_t=\frac{\gamma_1}{\gamma_0}x_{t-1} = \tfrac{1}{2}x_{t-1}.$$

Cascading $k>1$ such predictors we obtain

$$x_t=\frac{x_{t-k}}{2^k}.$$

However, the appropriate mean square k-step-ahead predictor error is $\sigma^2(\theta)=E\{(x_t-\theta x_{t-k})^2\}=\gamma_0(1+\theta^2)-\gamma_k(2\theta)$ which is minimized by $\theta_0=0$ for $k>1$, giving the optimal restricted complexity predictor as $\hat{x}_t=0$ for $k>1$.

Appendix B.

THEOREM B.1. *Under the conditions of Theorem* 6.1 *and using the notation introduced in the proof of that theorem*

$$S_n^{(ij)}(\theta_n^*) \xrightarrow{\text{Prob}} \phi_{ij}.$$

Proof. Since Z is a compact subset of $\mathcal{R}^{p+1}$ and $h^{(ij)}(\lambda, \theta)$ is continuous on this set, $h^{(ij)}(\lambda, \theta)$ is also uniformly continuous. Hence given $\varepsilon > 0$ a neighborhood N_1 and θ_0 can be found such that $N_1 \subset N_0$ and for which $|h^{(ij)}(\lambda, \theta) - h^{(ij)}(\lambda, \theta_0)| < \varepsilon$ uniformly in $\theta \in N_1$, $\lambda \in [-\pi, \pi]$.

Now since $\theta_n^* \xrightarrow{\text{Prob}} \theta_0$ the limit result may be derived on the assumption that $\theta_n^* \in N_1$. Now for $\theta_n^* \in N_1$:

$$\begin{aligned}|S_n^{(ij)}(\theta_n^*) - S_n^{(ij)}(\theta_0)| &= \left| \int_{-\pi}^{\pi} I_n(\lambda)\{h^{(ij)}(\lambda, \theta_n^*) - h^{(ij)}(\lambda, \theta_0)\}\, d\lambda \right| \\ &\leqq \varepsilon C_0.\end{aligned}$$

Now $C_0 \xrightarrow{\text{Prob}} \gamma_0$ and ε is arbitrarily small so that

$$|S_n^{(ij)}(\theta_n^*) - S_n^{(ij)}(\theta_0)| \xrightarrow{\text{Prob}} 0. \tag{B.1}$$

Let $q_M'(\lambda, \theta)$ denote the Cesaro sum of the Fourier series of $h^{(ij)}(\lambda, \theta)$ taken to M terms i.e.

$$q_M'(\lambda, \theta) = \sum_{s=-M}^{M} \alpha_s^{(ij)}(\theta)\left(1 - \frac{|s|}{M}\right) e^{is\lambda}.$$

(The derivatives exist by virtue of assumptions C1 and C2 which imply (2.5) can be differentiated under the integral sign.) Also let

$$J_1(\theta_0) = \left| S_n^{(ij)}(\theta_0) - \int_{-\pi}^{\pi} q_M'(\lambda, \theta_0) I_n(\lambda)\, d\lambda \right|,$$

$$J_2(\theta_0) = \left| \int_{-\pi}^{\pi} q_M'(\lambda, \theta_0) I_n(\lambda)\, d\lambda - \int_{-\pi}^{\pi} q_M'(\lambda, \theta_0) f(\lambda)\, d\lambda \right|,$$

$$J_3(\theta_0) = \left| \int_{-\pi}^{\pi} q_M(\lambda, \theta_0) f(\lambda)\, d\lambda - \phi_{ij} \right|.$$

Note $|S_n^{(ij)}(\theta_0) - \phi_{ij}| \leqq J_1(\theta_0) + J_2(\theta_0) + J_3(\theta_0)$.

Now given $\varepsilon_1 > 0$ we may fix M so large that $|h^{(ij)}(\lambda, \theta) - q_M(\lambda, \theta)| < \varepsilon_1$ uniformly in $(\lambda, \theta) \in Z$ since the Cesaro sum converges uniformly in $(\lambda, \theta) \in Z$.

$$\begin{aligned}J_1(\theta_0) &\leqq \int_{\pi}^{\pi} |h^{(ij)}(\lambda, \theta_0) - q_M'(\lambda, \theta_0)| I_n(\lambda)\, d\lambda \\ &\leqq \varepsilon_1 C_0.\end{aligned}$$

Similarly $J_3(\theta_0) \leqq \varepsilon_1 \gamma_0$,

$$J_2(\theta_0) = \left| \sum_{s=-M}^{M} \alpha_s^{ij}(\theta_0)\left(1 - \frac{|s|}{M}\right)(C_s - \gamma_s) \right|.$$

But $h^{(ij)}(\lambda, \theta)$ has been assumed to be continuous on a compact set, so $\max_{(\lambda,\theta)\in Z} |h^{(ij)}(\lambda, \theta)| \leqq k < \infty$. Thus independently of $\theta \in \Theta$, $|\alpha_s^{(ij)}(\theta)| \leqq k$ and

$$J_2(\theta_0) \leqq k \sum_{s=-M}^{M} |C_s - \gamma_s|.$$

But $C_s \xrightarrow{\text{Prob}} \gamma_s$, so that $J_2(\theta_0) \xrightarrow{\text{Prob}} 0$ (M fixed).

Hence combining these results $J_1(\theta_0)$, $J_2(\theta_0)$, $J_3(\theta)$ for given $\varepsilon > 0$, $\eta > 0$ we can fix M sufficiently large and find an $n(>M)$ sufficiently large that

$$P\{|S_n^{(ij)}(\theta_0) - \phi_{ij}| > \varepsilon\} < \frac{\eta}{2}.$$

Combining this with (B.1) yields the desired result. □

Appendix C.

LEMMA C.1. *Under assumptions* A2, A3, A5 *and* A6

(a) $E\{\varepsilon_i\} = 0$, $E\{\varepsilon_i \varepsilon_j\} = \sigma^2 \delta_{ij}$.

(b) $E\{\varepsilon_i \varepsilon_j \varepsilon_k \varepsilon_l\} = 0$ *for* $i > j > k > l$.

(c) $E\{\varepsilon_i^3 \varepsilon_k\} = 0$ *for* $i \neq k$.

(d) $E\{\varepsilon_i^2 \varepsilon_j \varepsilon_k\} = 0$ *for* $i \neq j$, $i \neq k$ *and* $j \neq k$.

(e) $E\{\varepsilon_i^2 \varepsilon_j^2\} = \sigma^4$ *for* $i \neq j$.

Proof. The proof is straightforward using properties of conditional expectations.

THEOREM C.1. *Under assumptions* A1, *to* A6 *we have the following implications*

(a) $\sum_u u|l_u| < \infty$ *implies* $\sum u l_u^2 < \infty$.

(b) $\sum_u u l_u^2 < \infty$ *implies* $\sum_u \gamma_t^2 < \infty$.

(c) $\sum_u u|l_u| < \infty$ *implies* $\sum_t |t\gamma_t| < \infty$.

(d) $\sum_t |t\gamma_t| < \infty$ *implies* $\sum_t |\gamma_t| < \infty$.

(e) $\sum_t |\gamma_t| < \infty$ *implies* $\sum_t \gamma_t^2 < \infty$.

(f) $\sum_t \gamma_t^2 < \infty$ *implies* $f(\lambda) = (1/2\pi) \sum \gamma_n e^{in\lambda}$ *in mean-square where*

$$\gamma_n \triangleq \int_{-\pi}^{\pi} f(\lambda)\, e^{-in\lambda}\, d\lambda.$$

(g) $\sum_t |\gamma_t| < \infty$ *implies* $f(\lambda)$ *is continuous and*

$$f(\lambda) = \frac{1}{2\pi} \sum_n \gamma_n e^{in\lambda}$$

pointwise.

(h) $\sum_t |t\gamma_t| < \infty$ *implies* $f'(\lambda)$ *is continuous for* $\lambda \in [-\pi, \pi]$.

Proof. (a) Suppose $\sum u|l_u| < \infty$ then clearly $|l_u| < 1$ for $u >$ some N but then $l_u^2 < |l_u|$. Hence $\sum_u u l_u^2 < \infty$.

(b) Suppose $\sum_u u l_u^2 < \infty$ then

$$\begin{aligned}\sum_t |\gamma_t|^2 &= \sum_t \left|\sum_u l_u l_{u+t}\right|^2 \sigma^4 \\ &\leqq \sum_t \left(\sum_u l_u^2\right)\left(\sum_{v=0}^{\infty} l_{v+t}^2\right)\sigma^4 \\ &= \left(\sum_u l_u^2\right) \sum_t \sum_{v=0}^{\infty} l_{v+t}^2 \sigma^4 = \left(\sum_u l_u^2\right)\left(\sum_t t l_t^2\right)\sigma^4 < \infty.\end{aligned}$$

(c) Suppose $\sum_u u|l_u| < \infty$. Now for $t > 0$

$$\sum_{u=1}^{\infty} u|l_{u+t}| \leqq \sum_{u=1}^{\infty} (u+t)|l_{u+t}| \leqq \sum_{u=1}^{\infty} u|l_u| < \infty.$$

Hence

$$\sum_t t|\gamma_t| = 2 \sum_{t=1}^{\infty} t\left|\sum_u l_u l_{u+t}\right| \sigma^2$$

$$\leqq 2 \sum_u |l_u| \sum_{t=1}^{\infty} t|l_t| < \infty.$$

(d) and (e) are elementary.

(f) By the Riesz–Fischer theorem $\sum_t \gamma_t^2 < \infty \Rightarrow f(\lambda) = \sum_n \gamma_n e^{in\lambda}$ in mean-square.

(g) By the Weistrass M-test $\sum_t |\gamma_t| < \infty$ implies that $\sum_n \gamma_n e^{in\lambda}$ converges uniformly. Hence $f(\lambda)$ is continuous and $f(\lambda) = \sum_t \gamma_n e^{in\lambda}$ pointwise.

(h) Suppose $\sum_t |t\gamma_t| < \infty$. Now if we define

$$g(z) \triangleq \frac{1}{2\pi} \sum_s \gamma_s e^{izs} - \frac{1}{2\pi} \sum_s \gamma_s e^{-i\pi s}$$

then we see that

$$g(z) = \int_{-\pi}^{z} k(\lambda)\, d\lambda \quad \text{where } k(\lambda) \triangleq \frac{1}{2\pi} \sum_s (is)\gamma_s e^{i\lambda s}.$$

Clearly $k(\lambda)$ is continuous on $[-\pi, \pi]$. But $g(z) = f(z) + \text{constant}$, $f'(z) = g'(z)$ so that $f'(z) = k(z)$. Consequently $f'(z)$ is continuous for $\lambda \in [-\pi, \pi]$. □

RESULT C.1. *Under assumptions* A1, A2, A3, A5, A6, B3, *let* $\kappa_4 \triangleq \mu_4 - 3\sigma^4$ (κ_4 *is the fourth cumulant of* ε_t). *Then*

$$\text{(A)} \qquad E\{x_i x_j x_k x_l\} = \gamma_{i-j}\gamma_{k-l} + \gamma_{i-k}\gamma_{j-l} + \gamma_{i-l}\gamma_{j-k} + \kappa_4 \sum_a l_a l_{a+(j-i)} l_{a+(k-i)} l_{a+(l-i)}$$

$$\text{(B)} \qquad n \operatorname{cov}(c_t, c_r) = \frac{1}{n} \sum_{i=1}^{n-t} \sum_{j=1}^{n-r} \{\gamma_{i-j}\gamma_{i+t-j-r} + \gamma_{i-j-r}\gamma_{i+t-j} + \kappa_4 \sum_a l_a l_{a+t} l_{a+(j-i)} l_{a+j+r-i}\}$$

$$\text{(C)} \qquad |n \operatorname{cov}(c_t, c_r)| \leqq \sum_v \{|\gamma_v \gamma_{v+t-r}| + |\gamma_{v-r}\gamma_{v+t}| + |\kappa_4| \sum_a |l_a l_{a+r} l_{a-v} l_{a+r-v}|\}.$$

Proof. (A): the proof is almost identical to pp. 466–467 of Anderson [20]. (B) follows by calculation from (A), and (C) is an immediate consequence of (B). □

RESULT C.2. *Suppose* $x_t = \sum_u l_u \varepsilon_{t-u}$ *with* $E\{\varepsilon_t\} = 0$, $E\{\varepsilon_n \varepsilon_m\} = \sigma^2 \delta_{nm}$, $E\{\varepsilon_t \varepsilon_s \varepsilon_q \varepsilon_r\} = 0$, $t \neq s$, $t \neq q$, $t \neq r$, $E\{\varepsilon^2 \varepsilon^2\} = \sigma^4 r \neq s$, $E\{\varepsilon_t^4\} = \mu_4$ *and* $\sum_t |\gamma_t| < \infty$.

$$\lim_n n \operatorname{cov}(c_t, c_r) = \sum_{s=-\infty}^{\infty} \{\gamma_s \gamma_{s-r+t} + \gamma_{s-r}\gamma_{s+t}\} + \frac{\kappa_4}{\sigma^4} \gamma_t \gamma_r.$$

Proof. The proof is essentially as on pp. 467–468 of Anderson [20]. □

THEOREM C.2. *Suppose* $x_t = \sum_u l_u \varepsilon_{t-u}$ *with* $E\{\varepsilon_t\} = 0$, $E\{\varepsilon_n \varepsilon_m\} = \sigma^2 \delta_{nm}$, $E\{\varepsilon_t \varepsilon_s \varepsilon_q \varepsilon_r\} = 0$, $t \neq s$, $t \neq q$, $t \neq r$, $E\{\varepsilon_t^2 \varepsilon_s^2\} = \sigma^4$, $t \neq s$, $E\{\varepsilon_t^4\} = \mu_4$ *and* $\sum_t |\gamma_t| < \infty$. *Let*

$W_1(\lambda)$, $W_2(\lambda)$ be even and bounded with at most a finite number of discontinuities for $\lambda \in [-\pi, \pi]$. Now let

$$T_m = \sum_{s=-m}^{m} \delta_s c_s, \qquad V_m = \sum_{r=-m}^{m} \beta_r c_r$$

where

$$\delta_s \triangleq \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i\lambda s} W_1(\lambda)\, d\lambda, \qquad \beta_r = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i\lambda r} W_2(\lambda)\, d\lambda$$

and

$$a \triangleq n^{1/2}[T_m - E\{T_m\}], \qquad b \triangleq n^{1/2}[V_m - E\{V_m\}];$$

then

$$\lim_m \lim_n E\{ab\} = 4\pi \int_{-\pi}^{\pi} W_1(\lambda) W_2(\lambda) f^2(\lambda)\, d\lambda$$
$$+ \frac{\kappa_4}{\sigma^4} \cdot \int_{-\pi}^{\pi} W_1(\lambda) f(\lambda)\, d\lambda \cdot \int_{-\pi}^{\pi} W_2(\lambda) f(\lambda)\, d\lambda.$$

Proof. From the definition of a and b

$$a = n^{1/2} \sum_{t=-m}^{m} \delta_t [c_t - E\{c_t\}],$$

$$b = n^{1/2} \sum_{r=-m}^{m} \beta_r [c_r - E\{c_r\}]$$

hence

$$E\{ab\} = \sum_{t=-m}^{m} \sum_{r=-m}^{m} \delta_t \beta_r n \operatorname{cov}(c_t, c_r).$$

By Result C.2

$$\lim_n E\{ab\} = \sum_{r=-m}^{m} \sum_{t=-m}^{m} \delta_t \beta_r \left[\sum_s \{\gamma_s \gamma_{s-t+r} + \gamma_{s+r}\gamma_{s-t}\} + \frac{\kappa_4}{\sigma^4} \gamma_t \gamma_r \right].$$

Now since $\beta_{-r} = \beta_r$

$$\lim_n E\{ab\} = 2 \sum_{t=-m}^{m} \sum_{r=-m}^{m} \delta_t \beta_r \sum_s \gamma_s \gamma_{s-t+r}$$
$$+ \frac{\kappa_4}{\sigma^4} \sum_s \sum_{t=-m}^{m} \sum_{r=-m}^{m} \delta_t \beta_r \gamma_t \gamma_r.$$

By Parseval's theorem

$$\lim_m \lim_n E\{ab\} = 4\pi \int_{-\pi}^{\pi} W_1(\lambda) W_2(\lambda) f^2(\lambda)\, d\lambda$$
$$+ \frac{\kappa_4}{\sigma^4} \int_{-\pi}^{\pi} W_1(\lambda) f(\lambda)\, d\lambda \cdot \int_{-\pi}^{\pi} W_2(\lambda) f(\lambda)\, d\lambda. \quad \square$$

THEOREM C.3. *Under Assumptions* A1, A2, A3, A5, A6 *and supposing* $\sum_t |\gamma_t| < \infty$ (*which is implied by Assumption* B3 (*see Theorem* C.1)) $n^{1/2} c_t$, $0 \leqq t \leqq m$, *are asymptotically normally distributed.*

Proof. The proof is in a manner similar to Hannan and Heyde [18], pp. 2062–2063. □

THEOREM C.4. *Suppose Assumptions* A1, A2, A3, A5, A6 *and* B3 *hold. Introduce now* $g_1(\lambda), \cdots, g_q(\lambda)$ *where each* $g_i(\lambda)$ *is even and for which*

$$\sum_s |\alpha_s^i| < \infty; \qquad \alpha_n^i \triangleq \frac{1}{2\pi}\int_{-\pi}^{\pi} g_i(\lambda)\, e^{-in\lambda}\, d\lambda.$$

Now define

$$S_n^i \triangleq \int_{-\pi}^{\pi} g_i(\lambda) I_n(\lambda)\, d\lambda = \sum_{t=-n+1}^{n-1} \alpha_t^i c_t.$$

Then $n^{1/2}\{[S_n^1, \cdots, S_n^q]^T - E[S_n^1, \cdots, S_n^q]^T\}$ *is in the limit distributed as* $N(0, Z)$ *where* $Z = (z_{ij})$ *and*

$$z_{ij} = 4\pi \int_{-\pi}^{\pi} g_i(\lambda) g_j(\lambda) f^2(\lambda)\, d\lambda + \frac{\kappa_4}{\sigma^4} \cdot \int_{-\pi}^{\pi} g_i(\lambda) f(\lambda)\, d\lambda \cdot \int_{-\pi}^{\pi} g_j(\lambda) f(\lambda)\, d\lambda.$$

Proof. Now

$$n^{1/2}[S_n^i - E\{S_n^i\}] = n^{1/2} \sum_{t=-n+1}^{n-1} \alpha_t^i [c_t - E\{c_t\}].$$

Now define

$$Z_{mn}^j \triangleq n^{1/2} \sum_{|t| \leqq m} \alpha_t^j [c_t - E\{c_t\}],$$

$$R_{mn}^j = n^{1/2} \sum_{m<|t|<n} \alpha_t^j [c_t - E\{c_t\}], \qquad m < n-1.$$

By Theorem C.3 the limiting distribution of $n^{1/2}[c_t - E\{c_t\}]$ $0 \leqq t \leqq m$ is normal. Also, from Theorem C.2

$$\lim_m \lim_n \operatorname{cov}(Z_{mn}^j, Z_{mn}^k) = 4\pi \int_{-\pi}^{\pi} g_j(\lambda) g_k(\lambda) f^2(\lambda)\, d\lambda$$
$$+ \frac{\kappa_4}{\sigma^4} \cdot \int_{-\pi}^{\pi} g_j(\lambda) f(\lambda)\, d\lambda \cdot \int_{-\pi}^{\pi} g_j(\lambda) f(\lambda)\, d\lambda.$$

Also

$$E|R_{mn}^i|^2 \leqq 4 \sum_{t=m+1}^{\infty} \sum_{r=m+1}^{\infty} |\alpha_t^i \alpha_r^j|\, |n \operatorname{cov}(c_t, c_r)|$$
$$\leqq 4 \sum_{t=m+1}^{\infty} \sum_{r=m+1}^{\infty} |\alpha_t^i \alpha_r^j| \sum_v \left\{ |\gamma_{v+r}\gamma_{v+t}| + |\gamma_{v-r}\gamma_{v+t}| \right.$$
$$\left. + \frac{|\kappa_4|}{\sigma^4} \sum_a |l_a l_{a+t} l_{a-v} l_{a+r-v}| \right\} \quad \text{by Result C.1.}$$

Now

$$\sum_{t=m+1}^{\infty} \sum_{r=m+1}^{\infty} |\alpha_t^i \alpha_r^j| \sum_v |\gamma_{v+r}\gamma_{v+t}| \leqq \sum_{t=m+1}^{\infty} \sum_{r=m+1}^{\infty} |\alpha_t^i \alpha_r^j| \sum_v \gamma_v^2$$
$$= \sum_{t=m+1}^{\infty} |\alpha_t^i| \sum_{r=m+1}^{\infty} |\alpha_r^j| \sum_v \gamma_v^2.$$

Also

$$\sum_a |l_a l_{a+t} l_{a-v} l_{a+r-v}| \leqq \left(\sum_a l_a^2 l_{a+t}^2\right)^{1/2} \left(\sum_b l_{b-v}^2 l_{b+r-v}^2\right)^{1/2}$$

$$\leqq \sum_a l_a^4.$$

Consequently $\lim_m E|R^j_{mn}|^2 = 0$. Therefore by a result of Diananda [20] $Z^j_{mn} + R^j_{mn}$, $j = 1, \cdots, p$ each have a limiting distribution which is, respectively, $N(0, z_{jj})$.

Similarly, it can be shown that the limiting distribution of $n^{1/2} \sum_{j=1}^q \beta_j[S^j_n - E\{S^j_n\}]$, where $\beta_1, \cdots, \beta_q$ are arbitrary constants, is $N(0, \sum_{i,j=1}^q \beta_i\beta_j z_{ij})$. By the Cramer–Wold [12] theorem the limiting joint distributions of $n^{1/2}[S^j_n - E\{S^j_n\}]$, $1 \leqq j \leqq q$ is $N(0, Z)$. □

RESULT C.3. *Suppose $\{x_t\}$ is weakly stationary with absolutely continuous spectral distribution and that $f(\lambda)$ has a derivative which is bounded for $-\pi \leqq \lambda \leqq \pi$. Suppose also that $W(\lambda)$ is any bounded even function with at most a finite number of discontinuities then*

$$E\left\{\int_{-\pi}^{\pi} I_n(\lambda) W(\lambda)\, d\lambda\right\} = \int_{-\pi}^{\pi} f(\lambda) W(\lambda)\, d\lambda + O\left(\frac{\log n}{n}\right).$$

Proof. The result is stated and proved by Granander and Rosenblatt [23]. Their proof depends essentially on a result due to Fejer [21].

REFERENCES

[1] L. LJUNG, *On consistency and identifiability*, Mathematical Programming Studies No. 5, North-Holland, Amsterdam, 1976.

[2] P. V. KABAILA AND G. C. GOODWIN, *On the large sample properties of a least squares procedure*, Tech. Report, University of Newcastle, New South Wales, Australia, January 1977.

[3] L. LJUNG, *Prediction error identification methods*, Report LiTH-ISY-I-0139, University of Linkoping, 1977.

[4] P. E. CAINES, *The identification of stochastic process representations using prediction error methods*, Tech. Report, Division of Applied Sciences, Harvard University, Boston, MA, 1977.

[5] ———, *Stationary linear and nonlinear system identification and predictor set completeness*, IEEE Transactions on Automatic Control, 1978, to appear.

[6] B. D. O. ANDERSON, J. B. MOORE AND R. M. HAWKES, *Model approximation via prediction error methods*, Ibid., 1978, to appear.

[7] H. B. MANN AND A. WALD, *On the statistical treatment of linear stochastic difference equations*, Econometrica, 11 (1943), pp. 173–220.

[8] P. WHITTLE, *Prediction and Regulation by Linear Least Squares Methods*, the English Universities Press, 1963.

[9] ———, *Estimation and information in stationary time series*, Ark. Mat. 2 (1953), pp. 423–434.

[10] ———, *Some recent contributions to the theory of stationary processes*, Appendix 2, A Study in the Analysis of Stationary Time Series, H. Wold, ed., 2nd Ed. Almquist and Wiksell, Stockholm, 1954.

[11] ———, *Gaussian estimation in time series*, Bull ISI, 39 (1962), pp. 105–129.

[12] A. M. WALKER, *Asymptotic properties of least squares estimates of parameters of the spectrum of a stationary non-deterministic time series*, Aust. Math. Soc. J., 4 (1964), pp. 363–384.

[13] E. J. HANNAN, *The asymptotic theory of linear time series models*, J. Appl. Prob., 10 (1973), pp. 130–145.

[14] J. L. DOOB, *Stochastic Processes*, John Wiley and Sons, New York, 1953.

[15] P. L. BUTZER AND R. J. NESSEL, *Fourier Analysis and Approximation*, Birkhäuser Verlag, Basel, 1971.

[16] H. CRAMER, *Mathematical Methods of Statistics*, Princeton University Press, 1946.

[17] E. PARZEN, *Stochastic Processes*, Holden Day, San Francisco, 1970.

[18] E. J. HANNAN AND C. C. HEYDE, *On limit theorems for quadratic forms of discrete time-series*, Ann. Math. Statist., 43 (1972), pp. 2058–2068.

[19] P. V. KABAILA, *Ultimate objective in estimation and structure choice*, Ph.D. dissertation, University of Newcastle, New South Wales, Australia, 1978.

[20] T. W. ANDERSON, *The Statistical Analysis of Time Series*, John Wiley, New York, 1971.

[21] L. FEJER, *Lebesguesche Konstantenund divergente fourier-reihen*, J. Reine. Angew. Math., 138 (1910), pp. 54–59.

[22] C. W. BURRILL, *Measure, Integration and Probability*, McGraw-Hill, New York, 1972.

[23] U. GRENANDER AND M. ROSENBLATT, *Statistical spectral analysis of time series arising from stationary stochastic process*, Ann. Math. Statist., 24 (1953), pp. 537–539.

[24] U. GRENANDER, *On the estimation of regression coefficients in the case of an autocorrelated disturbance*, Ibid., 25 (1954), pp. 252–272.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 2, March 1980

0363-0129/80/1802-0004 $01.00/0

BOUNDARY CONTROL FOR THE HEAT EQUATION WITH STEADY-STATE TARGETS*

E. J. P. GEORG SCHMIDT†

Abstract. Let Ω be a given domain in R^n, and $u(x, t)$ denote the temperature distribution of Ω at time t. The evolution of $u(x, t)$ is governed by an initial boundary value problem for the heat equation; the boundary value can be regarded as a control function. Within this context, given initial and target temperature distributions $u_0(x)$ and $u_1(x)$, traditional questions of control theory—controllability, optimal controllability and the characterization of optimal controls—have been extensively studied. Here these topics are considered with particular reference to steady state distributions, that is solutions of the heat equation which do not depend on time. We show that any $u_0(x)$ can be controlled exactly to any steady state target $u_1(x)$, and that the corresponding time optimal problem (with bounded controls) has as solution a "bang-bang" control. For controlling from $c_0v(x)$ to $c_1v(x)$ (where $v(x)$ is a steady state with boundary value $g(x)$, and c_0 and c_1 are constants) the restricted class of controls of the form $h(t)g(x)$ is considered. Controllability results (including a necessary and sufficient condition for exact controllability within that restricted class of controls) are proved. Moreover, we show that a certain time-optimal problem, in which the target is a neighborhood of $c_1v(x)$, has a unique solution $h(t)g(x)$ with $h(t)$ "bang-bang". These results apply in particular to the problem of controlling from c_0 to c_1 using controls dependent on time alone.

1. Introduction. Let Ω be a bounded domain in R^n whose boundary $\partial\Omega$ is a C^∞ manifold. Let Δ denote the Laplacian operation on R^n, $\partial/\partial\nu$ denote differentiation with respect to the outward pointing normal ν to $\partial\Omega$, a be a nonnegative constant, and $B'' = ''a(\partial/\partial\nu)+1$. We consider the following initial boundary value problem:

$$
\begin{aligned}
&\frac{\partial u}{\partial t}(x, t) = \Delta u(x, t) && \text{for } x \in \Omega, && t \in (0, \infty),\\
(1)\qquad &Bu(x, t) = f(x, t) && \text{for } x \in \partial\Omega, && t \in (0, \infty),\\
&u(x, 0) = u_0(x) && \text{for } x \in \Omega.
\end{aligned}
$$

It can be shown that, given u_0 in $H = L_2(\Omega)$ and f in $L_\infty = L_\infty(\partial\Omega \times (0, \infty))$, (1) has a unique solution u in a certain weak sense to be specified later. Moreover $u(\cdot, t)$ lies in H for each $t > 0$.

Let $u_0 \in H$ be given, and t be a fixed positive time. For any subclass L of L_∞ we define

$$R_t(u_0; L) = \{u(\cdot, t)\text{: there exists } f \in L \text{ with } u \text{ the corresponding solution of (1)}\}.$$

Controllability involves the study of these sets. It is well known that $R_t(u_0, L_\infty)$ is dense in (but not equal to) H, i.e., that the system (1) is *approximately* controllable. (See, for example, [9]). Exact controllability involves identifying elements of $R_t(u_0; L)$. Certain rather stringent sufficient conditions for u_1 to lie in $R_t(u_0; L_\infty)$ have been developed by Fattorini and Russell (see [3], [4] and [11]); in particular it follows that $0 \in R_t(u_0, L_\infty)$ for all $t > 0$, a property known as *null controllability*.

We introduce a class of temperature distributions on Ω which generalize the constant temperature distributions, and share many of their desirable properties.

* Received by the editors September 13, 1978 and in revised form March 12, 1979.

† Department of Mathematics, McGill University, Montreal, Quebec H3C 3GI, Canada. This research was supported in part by the National Research Council of Canada Grant A7271 and by an Alexander von Humboldt Research Fellowship.

A steady state for the system (1) is a weak solution $v \in H$ of

$$(2)\qquad \begin{aligned} \Delta v(x) &= 0, && \text{for } x \in \Omega, \\ Bv(x) &= g(x), && \text{for } x \in \partial\Omega, \end{aligned}$$

where $g \in L_\infty(\partial\Omega)$. From null controllability it follows that each steady state v is "reachable", i.e., $v \in R_t(u_0; L_\infty)$ for any $u_0 \in H$ and any $t > 0$. Steady states have other desirable features too.

To actually find a control which takes one from a given initial state u_0 to an arbitrary reachable target state u_1 is extremely difficult. However when u_1 is a steady state with $Bu_1 = g$, the control function $f(x, t) = g(x)$ yields a solution u of (1) with the property that $u(\cdot, t)$ converges to u_1 as $t \to \infty$. Thus, given $\varepsilon > 0$, one can explicitly find a control such that for t sufficiently large $\|u(\cdot, t) - u_1\| < \varepsilon$ (where $\|\cdot\|$ denotes the norm in H). This suggests introducing the class of control functions

$$L_\infty^g = \{f(x, t) = h(t)g(x) : h \in L_\infty(0, \infty)\}.$$

Such control functions were already studied by Glashoff and Weck in [6], but not in connection with steady states. We obtain a characterization of the closure of $R_t(0; L_\infty^g)$ in H. We then also use a result of Galchuk [5] to derive a necessary and sufficient condition for $c_1 v$ to lie in $R_t(c_0 v; L_{m,M}^g)$, where $L_{m,M}^g = \{f = gh \in L_\infty^g : m \leqq h(t) \leqq M \text{ a.e.}\}$. These results apply in particular when $v \equiv 1$, in which case the control function depends on t alone, i.e., the control of the temperature of the "body" Ω is by means of the ambient temperature $h(t)$, and the aim is to cool (or heat) the body from c_0 to c_1. That situation originally motivated this paper.

The paper finally deals with optimal problems related to the above mentioned controllability results, with special reference to the bang-bang property of optimal controls.

2. Some facts about the heat equation. We shall need some facts concerning the spectral theory of the Laplacian. A self-adjoint operator L can be defined in H as the Laplacian acting on a suitable domain of functions satisfying $Bu = 0$ on $\partial\Omega$. It is well known (see Agmon [1]) that L has a complete, orthonormal system of eigenfunctions $\{\varphi_k\}_{k \in N}$ (N the natural numbers) corresponding to negative eigenvalues $\{-\lambda_k\}_{k \in N}^\infty$:

$$(3)\qquad \Delta\varphi_k = -\lambda_k \varphi_k, \qquad B\varphi_k = 0.$$

For the eigenvalues one has the asymptotic estimate

$$(4)\qquad \lambda_k \sim Ck^{2/n} \qquad (C \text{ a constant}).$$

The eigenfunctions lie in $C^\infty(\bar{\Omega})$, and, letting D^r denote an arbitrary partial derivative of order r,

$$(5)\qquad |D^r \varphi_k(x)| \leqq C_r \lambda_k^{m_r},$$

where C_r and m_r are positive constants.

Following Fattorini [2] who treated the case $a = 0$, we shall say that a function $u(x, t)$ which belongs to $L_2(\Omega \times (0, T))$ for each $T > 0$, is a weak solution of (1) if

$$(6)\qquad \begin{aligned} &\int_\Omega \int_0^T u(x, t)\left[\frac{\partial\varphi}{\partial t}(x, t) + \Delta\varphi(x, t)\right] dx\, dt + \int_\Omega u_0(x)\varphi(x, 0)\, dx \\ &\qquad + \int_{\partial\Omega} \int_0^T f(x, t)\varphi^\partial(x, t)\, dS_x\, dt = 0, \end{aligned}$$

where dS_x denotes an element of surface area of $\partial\Omega$, φ belongs to the space of test functions

$$D_T=\{\varphi\in C^\infty(\partial\Omega\times[0,T]):\varphi(x,T)\equiv 0, B\varphi(x,t)\equiv 0\}$$

and

$$\varphi^\partial(x,t)=\begin{cases} a^{-1}\varphi(x,t) & \text{for } x\in\partial\Omega,\ t>0, \quad \text{if } a>0,\\ -\dfrac{\partial\varphi}{\partial\nu}(x,t) & \text{for } x\in\partial\Omega,\ t>0, \quad \text{if } a=0.\end{cases} \tag{7}$$

Combining Fattorini's results with those of Glashoff and Weck [6], it is not difficult to prove the following theorem in which the main facts relevant to this paper are summarized.

THEOREM 2.1. *Given $u_0\in H$ and $f\in L_\infty$ there exists a unique weak solution u to* (1). *That solution belongs to $C^\infty(\Omega\times(0,\infty))$ as well as to $L_2(\Omega\times(0,T))$ for each $T>0$. Moreover for each $t\geqq 0$ one has*

$$u(\cdot\,,t)=V_tu_0+S_tf, \tag{8}$$

where

(a) *$\{V_t\}_{t\geqq 0}$ is a strongly continuous semigroup of bounded linear operators on H;*

(b) *S_t: $L_\infty\to H$ is continuous from the weak*-topology on L_∞ to H;*

(c)

$$V_tu_0=\sum_{k\in N}e^{-\lambda_k t}(u_0,\varphi_k)\varphi_k, \tag{9}$$

with $(\cdot\,,\cdot)$ the inner product on H;

(d)

$$S_tf=\sum_{k\in N}\left[\int_0^t\int_{\partial\Omega}e^{-\lambda_k(t-s)}\varphi_k^\partial(y)f(y,s)\,dS_y\,ds\right]\varphi_k, \tag{10}$$

where $\varphi_k^\partial(y)$ is equal to $\varphi_k(y)/a$ if $a>0$, and $-\partial\varphi_k/\partial\nu$ if $a=0$.

(e) *If $u_0(x)$ and $f(x,t)$ are essentially bounded below by m (or above by M) the same is true for $u(x,t)$.*

3. Steady states for the heat equation.

DEFINITION. Let g be in $L_\infty(\partial\Omega)$. A function $u\in H$ is said to be a *steady state* (for the heat equation) *holdable* by g, if it is a weak solution of (2); i.e., if for each φ in $\{\varphi\in C^\infty(\bar\Omega):B\varphi\equiv 0 \text{ on } \partial\Omega\}$

$$\int_\Omega v(x)\,\Delta\varphi(x)\,dx+\int_{\partial\Omega}g(x)\varphi^\partial(x)\,dS=0. \tag{11}$$

Let S be the subspace of H consisting of all such steady states (elliptic theory implies the existence of steady states corresponding to each g; see Nečas [10]).

The role of steady states as targets is illuminated by the next result.

THEOREM 3.1. *Let $u_0\in H$ be given.*

(a) *Suppose $u_1\in H$ has the property that for some $\varepsilon>0$ there exists $f^\varepsilon\in L_\infty$ and $f^\varepsilon>0$ such that*

$$\|u_1-(V_tu_0+S_tf^\varepsilon)\|\leqq\varepsilon \quad \textit{for } t>t_\varepsilon.$$

Then there exists a steady state v with $\|u_1-v\|\leqq\varepsilon$.

(b) *Suppose the hypothesis of* (a) *holds for each $\varepsilon>0$ and that* $\sup_{\varepsilon>0}\|f^\varepsilon\|_\infty<\infty$.

Then u_1 is a steady state.

(c) *Suppose there exists f in L_∞ such that $\|u_1-(V_tu_0+S_tf)\|\to 0$ as $t\to\infty$. Then u_1 is a steady state.*

Proof. We prove (a). Since $u=u^\varepsilon$ is a weak solution of (1) (with $f=f^\varepsilon$), it follows from (6) that for any $p>0$, $T>0$ and φ in $C^\infty(\bar\Omega)$ with $B\varphi\equiv 0$ on $\partial\Omega$,

$$\int_\Omega\int_T^{T+p} u(x,t)\,\Delta\varphi(x)\,dx\,dt+\int_\Omega [u(x,T)-u(x,T+p)]\varphi(x)\,dx$$
$$+\int_{\partial\Omega}\int_T^{T+p} f(x,t)\varphi^\partial(x)\,dS_x\,dt=0.$$

Let $T=t^\varepsilon$, and define

$$u_p(x)=p^{-1}\int_T^{T+p} u(x,t)\,dt,\qquad f_p(x)=p^{-1}\int_T^{T+p} f(t,x)\,dt.$$

Then

$$\int_\Omega u_p(x)\,\Delta\varphi(x)\,dx+\int_{\partial\Omega} f_p(x)\varphi^\partial(x)\,dS_x \tag{12}$$
$$=p^{-1}\int_\Omega [u(x,T+p)-u(x,T)]\,dx.$$

Now, since $\|u(\cdot,T+p)\|_\infty\leqq \max(\|f\|_\infty,\|u(\cdot,T)\|_\infty)$, the right side of (12) converges to 0 as $p\to\infty$. Moreover, because $\|f_p\|_\infty\leqq\|f\|_\infty$, and

$$\|u_p-u_1\|\leqq\left(p^{-1}\int_T^{T+p}\|u(\cdot,t)-u_1\|^2\,dt\right)^{1/2}\leqq\varepsilon, \tag{13}$$

one can pick a sequence $p_n\uparrow\infty$, such that f_{p_n} converges weak* in $L_\infty(\partial\Omega)$ to $g_1=g_1^\varepsilon$ and u_{p_n} converges weakly in $L_2(\Omega)$ to $v=v^\varepsilon$. Passing to the limit in (13), and noting (14) one sees that v is a steady state (holdable by g_1, with $\|g_1\|_\infty\leqq\|f^\varepsilon\|_\infty$), and that $\|u_1-v\|\leqq\varepsilon$.

That (b) implies (c) is trivial. That (b) follows from (a) is proved by passing to the limit in

$$\int_\Omega v^\varepsilon(x)\,\Delta\varphi(x)\,dx=\int_{\partial\Omega} g_1^\varepsilon(x)\varphi^\partial(x)\,dS_x,$$

noting that $v^\varepsilon(x)\to u_1$ as $\varepsilon\to 0$, while g_1^ε has a weak* convergent subsequence because $\|g_1^\varepsilon\|_\infty\leqq\sup_{\varepsilon>0}\|f^\varepsilon\|_\infty<\infty$.

Most of the desirable properties of steady states depend on the following trivial lemma, which is obtained by setting $\varphi=\varphi_k=-\lambda_k^{-1}\,\Delta\varphi_k$ in (11).

LEMMA 3.2. *Let v be a steady state holdable by g. Then*

$$\int_\Omega v(x)\varphi_k(x)\,dx=\lambda_k^{-1}\int_{\partial\Omega} g(x)\varphi_k^\partial(x)\,dS_x. \tag{14}$$

4. Controllability results. We prove that

THEOREM 4.1. (a) *For any $u_0\in H$ and $t>0$, $S\subset R_t(u_0;L_\infty)$.*

(b) *Let u_1 be a steady state holdable by g. Suppose that $m<\operatorname{ess\,inf} g$, $\operatorname{ess\,sup} g<M$. Then, for any $u_0\in H$, $u_1\in R_t(u_0;L^g_{m,M})$ for t sufficiently large.*

Proof. The trivial proof depends on the less trivial fact that $0\in R_t(u_0;L_\infty)$. This property (null controllability) was proved for general Ω by Russell in [11], but can also be proved following Seidman [14], by more elementary means. Let v be a steady state holdable by g. To prove $v \in R_t(u_0; L_\infty)$ one needs to find $f \in L_\infty$ such that $V_t u_0 + S_t f = v$. Since obviously $v = V_t v + S_t g$ this requirement on f can be rewritten as $V_t(u_0 - v) + S_t(f - g) = 0$. Since $0 \in R_t(u_0 - v, L_\infty)$ such an f can be found, thus proving (a). The proof of (b) depends on the fact proved by Russell, that for any $u_0 \in H$ and $M > 0$, $0 \in R_t(u_0, L_M)$ for t sufficiently large. Thus the equation $V_t(u_0 - v) + S_t(f - g) = 0$ occurring above can be solved (for t sufficiently large) with $f - g$ arbitrarily small i.e., certainly with $f \in L_{m,M}$.

We show also some controllability results using the restricted classes of controls L_∞^g and $L_{m,M}^g$ defined in the introduction. Before evaluating $S_t f$ with $f \in L_\infty^g$ we introduce some notation. Let $\{-\mu_l\}_{l \in N}$ denote the *distinct* eigenvalues of L in decreasing order, $M_l = \{k \in N : \lambda_k = \mu_l\}$ and $P_l = \sum_{k \in M_l} (\cdot\,, \varphi_k)\varphi_k$ be the projection operator onto the eigenspace corresponding to μ_l. From the representation (10) and Lemma 3.2 it easily follows that, if $f(x, t) = g(x)h(t)$.

$$S_t f = \sum_{l \in N} \left(\mu_l \int_0^t e^{-\mu_l(t-s)} h(s)\, ds\right) P_l v, \tag{15}$$

where v is the steady state corresponding to g.

THEOREM 4.2. *Let v be the steady state corresponding to g, and $H_v = \{u \in H : u = \sum_{l=1}^\infty c_l P_l v\}$. ($H_v$ is a closed subspace of H). Then for any $u_0 \in H_v$, and any $t > 0$, $R_t(u_0, L_\infty^g)$ is dense in H_v.*

Proof. It is enough to consider the case $u_0 = 0$, since it follows from (9) that H_v is invariant under $\{V_t\}_{t \geqq 0}$. Suppose $R_t(0; L_\infty^g)$ is not dense in H_v. Then there exists a nonzero element $\sum_{l \in N} c_l P_l v$ in H_v satisfying the following identity for all $h \in L_\infty(0, \infty)$:

$$\left(\sum_{j \in N} c_j P_j v, \sum_{l \in N} \mu_l \int_0^t e^{-u_l(t-s)} h(s)\, ds\, P_l v\right) = 0.$$

Letting $h(s)$ be the characteristic function of $[0, r]$ (with $r < t$), and noting that $(P_j v, P_l v) = \delta_{jl} \|P_l v\|^2$, one gets, after integrating and then differentiating with respect to r, (which is permitted, since, as a consequence of (4), the series converges uniformly for $r \in [0, t - \varepsilon]$, with any $\varepsilon > 0$),

$$\sum_{l \in N} c_l \|P_l v\|^2 \mu_l e^{-\mu_l(t-r)} = 0.$$

It follows, by a standard argument, that $c_l \|P_l v\|^2 \mu_l = 0$ for each l, so that $\sum_{l \in N} c_l P_l v = 0$, a contradiction which completes the proof.

We have not been able to resolve the problem of null controllability within H_v using the controls L_∞^g. However a deep result of Galchuk [5], for which we give a new proof and a slight extension in an appendix, does allow us to characterize the situations in which it is possible to control exactly from $u_0 = c_0 v$ to $u_1 = c_1 v$ with controls in L_∞^g or $L_{m,M}^g$.

THEOREM 4.3. *Let v be a steady state holdable by g, c_0 and c_1 be constants. Then*

$$\sum_{l \in S} \mu_l^{-1} < \infty, \quad \textit{where } S = \{l \in N : P_l v \neq 0\}, \tag{16}$$

is a necessary and sufficient condition for the validity of each of the following statements:

(a) *$c_1 v \in R_t(c_0 v, L_\infty^g)$ for each $t > 0$;*

(b) *if $m < c_0 - |c_1 - c_0|$ and $M > c_0 + |c_1 - c_0|$ then $c_1 v \in R_t(c_0 v; L_{m,M}^g)$ for t sufficiently large.*

Proof. We prove first the equivalence of (16) and (b). We suppose that (16) holds. Then, using the representations (9) and (15) the requirement $c_1 v = V_t(c_0 v) + S_t f$ (with $f(x, t) = g(x)h(t)$) is equivalent to

$$c_1 = c_0\, e^{-\mu_l t} + \mu_l \int_0^{t\cdot} e^{-\mu_l(t-s)} h(s)\, ds \quad \text{for } l \text{ in } S$$

or, equivalently, to

$$\mu_l \int_0^t e^{-\mu_l s}(h(t-s) - c_0)\, ds = c_1 - c_0 \quad \text{for } l \text{ in } S. \tag{17}$$

Now, since (16) holds, the theorem of Galchuk (Appendix; A.1, part (b)) assures that for given ε, and t sufficiently large, the moment problem (17) has a solution satisfying

$$|h(t-s) - c_0| \leqq |c_1 - c_0| + \varepsilon.$$

If ε is sufficiently small this ensures that $f \in L^g_{m,M}$ so that indeed $c_1 v \in R_t(c_0 v; L^g_{m,M})$.

Suppose, conversely, that $c_1 v$ lies in $R_t(c_0 v; L^g_{m,M})$, or equivalently that (17) has a solution. Let $c = c_1 - c_0$, and $k(s) = h(t-s) - c_0$ for s in $(0, t)$ and $k(s) = 0$ for $s > t$. Then, for each l in S,

$$\int_0^\infty e^{-\mu_l s} k(s)\, ds = c\mu_l^{-1} = c \int_0^\infty e^{-\mu_l s}\, ds.$$

From this it follows that

$$\int_0^\infty P(s)(k(s) - c)\, ds = 0$$

for each $P(s)$ which is a finite linear combination of real exponential functions $e^{-\mu_l s}$, with l in S. Now $k(s) - c$ does not vanish identically, and lies in $L_\infty(0, \infty)$. Hence the class of the above exponential "polynomials" is not dense in $L_1(0, \infty)$. From well known results (see, for example, Schwartz [13]) it follows that (16) holds. The equivalence of (16) and assertion (a) is proved similarly using Theorem A.1, part (a).

It is interesting to apply this theorem in the case that $v \equiv 1$. One can check that (16) is then satisfied if Ω is a ball, but not if it is a parallelopipedon. In the former (but not in the latter) case one can therefore control exactly from one constant temperature to another using controls dependent on time alone.

5. Results on optimal controllability. In connection with Theorem 4.1 we have

THEOREM 5.1. *Let $u_0 \in H$ and u_1 be a steady state holdable by g. Suppose $m <$ ess $\inf_{x\in\partial\Omega} g(x)$, ess $\sup_{x\in\partial\Omega} g(x) < M$. Then there exists a unique $f_* \in L_\infty$ such that*

$$u_1 = V_{t_*} u_0 + S_{t_*} f_* \quad \textit{with } t_* = \inf\{t : u_1 \in R_t(u_0; L_{m,M})\}.$$

Moreover $f_*(x, t) = m$ or M a.e. on $\partial\Omega \times (0, t_*)$.

We remark that the fact that t_* is finite follows from (b) of Theorem 4.1, while the existence of f_* is a consequence of certain continuity properties of V_t and S_t. The "bang-bang" property of f_*, from which the uniqueness also follows, was proved in a previous version of this paper; that proof has however been generalized (also using ideas occurring in Fattorini [2] and Henry [7]) in Schmidt [12]; we refer to the latter paper for proofs and for greater precision in the formulation of the "bang-bang" property.

The final results deal with problems using the restricted control class $L^g_{m,M}$. These are much more easily proved than the deep Theorem 5.1. They can, unlike the latter theorem, be derived by the standard separation argument (used already in Yegorov [15]) as systematized with great generality by Knowles in [8]. For the sake of completeness we give a different direct proof.

THEOREM 5.2. *Let v be a steady state holdable by g, u_0 and u_1 belong to H_v, t be a fixed positive number and suppose that*

$$\inf\{\|V_t u_0 + S_t f - u_1\| : f \in L^g_{m,M}\} = \delta > 0. \tag{18}$$

Then there exists a unique $h_* \in L_\infty(0, t)$ with $m \leqq h_*(t) \leqq M$ such that, letting $f_*(x, t) = h_*(t)g(x)$, one has

$$\|V_t u_0 + S_t f_* - u_1\| = \delta.$$

Moreover, h^* takes on only the values m and M, with a finite number of "jumps" in each interval $(0, t-\varepsilon)$.

Proof. The existence of a minimizing control $f_*(x, t) = h_*(t)g(x)$ is standard. Let $u_0 = \sum_{l \in N} b_l P_l v$, $u_1 = \sum_{l \in N} c_l P_l v$. Then, using the representations (9) and (15) one has, for $f(x, t) = h(t)g(x)$,

$$\|V_t u_0 + S_t f - u_1\|^2 = \sum_{l \in N} \|P_l v\|^2 A_l(h)^2$$

with

$$A_l(h) = b_l e^{-\mu_l t} - c_l + \mu_l \int_0^t e^{-\mu_l(t-s)} h(s)\, ds.$$

Now h^* has to minimize the functional $J(h) = \sum_{l \in N} \|P_l v\|^2 (A_l(h))^2$ subject to the constraint $m \leqq h(t) \leqq M$. Hence, for any h satisfying the latter constraint, one must have that

$$J'(h^*)(h - h_*) \geqq 0,$$

where J' denotes the Fréchet derivative of the function J. More explicitly the latter condition is

$$\sum_{l \in N} \|P_l v\|^2 A_l(h_*) \mu_l \int_0^t e^{-\mu_l(t-s)} [h(s) - h_*(s)]\, ds \geqq 0. \tag{19}$$

Let c be any constant in (m, M) and

$$h(s) = (1 - \chi_{(r,r+\varepsilon)}(s)) h_*(s) + c \chi_{(r,r+\varepsilon)}(s),$$

where $\chi_{(r,r+\varepsilon)}(s)$ is the characteristic function of $(r, r+\varepsilon)$. Substituting h in (19), and letting $\varepsilon \downarrow 0$, one obtains for a.e. r in $(0, t)$ and for c in (m, M)

$$\sum \|P_l v\|^2 A_l(h_*) \mu_l e^{-\mu_l(t-r)} (c - h_*(r)) \geqq 0.$$

Let

$$\eta(r) = \sum_{l \in N} \|P_l v\|^2 A_l(h_*) \mu_l e^{-\mu_l(t-r)}.$$

This cannot vanish on a set having an accumulation point in $[0, t)$, for then one would have $\|P_l v\|^2 A_l(h_*) = 0$, for each l, in which case $J(h_*) = 0$, contradicting the assumption $\delta > 0$. From this fact, together with the inequality $\eta(t)(c - h^*(t)) \geqq 0$, the "bang-bang" property of h_* (and hence its uniqueness) follows.

This leads directly to a result on time-optimal control.

COROLLARY 5.3. *Let v be a steady state holdable by g, u_0 and u_1 belong to H_v. Let $\delta > 0$ be given along with constants m and M such that*

$$t_* = \inf\{t\colon \text{there exists } f \in L^g_{m,M} \text{ with } \|V_t u_0 + S_t f - u_1\| \leqq \delta\}$$

is finite. Then there exists a unique $f_*(x, t) = h_*(t)g(x)$ in $L^g_{m,M}$ with $\|V_{t_*}u_0 + S_{t_*}f_* - u_1\| = \delta$. The control h_* has the "bang-bang" property described in Theorem 5.2.

Proof. The existence of f_* is standard. The "bang-bang" property is proved by noting that f_* is also a solution of the norm approximation problem in Theorem 5.2 with $t = t_*$.

Remarks. (1) If $u_1 = c_1 v$ (c_1 a constant) it follows from (9) and (15) that, setting $f(x, t) = c_1 g(x)$, $V_t u_0 + S_t f \to u_1$ as $t \to \infty$. In this case therefore if $m < c_1 < M$, t_* is finite and the conclusion of the corollary holds.

(2) We have not been able to decide what happens in the case $\delta = 0$, i.e., when $t_* = \inf\{t\colon u_1 \in R_t(u_0; L^g_{m,M})\}$. This is related to the question of null controllability in H_v with controls in L^g_∞; null controllability would imply the "bang-bang" property.

Appendix. On the moment problem treated by Galchuk. In [5] Galchuk proved assertion (b) of the following theorem about the moment problem

$$\text{(A.1)} \qquad \int_0^t e^{-\mu_l s} f(s)\, ds = \mu_l^{-1} \quad \text{for } l \in N.$$

THEOREM A.1. *Let $\{\mu_l\}_{l\in N}$ be an increasing sequence of positive numbers such that $\sum_{l\in N}\mu_l^{-1} < \infty$. Then*

(a) *for each $t > 0$ the moment problem* (A.1) *has a solution f in $L_\infty(0, t)$;*

(b) *let $C > 1$ be given; then for all t sufficiently large* (A.1) *has a solution f in $L_\infty(0, t)$, with $\|f\|_\infty \leqq C$.*

We provide a proof which is considerably shorter and more elementary than that of Galchuk, at least if certain results on real exponentials due to Schwarz are assumed; moreover this proof yields (a) which was not proved by Galchuk. The following proposition is just a special case of a theorem due to Banach and Riesz.

PROPOSITION A.2. *The moment problem*

$$\text{(A.2)} \qquad \int_0^t e^{-\mu_l s} f(s)\, ds = c_l,$$

has a solution f in $L_\infty(0, t)$ if and only if there exists a positive constant C such that for each $\{\xi_k\}_{l\in N}$ with only a finite number of nonzero terms (the set of such sequences is denoted by S_F) one has

$$\text{(A.3)} \qquad \left|\sum_{l\in N} c_l \xi_l\right| \leqq C \int_0^t \left|\sum_{l\in N} \xi_l e^{-\mu_l s}\right| ds.$$

Moreover if C is the smallest constant for which (A.3) *holds one has*

$$C = \inf\{\|f\|_\infty : f \text{ is a solution of (A.2)}\}.$$

Proof. An easy estimate shows that if f is a solution of (A.2) one has (A.3) with $C \leqq \|f\|_\infty$.

Conversely, if (A.3) holds, one can define a nonvanishing functional on $L_1(0, t)$ by setting $F(\sum_{l\in N}\xi_l e^{-\mu_l s}) = \sum_{l\in N} c_l \xi_l$ (where $\{\xi_l\}_{l\in N} \in S_F$), extending that functional by continuity to the subspace generated by $\{e^{-\mu_l s}\}_{l\in N}$ and to all of $L_1(0, t)$ by the Hahn–Banach theorem. That functional has norm bounded by C and can be represented by a function $f \in L_\infty(0, t)$ with $\|f\|_\infty \leqq C$: f is a solution of (A.2).

The following result is a refinement of a theorem proved by L. Schwartz in [13]; P. Koosis showed me the proof.

THEOREM A.3. *Suppose* $\sum_{l\in N}\mu_l^{-1}<\infty$. *Then for each* $t>0$ *there exists a* (*least*) *constant* $C_t>0$ *such that*

$$\text{(A.4)}\qquad \int_0^\infty \Big|\sum_{l\in N}\xi_l e^{-\mu_l s}\Big|\,ds\leqq C_t\int_0^t\Big|\sum_{l\in N}\xi_l e^{-\mu_l s}\Big|\,ds$$

for each $\{\xi_l\}_{l\in N}\in S_F$; *moreover* $C_t\to 1$ *as* $t\to\infty$.

Proof. The existence of such a constant C_t for each $t>0$ is proved by Schwartz, but he does not discuss the asymptotic behavior. We show that $C_t\to 1$ as $t\to\infty$. Set $t=1$, clearly $C_1>1$. Now

$$\int_1^\infty\Big|\sum_{l\in N}\xi_l e^{-\mu_l s}\Big|\,ds=\int_0^\infty\Big|\sum_{l\in N}\xi_l e^{-\mu_l s}\Big|\,ds-\int_0^1\Big|\sum_{l\in N}\xi_l e^{-\mu_l s}\Big|\,ds$$
$$\leqq\rho\int_0^\infty\Big|\sum_{l\in N}\xi_l e^{-\mu_l s}\Big|\,ds,$$

where $\rho=1-(1/C_1)$ satisfies $0<\rho<1$. Using the identity

$$\int_{n+1}^\infty\Big|\sum_{l\in N}\xi_l e^{-\mu_l s}\Big|\,ds=\int_n^\infty\Big|\sum_{l\in N}(\xi_l e^{-\mu_l})\,e^{-\mu_l s}\Big|\,ds$$

one can easily prove inductively that

$$\int_n^\infty\Big|\sum_{l\in N}\xi_l e^{-\mu_l s}\Big|\,ds\leqq\rho^n\int_0^\infty\Big|\sum_{l\in N}\xi_l e^{-\mu_l s}\Big|\,ds.$$

Hence it follows that

$$\int_0^\infty\Big|\sum_{l\in N}\xi_l e^{-\mu_l s}\Big|\,ds\leqq(1-\rho^n)^{-1}\int_0^n\Big|\sum_{l\in N}\xi_l e^{-\mu_l s}\Big|\,ds,$$

so that $C_n\to 1$ as $n\to\infty$. Since C_t is monotone decreasing one also has $C_t\to 1$ as $t\to\infty$.

Now we can prove Theorem A.1. To prove the solvability of the moment problem (A.1) using an estimate of the form (A.3) we note that for $\{\xi_l\}_{l\in N}\in S_F$

$$\Big|\sum_{l\in N}\mu_l^{-1}\xi_l\Big|=\Big|\int_0^\infty\sum_{l\in N}\xi_l e^{-\mu_l s}\,ds\Big|$$
$$\leqq\int_0^\infty\Big|\sum_{l\in N}\xi_l e^{-\mu_l s}\Big|\,ds\leqq C_t\int_0^t\Big|\sum_{l\in N}\xi_l e^{-\mu_l s}\Big|\,ds$$

and hence (A.1) has a solution f with $\|f\|_\infty\leqq C_t$; thus (a) is proved and (b) follows also since $C_t\to 1$ as $t\to\infty$.

Remarks. (1) Previously M. von Golitschek has assisted the author in proving that $C_t=1+O(e^{-\mu_l t})$ as $t\to\infty$, under the additional hypothesis that $\mu_{l+1}-\mu_l\geqq\mu>0$ for all l and some μ.

(2) The problem of finding biorthogonal series $\{f_l\}_{l\in N}(\subset L_\infty(0,t))$ to $\{e^{-\mu_l s}\}_{l\in N}(\subset L_1(0,t))$, i.e., solutions to $\int_0^t e^{-\mu_l s}f_k(s)\,ds=\delta_{kl}$, can also be solved using Proposition A.2; estimates for $\|f_l\|_\infty$ are also given by this approach. This provides an alternative approach to that of Fattorini and Russell (who use the Hilbert space structure of $L_2(0,t)$ to construct biorthogonal functions in that space, for which they then obtain uniform estimates), but does not yield new results.

Acknowledgment. The author would like to acknowledge the hospitality of the Institute for Applied Mathematics and Statistics at the University of Würzburg, where most of the work was done.

REFERENCES

[1] S. AGMON, *Lectures on Elliptic Boundary Value Problems*, Van Nostrand, New York, 1965.

[2] 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.

[3] H. O. FATTORINI AND D. L. RUSSELL, *Exact controllability theorems for linear parabolic equations in one space dimension*, Arch. Rat. Mech. Anal., 4 (1971), pp. 272–292.

[4] ———, *Uniform bounds on biorthogonal functions for real exponentials with an application to the control theory of parabolic equations*, Quart. Appl. Math., 32 (1971), pp. 45–69.

[5] L. I. GALCHUK, *Optimal control of systems described by parabolic equations*, this Journal, 7 (1969), pp. 546–558.

[6] K. GLASHOFF AND N. WECK, *Boundary control of parabolic differential equations in arbitrary dimensions; supremum norm problems*, this Journal, 14 (1976), pp. 662–681.

[7] J. HENRY, *Controle en temps optimal pour les systemes gouvernés par une équation de type parabolique*, preprint.

[8] G. KNOWLES, *Time optimal control of infinite dimensional systems*, this Journal, 14 (1976), pp. 919–933.

[9] R. C. MACCARMY, V. J. MIZEL AND T. I. SEIDMAN, *Approximate boundary controllability for the heat equation*, J. Math. Anal. Appl., 23 (1968), pp. 699–703.

[10] J. NECAS, *Les Méthodes Directes en Théorie des Equations Elliptiques*, Masson et Cie., Paris, 1967.

[11] D. L. RUSSELL, *A unified boundary controllability theory for hyperbolic and parabolic partial differential equations*, Studies in Appl. Math., 52 (1973), pp. 189–211.

[12] G. SCHMIDT, *The "bang-bang" principle for the time optimal problem in boundary control of the heat equation*, this Journal, to appear.

[13] L. SCHWARTZ, *Etude des sommes d'exponentielles*, 2^{eme} edition, Hermann, Paris, 1959.

[14] T. I. SEIDMAN, *A well-posed problem for the heat equation*, Bull. Amer. Math. Soc., 80 (1974), pp. 901–902.

[15] J. V. YEGOROV, *Some problems in the theory of optimal controls*, U.S.S.R. Computational, Math. and Math. Phys., 3 (1963), pp. 1209–1232.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 2, March 1980

STOCHASTIC REALIZATION AND INVARIANT DIRECTIONS OF THE MATRIX RICCATI EQUATION*

MICHELE PAVON†

Abstract. Invariant directions of the Riccati difference equation of Kalman filtering are shown to occur in a large class of prediction problems and to be related to a certain invariant subspace of the transpose of the feedback matrix. The discrete time stochastic realization problem is studied in its deterministic as well as probabilistic aspects. In particular a new derivation of the classification of the minimal Markovian representations of the given process z is presented which is based on a certain backward filter of the innovations. For each Markovian representation which can be determined from z the space of invariant directions is decomposed into two subspaces, one on which it is possible to predict the state process without error forward in time and one on which this can be done backward in time.

Introduction. The aim of this paper is to extend the theory of invariant directions of the matrix Riccati equation to a large class of filtering problems, to present some new results on the deterministic and probabilistic aspects of the discrete time stochastic realization problem and to illustrate the particular features introduced in stochastic realization by the presence of invariant directions.

Section 1 of the paper is concerned with characterizing invariant vectors for the usual linear least squares estimation problem in additive white noise. We extend the previous results on the colored noise problem [8], [14], [29] to our more general setting and present some new ones. The main result of this part is Theorem 1.6 which provides different necessary and sufficient conditions for invariance. These conditions are phrased in terms of the convolution of two weighting patterns, of the optimal control of the dual problem, of the best one step predictor and of the feedback matrix $\Gamma(t)$ of the Kalman filter. The latter characterization appears here for the first time. Indeed, the space of all invariant directions is simply the invariant subspace related to the eigenvalue zero of the transpose of $\Gamma(t)$ for t larger than a certain value. This interpretation turns out to be quite useful and enlightening, since $\Gamma(\cdot)$ is the transition matrix of the estimation error and it is essential in classifying Markovian representations in the stochastic realization setting (see e.g., Theorem 2.8). Also the fact that invariant vectors are generalized eigenvectors sheds new light on the proof techniques employed in [8], [9], [29]. The paper by Clements and B. D. O. Anderson [9], which contains results closely related to conditions (ii) and (iii) of Theorem 1.6, became available to us right after the first version of this paper was submitted. The emphasis in [9], however, is somewhat different from ours in that the authors seek to characterize invariance for a very general form of the linear quadratic regulator problem, whereas our main interest lies in the stochastic implications of this phenomenon.

The second part of the paper deals with discrete time stochastic realization theory. Given a wide sense stationary vector process z with rational spectral density Φ, such that $\Phi(\infty)$ is finite and $\Phi(e^{i\omega})$ is positive definite for all ω, and a Hilbert space H containing the components of $z(t)$ for all t, consider the problem of determining all minimal Markovian representations of z (*stochastic realizations*) driven by a white noise with components in H. We solve the problem in the following way. First the second order properties of the stochastic realizations are described. Our results

* Received by the editors August 29, 1977, and in final revised form May 29, 1979.

† Graduate School of Mathematics, University of Kentucky, Lexington, Kentucky, 40506. This work was supported in part by the C.N.R. of Italy, the National Science Foundation under Grants MPS75-07028 and MCS77-00241 and the Air Force Office of Scientific Research under Grant AFOSR-78-3519.

integrate those of B. D. O. Anderson [3]–[5], Faurre [11], [12] and Ruckebusch [33], [34]. In particular, we show that the correspondence in [33, p. 70] between realizations with square transfer function and real symmetric solutions of a certain algebraic matrix equation of the Riccati type holds without any assumption on the feedback matrix. Our analysis on this aspect of the stochastic realization problem parallels in some respects the continuous time work of Lindquist and Picci [19].

Then we turn to the probabilistic side of the problem which has received considerable attention in recent years [1], [2], [18]–[23], [27], [32]–[36]. A tool for this study is provided to us by Theorem 2.5, which establishes a correspondence between the deterministic as well as stochastic elements of realizations evolving forward and backward in time. The last two subsections of § 2 are devoted to a new derivation of the classification of the state processes of stochastic realizations due to Ruckebusch [33] in discrete time and Lindquist and Picci in continuous time [19]. Our approach makes essential use of Markovian representations of the innovation process with the estimation error as the state. Ruckebusch has used the error process in finite and infinite dimensional stochastic realization to derive a number of results [33]–[35], but our idea of associating it with a stochastic realization of the innovations appears to be new. Tackling the problem in this way we not only derive the main results in a rather simple manner, but we also gain insight into their meaning. For instance, the important result that realizations which can be constructed from only the process z (*internal*) are in one to one correspondence with the invariant subspaces of the feedback matrix Γ_* (Theorem 2.8) can be given a natural explanation in terms of the backward filter of the innovations (see Remark 2.10). Last, but not least, these stochastic realizations of the innovation process provide a key to understanding the relationship between the invariant subspaces of Γ_* and a certain class of *inner* functions in terms of which it is possible to describe the realizations of z [21], [35], [36]. Our results on this subject, however, will be presented elsewhere.

Section 3 is the natural continuation of §§ 1 and 2 in that it explores how invariant directions affect the family of stochastic realizations. Indeed the space of invariant vectors $\mathcal{I}$ is the same for all realizations and is nontrivial if and only if $\Phi(\infty)$ is singular. The characterization of $\mathcal{I}$ as the invariant subspace of the transpose of Γ_* relative to zero is important in establishing the two principal results of § 3. The first is Theorem 3.8 which says, loosely speaking, that in an invariant direction we can either predict or smooth the state of an internal realization exactly (i.e., without error), showing that $\mathcal{I}$ is closely related to the *germ space* of z [23]. The second is Theorem 3.9 which embeds every internal realization in a chain of internal realizations (totally ordered with respect to state covariances) whose minimum element has a full set of *predictable* directions [14] and whose maximum one has a full set of *smoothable* directions (Definition 3.7).

The last subsection of § 3 is devoted to comparing two possible approaches to discrete time stochastic realization based on different factorizations of the covariance operator. We show that the factorization leading to Markovian representations without noise in the output [1], [11] considerably narrows, compared with the other approach, the solution class of the stochastic realization problem when $\Phi(\infty)$ is singular. This deficiency of the first method makes it advisable to seek Markovian representations of the type considered in this paper unless nonsingularity of $\Phi(\infty)$ is guaranteed.

It is worthwhile remarking that the assumptions made on the process z in §§ 2 and 3 are mostly for simplicity. Indeed many of the central results can be established, in a suitably modified form, in the nonstationary case under mild assumptions on z, albeit the derivation becomes more involved. This explains why we refrain from introducing backward realizations and related concepts, like that of smoothable direction, in the setting of § 1. Our results on this matter will be presented somewhere else.

The scalar case has some interesting features for which we refer the reader to [23].

1. Invariant directions of the matrix Riccati equation.

1.1. Basic notation and formulation of the problem. We use standard vector-matrix notation, with the following conventions. The unit matrix is denoted by I, the transpose of a matrix by prime. All vectors without prime are column vectors. $\mathcal{N}(R)$ indicates the null space of the matrix R. If R is symmetric, $R>0$ ($R\geqq 0$) means R positive (nonnegative) definite. If $R\geqq 0$, $R^{1/2}$ is the unique nonnegative square root of R. The Moore–Penrose pseudoinverse [26] is denoted by $\#$. The trace operator is indicated by tr. The cone of symmetric, nonnegative definite $n\times n$ matrices is denoted by $\mathcal{C}_n$. The Kronecker symbol is δ_{st}. The superscript o identifies "optimal."

Consider the linear stochastic model

$$x(t+1)=Ax(t)+Bw(t), \tag{1.1}$$

$$y(t)=Cx(t)+Dw(t) \tag{1.2}$$

with initial condition $x(0)=x_0$, where A, B, C and D are constant matrices of dimensions $n\times n$, $n\times p$, $m\times n$ and $m\times p$, x_0 is an n-dimensional zero-mean random vector, the input w is a p-dimensional zero-mean white noise sequence uncorrelated with x_0, $E\{x_0x_0'\}=P_0$ and $E\{w(s)w(t)'\}=I\delta_{st}$.

As is well-known, the best linear least-squares estimate $\hat{x}(t)$ of $x(t)$, given the data $\{y(0),\cdots,y(t-1)\}$, is generated recursively by the Kalman filter

$$\hat{x}(t+1)=A\hat{x}(t)+K(t)[y(t)-C\hat{x}(t)],\qquad \hat{x}(0)=0, \tag{1.3}$$

where $K(t)$ is given by

$$K(t)=(A\Sigma(t)C'+BD')(C\Sigma(t)C'+DD')^{\#} \tag{1.4}$$

and $\Sigma(t)$ satisfies the Riccati difference equation

$$\begin{aligned}\Sigma(t+1)=A\Sigma(t)A'-(A\Sigma(t)C'+BD')(C\Sigma(t)C'+DD')^{\#}(C\Sigma(t)A'+DB')&\\ +BB',&\\ \Sigma(0)=P_0.&\end{aligned} \tag{1.5}$$

We shall indicate the solution of (1.5) at time s by $\Sigma(s;P_0)$ when we intend to emphasize the dependence on the initial condition P_0.

DEFINITION 1.1 ([8]). The n-dimensional vector a is called an *s-invariant direction* of (1.5) if $a'\Sigma(t;P_0)=a'\Sigma(s;0)$ for all $t\geqq s$ and all $P_0\in\mathcal{C}_n$.

We shall study the problem of characterizing all invariant directions of (1.5).

1.2. Preliminaries. In this section we transcribe some well known results of duality between estimation and control into a form best suited to our problem. We refer the reader to [24] for the variational principles underlying this duality.

Since $\hat{x}(t+1)$ is in the linear span of $y(0),\cdots,y(t)$ there exist matrices $U(s,t)^{\circ}$ for $s=0,\cdots,t$ such that $\hat{x}(t+1)=-\sum_{s=0}^{t}(U(s,t)^{\circ})'y(s)$. Such sequence is optimal for the following *dual problem*: find $U(t)=(U(0,t),\cdots,U(t,t))$ which minimizes

$$\operatorname{tr}\{J[U(t)]\}=\operatorname{tr}\{Q(-1,t)'P_0Q(-1,t)+\sum_{s=0}^{t}Z(s,t)'Z(s,t)\}, \tag{1.6}$$

where

$$Q(s-1,t)=A'Q(s,t)+C'U(s,t), \qquad Q(t,t)=I, \tag{1.7}$$

$$Z(s,t)=B'Q(s,t)+D'U(s,t). \tag{1.8}$$

A standard argument yields the closed-loop form of the optimal control

$$U(s,t)^{\circ}=-K(s)'Q(s,t)^{\circ}, \qquad s=0,\cdots,t. \tag{1.9}$$

Consider the linear estimator of $x(t+1)$ given by $\gamma(t+1)=-\sum_{s=0}^{t} U(s,t)'y(s)$. Then it is easily seen that

$$x(t+1)-\gamma(t+1)=Q(-1,t)'x_0+\sum_{s=0}^{t} Z(s,t)'w(s). \tag{1.10}$$

Introducing the quantities $P(s,t)=E\{x(s)[x(t+1)-\gamma(t+1)]'\}$, $R(s,t)=E\{y(s)[x(t+1)-\gamma(t+1)]'\}$ and applying the operator $E\{\cdot[x(t+1)-\gamma(t+1)]'\}$ to both sides of (1.1)–(1.2) we obtain, in view of (1.10), the following *adjoint system*

$$P(s+1,t)=AP(s,t)+BZ(s,t), \qquad P(0,t)=P_0Q(-1,t), \tag{1.11}$$

$$R(s,t)=CP(s,t)+DZ(s,t). \tag{1.12}$$

The terminology is justified by the fact that, setting up the discrete minimum principle for the dual problem (1.11) are seen to be, with the appropriate normalization, the adjoint equations. Let us note that

$$R(s,t)=0, \qquad s=0,\cdots,t \tag{1.13}$$

is a necessary and sufficient condition for optimality of the $U(t)$ sequence. Whenever A is nonsingular we can rewrite (1.7) in the form

$$Q(s,t)=(A')^{-1}Q(s-1,t)-(A')^{-1}CU(s,t), \qquad Q(t,t)=I. \tag{1.14}$$

Hence we have the following input-output relations:

$$Z(s,t)=\sum_{i=0}^{s} \tilde{T}(i)U(s-i,t)+B'(A')^{-s-1}Q(-1,t), \tag{1.15}$$

$$R(s,t)=\sum_{i=0}^{s} T(i)Z(s-i,t)+CA^{s}P_oQ(-1,t), \tag{1.16}$$

where the weighting patterns $\tilde{T}(\cdot)$ and $T(\cdot)$ are defined by

$$\tilde{T}(i)=\begin{cases} D'-B'(A')^{-1}C', & i=0,\\ -B'(A')^{-i-1}C', & i>0,\end{cases} \tag{1.17}$$

$$T(i)=\begin{cases} D, & i=0,\\ CA^{i-1}B, & i>0.\end{cases} \tag{1.18}$$

Combining (1.14) and (1.15) leads us to the *Hamiltonian system*

$$\begin{gathered}\begin{pmatrix} Q(s,t)\\ P(s+1,t)\end{pmatrix}=\begin{bmatrix} (A')^{-1} & 0\\ BB'(A')^{-1} & A\end{bmatrix}\begin{pmatrix} Q(s-1,t)\\ P(s,t)\end{pmatrix}\\ +\begin{bmatrix} -(A')^{-1}C\\ BD'-BB'(A')^{-1}C'\end{bmatrix}U(s,t),\\ \begin{pmatrix} Q(-1,t)\\ P(0,t)\end{pmatrix}=\begin{pmatrix} I\\ P_o\end{pmatrix}Q(-1,t),\end{gathered} \tag{1.19}$$

$$(1.20) \qquad R(s,t)=[DB'(A')^{-1} \quad C]\begin{pmatrix} Q(s-1,t) \\ P(s,t) \end{pmatrix}+[DD'-DB'(A')^{-1}C']U(s,t),$$

where $Q(-1,t)=(A')^t+\sum_{i=0}^{t}(A')^iC'U(i,t)$. It is clear that the weighting pattern $T_H(\cdot)$ of the Hamiltonian system is just the convolution of $T(\cdot)$ and $\tilde{T}(\cdot)$.

$$(1.21) \qquad T_H(i)=[T*\tilde{T}](i)=\sum_{j=0}^{i} T(i-j)\tilde{T}(j).$$

The matrices $T_H(0),\cdots,T_H(n-1)$ will play a central role in establishing necessary and sufficient conditions for invariance.

1.3. Characterization of invariant directions. We study the case where A is nonsingular. This assumption enables us to derive explicit expressions for the invariant vectors. (The case where no restriction is placed on A and on the definitness of the criterion matrices has been recently investigated in [9]). The three following lemmas extend known results to our more general setting.

LEMMA 1.2. *The vector a is an s-invariant direction of* (1.5) *if and only if*

$$(1.22) \qquad a\in\mathcal{N}(Q(t-s,t)^\circ) \quad \textit{for all } t\geqq s-1 \textit{ and all } P_0\in\mathscr{C}_n.$$

Proof. Observe that a control $U(t)$ is optimal for the dual problem if and only if it minimizes $a'J[U(t)]a$ for all $a\in\mathsf{R}^n$. The result now follows from a straightforward modification of the argument of Theorem 3 in [29]. □

Notice that optimal quantities in the dual problem depend on the terminal weight P_0. To keep notations simple, we shall refrain from explicitly exhibiting this dependence.

Remark 1.3. The proof of the sufficiency part in Lemma 1.2 relies on the fact that, under condition (1.22), $U(t-i,t)^\circ a$ is invariant over $t\geqq s$ for $i=0,\cdots,s-1$. Moreover, when (1.22) holds, it is easily seen using (1.7)–(1.9) that $a\in\mathcal{N}(U(i,t)^\circ)\cap\mathcal{N}(Z(i,t)^\circ)$ for $i=0,\cdots,t-s$. In particular it follows from (1.10) that $a'\tilde{x}(t+1)=a'\sum_{i=t-s+1}^{t}(Z(i,t)^\circ)'w(i)$, where $\tilde{x}(t)=x(t)-\hat{x}(t)$ is the estimation error.

The mathematical framework set up in the previous section will be useful in proving the following result.

LEMMA 1.4. *The vector a satisfies* (1.22) *if and only if*

$$(1.23) \qquad a=-\sum_{i=1}^{s}(A')^{-i}C'\lambda_i,$$

where the m-dimensional vectors $\lambda_1,\lambda_2,\cdots,\lambda_s$ *are such that*

$$(1.24) \qquad \sum_{i=0}^{s-j}T_H(i)\lambda_{j+i}=0, \qquad j=1,\cdots,s.$$

In this case the optimal control satisfies

$$(1.25) \qquad U(t)^\circ a=(0,\cdots,0,\lambda_s,\cdots,\lambda_1).$$

Proof. Assume that (1.22) holds. In view of the time invariance discussed in Remark 1.3, we can set $\lambda_i=U(t-i+1,t)^\circ a$ for $i=1,\cdots,s$. Expression (1.23) can now be derived using (1.7) recursively. Let us consider the input-output relation of the Hamiltonian system

$$R(s,t)=[DB'(A')^{-1} \quad C]A_H^s\begin{pmatrix} I \\ P_0 \end{pmatrix}Q(-1,t)+\sum_{i=0}^{s}T_H(i)U(s-i,t),$$

where

$$A_H = \begin{bmatrix} (A')^{-1} & 0 \\ BB'(A')^{-1} & A \end{bmatrix}.$$

As observed in Remark 1.3, $a \in \mathcal{N}(Q(-1, t)^{\circ})$. Then (1.24) follows from the optimality conditions (1.13). Conversely suppose a is as in (1.23) with the λ_j: s satisfying (1.24). Using (1.9) and, recursively, (1.7), we obtain

$$U(k, t)^{\circ}a = -K(k)'\left[(A')^{t-k} + \sum_{i=1}^{t-k} (A')^{i-1}C'U(k+i, t)^{\circ}\right]a$$

which, together with (1.23) yields

$$U(k, t)^{\circ}a = -K(k)'\left\{-\sum_{i=1}^{s-t+k} (A')^{-i}C'\lambda_{t-k+1} + \sum_{i=1}^{t-k} (A')^{i-1}C'[U(k+i, t)^{\circ}a - \lambda_{t-k-i+1}]\right\}.$$

A calculation similar to that found in the proof of Theorem 8 in [29], i.e., using (1.4), (1.5) repeatedly and condition (1.24), shows that

$$K(k)' \sum_{i=1}^{s-t+k} (A')^{-i}C'\lambda_{t-k+i} = \lambda_{t-k+1} \tag{1.26}$$

which, inserted into the previous expression for $U(k, t)^{\circ}a$, enables us to derive $U(k, t)^{\circ}a = \lambda_{t-k+1}$ for $k = t-s+1, \cdots, t$ recursively. This and (1.7) yield $Q(t-s, t)^{\circ}a = 0$, i.e., condition (1.22). Also (1.25) now follows in view of Remark 1.3. This completes the proof. □

A straightforward extension of the proof of Theorem 8 in [29] establishes the following lemma.

LEMMA 1.5. *A vector a is s-invariant for* (1.5) *if and only if a is as in* (1.23) *and*

$$a'\hat{x}(t+1) = -\sum_{i=1}^{s} \lambda_i' y(t+1-i) \quad \text{for all } t \geqq s-1. \tag{1.27}$$

Let $\Gamma(t)$ denote the *feedback matrix* $A - K(t)C$.

THEOREM 1.6. *The following statements are equivalent*:

(i) *a is an s-invariant direction of* (1.5).

(ii) *a satisfies* (1.22).

(iii) *a is as in* (1.23) *and* (1.24) *holds.*

(iv) *a is as in* (1.23) *and* (1.27) *holds.*

(v) *a generates the same s-dimensional cyclic subspace of* $\Gamma(t)'$ *for all* $t \geqq s-1$ *and all* $P_0 \in \mathscr{C}_n$; *this invariant subspace of* $\Gamma(t)'$ *is associated with the eigenvalue zero, i.e.,* $(\Gamma(t)')^s a = 0$. *Moreover* $\Gamma(t-s+1)' \cdots \Gamma(t)'a = 0$ *for all* $t \geqq s-1$.

Proof. The equivalence of (i), (ii), (iii) and (iv) follows directly from Lemmas 1.2, 1.4 and 1.5. Suppose a satisfies (v) and observe that relations (1.7) and (1.9) yield the expression $Q(t-s, t)^{\circ} = \Gamma(t-s+1)' \cdots \Gamma(t)'$. By assumption $\Gamma(t-s+1)' \cdots \Gamma(t)'a = 0$ and (1.22) follows. Conversely, if we assume (iii), we derive from (1.26) and the last part of the proof of Lemma 1.4 the relation

$$\Gamma(t)' \sum_{i=1}^{s-j} (A')^{-i}C'\lambda_{i+j} = \sum_{i=1}^{s-j-1} (A')^{-i}C'\lambda_{i+j+1}$$

for all $t \geqq s-j-1$ and all $P_0 \in \mathscr{C}_n$, where $j = 1, \cdots, s-1$ and, for $j = s-1$, the right hand side is defined to be zero. This establishes (v). □

Condition (v) of this theorem is new. Its importance will completely surface in the stochastic realization setting.

Remark 1.7 ([8]). The sets I_s of s-invariant directions and $\mathscr{I} = \bigcup_{s=1}^{\infty} I_s$ of invariant directions are vector spaces. It follows from the previous theorem that $\mathscr{I} = \bigcup_{s=1}^{n} I_s$.

Remark 1.8. The dimension of the *invariant subspace* $\mathscr{I}$ can be easily determined in the single-output case $y(t) = c'x(t) + d'w(t)$. It is equal to the minimum between the rank of the observability matrix $[c \quad A'c, \cdots, (A')^{n-1}c]'$ and the first index j such that $T_H(j-1) = \cdots = T_H(0) = 0$ and $T_H(j) \neq 0$. The general case is rather involved. We shall not pursue here the extension of the results of [29] on this matter.

Let

$$W(z) = \sum_{i=0}^{\infty} T(i)z^{-i} = C(zI - A)^{-1}B + D \tag{1.28}$$

be the *transfer function* of (1.1)–(1.2) and

$$W_H(z) = \sum_{i=0}^{\infty} T_H(i)z^{-i} \tag{1.29}$$

the transfer function of the Hamiltonian system. The following characterization of $T_H(\cdot)$ will be helpful in the third part of the paper.

THEOREM 1.9. *Assume A nonsingular. Then*

$$W_H(z) = W(z)W(z^{-1})'. \tag{1.30}$$

If y in (1.2) *is stationary with spectral density $\Phi(z)$, we also have*

$$W_H(z) = \Phi(z). \tag{1.31}$$

Proof. Consider $W(z^{-1})' = B'(z^{-1}I - A')^{-1}C' + D' = -B'(A')^{-1}(I - z^{-1} \times (A')^{-1})^{-1}C' + D'$. Expand the last term in a neighborhood of infinity as follows:

$$\begin{aligned} &-B'(A')^{-1}(I - z^{-1}(A')^{-1})^{-1}C' + D' \\ &= D' - B'(A')^{-1}C' - B'(A')^{-2}C'z^{-1} - B'(A')^{-3}C'z^{-2} \cdots \\ &= \sum_{i=0}^{\infty} \tilde{T}(i)z^{-i}. \end{aligned} \tag{1.32}$$

Take the Cauchy product of the two series in (1.28) and (1.32) to get (1.30). In the case of a stationary y the well-known spectral factorization formula

$$\Phi(z) = W(z)W(z^{-1})' \tag{1.33}$$

yields (1.31). □

Notice that the calculations in the previous theorem make sense because the series in (1.28) and (1.29) converge respectively to $W(z)$ and to $W_H(z)$ in an appropriate neighborhood of infinity.

Let $\Delta(t, s) = E\{y(t)y(s)'\}$ be the covariance operator of the observations. It is a simple matter, using the expression $y(s) = CA^{-n}x(s+n) + \sum_{i=0}^{n-1} \tilde{T}(i)'w(s+i)$ which can be derived from (1.1)–(1.2), to see that the parameters $T_H(0), \cdots, T_H(n-1)$ determine the degree of "smoothness" of $\Delta(\cdot, \cdot)$, i.e., the number of differencing operations on $\Delta(\cdot, \cdot)$ necessary in each direction to produce a Kronecker delta. This number has been named in the scalar case *relative order of the covariance*, see [14] for example. This fact has its counterpart in the spectral domain in Theorem 1.9.

1.4. Predictable directions. The invariance properties of invariant directions have been pointed out by several authors [8], [14]. Indeed, as it is apparent from Theorem 1.6, the space $\mathcal{I}$ is invariant over models (1.1)–(1.2) having the same covariance of the output and the same (up to a change of basis in the state space) pair (A, C). However, if a is an s-invariant vector for (1.5) the value $a'\Sigma(s; P_0)$ does depend on the model. A special case of particular interest is when $a \in \mathcal{N}(\Sigma(s; P_0))$.

DEFINITION 1.10 ([14]). The n-dimensional vector a is called an *s-predictable direction* of (1.5) if $a'\Sigma(t; P_0) = a'\Sigma(s; P_0) = 0$ for all $t \geqq s$. The two following theorems extend some results of Gevers [14].

THEOREM 1.11. *The vector a is an s-predictable direction of* (1.5) *if and only if a is as in* (1.23) *with the* λ_i *satisfying*

$$\sum_{i=0}^{s-j} \tilde{T}(i)\lambda_{j+i} = 0, \qquad j = 1, \cdots, s. \tag{1.34}$$

Proof. If a is s-predictable $a'\tilde{x}(t+1) = 0$ for all $t \geqq s-1$. Using (1.10) with optimal quantities we see that $a \in \mathcal{N}(Q(-1, t)^{\circ})$ and $a \in \bigcap_{i=0,\cdots,t} \mathcal{N}(Z(i, t)^{\circ})$ for all $t \geqq s-1$. Again time invariance of the optimal control can be shown to hold and, identifying quantities as in (1.25), we get (1.23) from $Q(-1, s-1)^{\circ}a = 0$. Also (1.34) follows from (1.15). To prove the converse first observe that (1.34) implies (1.24). By Lemma 1.4 $a \in \mathcal{N}(Q(-1, t)^{\circ})$ and (1.25) holds. From (1.15) and (1.10) we conclude that $a'\tilde{x}(t+1) = 0$ for all $t \geqq s-1$, i.e., a is s-predictable. □

THEOREM 1.12. *Let* $\Sigma(s; P_0) > 0$. *Then* $\Sigma(t; P_0) > 0$ *for all* $t \geqq s$ *if and only if* $\tilde{T}(0)$ *has rank m.*

Proof. Let λ be such that $\tilde{T}(0)\lambda = 0$. Then $(A')^{-1}C'\lambda \in \mathcal{N}(\Sigma(t; P_0))$ for all $t \geqq 1$. To prove the other half we use induction. Suppose $\Sigma(t-1; P_0) > 0$ and $a \in \mathcal{N}(\Sigma(t; P_0))$. It follows from the principle of optimality that

$$\begin{aligned} 0 = a'\Sigma(t; P_0)a = \min_{\lambda \in R^m} \{&(a'A + \lambda'C)\Sigma(t-1; P_0)(A'a + C'\lambda) \\ &+ (a'B + \lambda'D)(B'a + D'\lambda)\}. \end{aligned} \tag{1.35}$$

Let λ° be the optimal value in (1.35). Since $\Sigma(t-1;\, P_0) > 0$ we get $a = -(A')^{-1}C'\lambda^{\circ}$, $B'a + D'\lambda^{\circ} = 0$ and finally $(D' - B'(A')^{-1}C')\lambda^{\circ} = 0$. If $\tilde{T}(0)$ has rank m this implies that $a = 0$. □

Remark 1.13. Theorem 1.12 agrees with the results obtained by Silverman et al. [25], [30], [38]. In fact, the presence of nontrivial predictable directions of (1.5) implies that the system (1.1)–(1.2) is not *strongly observable* [38]. However, it can well happen that it is completely observable (and controllable). In the third part of the paper we shall study a set of minimal realizations with a nontrivial invariant and, for some of them, predictable subspace.

1.5. Discussion. Our study has shown that invariant directions can occur in a more general situation than just the noise-free measurements case treated in [8], [14], [29]. Conditions (iv) and (v) of Theorem 1.6 provide us with a probabilistic interpretation of this phenomenon. In an invariant direction the optimal filter depends only on some of the last observation instead of the whole information available. This fact is strictly related to the invariant subspace of $\Gamma(t)'$ corresponding to zero. Moreover, in the case when y is stationary with rational spectral density, condition (iii) of Theorem 1.6 with Theorem 1.9 shows a precise connection between invariant vectors and the spectrum of y. All of this motivates the stochastic realization approach to the problem taken in § 3.

Finally we remark that this theory can be extended in a straightforward manner to the case when the system matrices are time-varying replacing the concept of invariant direction by that of *degenerate* direction [14]. A reduction of the order of the Riccati equation which has to be solved can be achieved along the lines of [8] whenever invariant (or degenerate) directions exist.

2. Discrete time stochastic realization: General theory.

2.1. Notation and problem formulation. Almost sure equality between random vectors is simply indicated as equality. If $\{\xi(t);\, t\in \mathsf{Z}\}$ is a second order vector process defined on the probability space $(\Omega, \mathscr{F}, P)$ and S a subset of the integers Z, we denote by $H_S(\xi)$ the closed linear hull in $L_2(\Omega, \mathscr{F}, P)$ of the components of $\xi(t)$, $t\in S$. We shall write $H(\xi)$, $H_t^-(\xi)$, $H_t^+(\xi)$ and $H(\xi(t))$ instead of $H_{\mathsf{Z}}(\xi)$, $H_{\{z\in \mathsf{Z}|z\leqq t\}}(\xi)$, $H_{\{z\in\mathsf{Z}|z\geqq t\}}(\xi)$ and $H_{\{t\}}(\xi)$ respectively. Let $\hat{E}\{\cdot|H_S(\xi)\}$ denote the orthogonal projection operator onto $H_S(\xi)$. We abbreviate $\hat{E}\{\cdot|H(\xi(t))\}$ as $\hat{E}\{\cdot|\xi(t)\}$. The process ξ is called a *wide sense vector Markov process* if

$$\hat{E}\{\xi(s)|H_t^-(\xi)\}=\hat{E}\{\xi(s)|\xi(t)\}\quad \text{for } s\geqq t,$$

or equivalently

$$\hat{E}\{\xi(s)|H_t^+(\xi)\}=\hat{E}\{\xi(s)|\xi(t)\}\quad \text{for } s\leqq t.$$

For the sake of brevity we shall use the word "Markov" instead of the expression "wide sense vector Markov."

We shall be concerned with a wide sense stationary, purely nondeterministic, m-dimensional stochastic process $\{z(t);\, t\in\mathsf{Z}\}$. The process z, defined on the probability space $(\Omega, \mathscr{F}, P)$, is assumed to be centered and to have a rational spectral density Φ such that $\Phi(\infty)<\infty$. The finiteness of $\Phi(\infty)$ is essential only in § 3 and is assumed here for simplicity. The matrix function $\Phi(\cdot)$ enjoys the following properties: each element of Φ is analytic on the unit circle, Φ is discrete para-Hermitian, i.e., $\Phi(z)'=\Phi(z^{-1})$ and $\Phi(e^{i\omega})\geqq 0$ Hermitian for all real ω. In addition we suppose that z is a *minimal process* [31] which, in view of the rationality of its spectral density, is equivalent to $\Phi(e^{i\omega})>0$ for all ω. This assumption too is made for convenience and can be removed without impairing the main results of §§ 2 and 3.

In many problems of estimation and optimal control, when given a non-Markov process z which models the information flow, it is necessary to resort to an auxiliary Markov process x which makes $\xi(t)=\begin{pmatrix} x(t)\\ z(t-1)\end{pmatrix}$ a Markov process. More precisely we are interested in the following two problems.

I. *Wide sense stochastic realization problem.* Determine, from the knowledge of Φ, all quadruplets $[A, B, C, D]$, with dimension of A minimal, such that the process y, generated by the dynamical system (1.1)–(1.2) driven by an arbitrary normalized white noise w, has the same spectral density Φ as z.

II. *Proper stochastic realization problem.* Let H be a Hilbert space such that $H(z)\subset H\subset L_2(\Omega, \mathscr{F}, P)$. Given H and the process z find all quintuplets $[A, B, C, D; w]$, with dimension of A minimal and w a normalized white noise satisfying $H(w)\subset H$, such that $y(t)$, generated by (1.1)–(1.2) and $z(t)$ are equivalent random vectors for all t.

We shall call a solution to problem I a *wide sense minimal stochastic realization* and a solution to problem II a *proper minimal*[1] *stochastic realization.* It is immediate that to

[1] From now on we shall leave the word minimal out. All realizations are to be intended to be minimal unless the opposite is explicitly stated.

each proper stochastic realization there corresponds a (unique) wide sense realization. The converse is false. To attack problem II we shall choose a route passing through the solution of problem I, with the intent of deriving some new results along the way. It is good to bear in mind, however, that a direct probabilistic approach to proper stochastic realization is possible and in a sense more natural [18], [20]–[22], [27], [35], [36].

2.2. Wide sense stochastic realizations. Our preliminaries on problem I are based on the important work of B. D. O. Anderson [3]–[5] and Faurre [11], [12]. Problem I is equivalent to the classical *spectral factorization problem.* Find all *minimal stable spectral factors* of Φ, i.e., all matrices W of real rational functions of minimal McMillan degree [6] and with all their poles inside the unit circle which satisfy (1.33). Indeed, if $[A, B, C, D]$ solves problem I, then $W(z)=C(zI-A)^{-1}B+D$ is a stable minimal spectral factor of Φ. Conversely, any such W yields a whole class of wide sense stochastic realizations. In fact, using one of the algorithms [16], [39], [41] available in the literature we can compute a minimal [6] realization $[A, B, C, D]$ of W. Then all minimal realizations of W given by

$$[T^{-1}AT, T^{-1}B, CT, D], \qquad T\in GL_{\dim A}(\mathrm{R}) \tag{2.1}$$

solve problem I. In view of this equivalence problem I can be solved as follows. Express Φ, by means of partial fractions, as

$$\Phi(z)=S(z)+S(z^{-1})', \tag{2.2}$$

where S is a *positive real*[2] and rational function. Let $[F, G, H, J]$ be a minimal realization of S. As observed before, several procedures are known to determine $[F, G, H, J]$ which is unique up to an equivalence such as in (2.1). The following simple lemma allows us to eliminate J in the sequel.

LEMMA 2.1. *Let S be the positive real function satisfying* (2.2) *and* $[F, G, H, J]$ *a minimal realization of S. If* $\dim F=n\geqq 1$, *then F is nonsingular and* $J+J'=G'(F')^{-1}H'+\Phi(\infty)$.

Proof. Taking limits in (2.2) we see that $\Phi(\infty)=J+J'+\lim_{z\to\infty} G'(z^{-1}I-F')^{-1}H'$, since $S(z)=H(zI-F)^{-1}G+J$. The conclusion now follows from the finiteness of $\Phi(\infty)$ and the minimality of $[F, G, H, J]$. □

To avoid trivialities, we shall assume from now on that z is not a white noise, i.e., $\dim F=n\geqq 1$. It follows from Lemma 2.1 and the celebrated positive real lemma (see e.g., [28]) that the set of all wide sense stochastic realizations is nonempty and given by

$$[A, B, C, D]=[T^{-1}FT, T^{-1}(B_1, B_2)V, HT, (R(P)^{1/2}, 0)V], \tag{2.3}$$

where $T\in GL_n(\mathrm{R})$, V is any $p\times p$ constant orthogonal matrix, B_1 is $n\times m$, B_2 is $n\times(p-m)$ (here $p\geqq m$ is arbitrary), P is $n\times n$, symmetric and positive definite, $R(P)$ is the nonnegative definite quantity $G'(F')^{-1}H'+\Phi(\infty)-HPH'$ and (P, B_1, B_2) solve the system

$$P=FPF'+B_1B_1'+B_2B_2', \tag{2.4}$$

$$G=FPH'+B_1R(P)^{1/2}. \tag{2.5}$$

[2] A real rational function with no pole on the unit circle is said to be (discrete) positive real if it has no poles outside the unit circle and $S(e^{i\omega})+S(e^{-i\omega})'\geqq 0$ Hermitian for all real ω.

It is no restriction to choose $T = I$ and $V = I$ in (2.3). In fact all other realizations can be obtained from realizations of the form

$$x(t+1) = Fx(t) + B_1u(t) + B_2v(t), \qquad w = \binom{u}{v}, \tag{2.6}$$

$$z(t) = Hx(t) + R(P)^{1/2}u(t) \tag{2.7}$$

by means of a change of basis and an orthogonal transformation of w. Hence, whenever convenient, we shall narrow our attention to realizations of the type (2.6)–(2.7). We shall write $\mathcal{P}$ for the set of all symmetric, positive definite P which solve (2.4)–(2.5) and $\mathcal{Q}$ for the subset of $\mathcal{P}$ consisting of those P such that $R(P)$ is singular. Notice that the realizations corresponding to elements of $\mathcal{Q}$ are precisely those which have singular intensity of the noise in the output equation. It can be shown [12] that $\mathcal{P}$ is compact, convex and forms a complete lattice when endowed with the natural partial order $P_1 \geqq P_2$ if and only if $P_1 - P_2 \geqq 0$. There exist a maximal and a minimal element P^* and P_* so that $P_* \leqq P \leqq P^*$ for all $P \in \mathcal{P}$. Moreover the minimality of the process z implies [13] that $P^* - P_*$ and $R(P_*)$ are positive definite. Hence $\mathcal{P}\backslash\mathcal{Q}$ is nonempty. The following result provides us with some information about the set $\mathcal{Q}$.

PROPOSITION 2.2. *The set $\mathcal{P}\backslash\mathcal{Q}$ is convex. For all $P \in \mathcal{P}\backslash\mathcal{Q}$, $Q \in \mathcal{Q}$ and $\lambda \in (0, 1]$ we have that $[\lambda P + (1-\lambda)Q] \in \mathcal{P}\backslash\mathcal{Q}$. The set $\mathcal{Q}$ is contained in the relative boundary of $\mathcal{P}$.*

Proof. The first two results follow at once from the fact that for $P_1, P_2 \in \mathcal{P}$, $\lambda \in [0, 1]$ we have $R(\lambda P_1 + (1-\lambda)P_2) = \lambda R(P_1) + (1-\lambda)R(P_2)$. They in turn imply that, if $P \in \mathcal{P}\backslash\mathcal{Q}$ and $Q \in \mathcal{Q}$, the segment $[P, Q]$ cannot be extended beyond Q without leaving $\mathcal{P}$. We conclude that Q belongs to the relative boundary of $\mathcal{P}$. □

Let us introduce the mapping $\Lambda: \mathrm{R}^{n\times n} \rightarrow \mathrm{R}^{n\times n}$ defined by

$$\Lambda(P) = -P + FPF' + (G - FPH')R(P)^{-1}(G' - HPF'). \tag{2.8}$$

The set $\mathcal{P}/\mathcal{Q}$ is contained in the domain of $\Lambda(\cdot)$. It is possible to extend $\Lambda(\cdot)$ to all of $\mathcal{P}$ since the points in $\mathcal{Q}$ constitute removable discontinuities. We can now derive an important alternative characterization of the set $\mathcal{P}$.

THEOREM 2.3. *Let $\Lambda(\cdot)$ be given by* (2.8). *Then $\mathcal{P} = \{P | P = P', \Lambda(P) \leqq 0\}$.*

Proof. Let (P, B_1, B_2) solve (2.5)–(2.6) with $P = P'$ and $P > 0$. Then if $P \in \mathcal{P}\backslash\mathcal{Q}$, we get immediately $\Lambda(P) = -B_2B_2'$. If $P \in \mathcal{Q}$, let $\{P_i\}_{i=1}^{\infty}$ be a sequence in $\mathcal{P}\backslash\mathcal{Q}$ converging to P. Then $\Lambda(P_i) \leqq 0$ and it follows that $\Lambda(P) = \lim_i \Lambda(P_i) \leqq 0$. This shows that $\mathcal{P} \subseteq \{P | P = P', \Lambda(P) \leqq 0\}$. The other inclusion can be proven by an argument akin to that used by B. D. O. Anderson [4, p. 140]. □

This result provides a bridge between the theory of positive real functions and the study of quadratic matrix inequalities and algebraic Riccati equations.

Let us introduce the set $\mathcal{P}_0 = \{P \in \mathcal{P} | \Lambda(P) = 0\}$. Clearly $\mathcal{P}_0$ consists of all $P \in \mathcal{P}$ for which $B_2 = 0$.

Remark 2.4. Since the eigenvalues of F lie in the open unit disc, elementary Lyapunov theory ensures that to each (B_1, B_2) there corresponds a unique P. The converse does not hold in general. However, for realizations of the form (2.6)–(2.7), to each P there corresponds a unique B_1. This is immediate from (2.5) for $P \in \mathcal{P}\backslash\mathcal{Q}$ and holds for all $P \in \mathcal{P}$ since points in $\mathcal{Q}$ appear as removable discontinuities of the map $P \rightarrow (G - FPH')R(P)^{-1/2}$. Hence there is a unique wide sense realization of the type (2.6)–(2.7) corresponding to each P in $\mathcal{P}_0$.

Both problems I and II seek to find dynamical systems evolving forward in time like (1.1)–(1.2) which is natural to call *forward representations* of the process z. Yet, there are other representations of interest. There exist situations, for example, in which it is more useful to consider a *backward representation* of the form

$$\bar{x}(t-1)=\bar{A}\bar{x}(t)+\bar{B}\bar{w}(t), \tag{2.9}$$

$$y(t)=\bar{C}\bar{x}(t)+\bar{D}\bar{w}(t), \tag{2.10}$$

where $\bar{w}$ is a normalized white noise such that $\bar{w}(t)$ is orthogonal to $H_t^+(\bar{x})$ for all t. This leads us to formulate the backward counterpart of problems I and II.

$\bar{\mathrm{I}}$. *Wide sense backward stochastic realization problem.* Determine, from the knowledge of Φ, all quadruplets $[\bar{A}, \bar{B}, \bar{C}, \bar{D}]$, with dimension of $\bar{A}$ minimal, such that the process y, generated by the dynamical system (2.9)–(2.10) driven by an arbitrary normalized white noise $\bar{w}$, has the same spectral density Φ as z.

$\bar{\mathrm{II}}$. *Proper backward stochastic realization problem.* Given H and z find all quintuplets $[\bar{A}, \bar{B}, \bar{C}, \bar{D}; \bar{w}]$, with dimension of $\bar{A}$ minimal and $\bar{w}$ a normalized white noise satisfying $H(\bar{w})\subset H$, such that $y(t)$ given by (2.9)–(2.10) and $z(t)$ are equivalent random vectors for all t.

Solutions to problems $\bar{\mathrm{I}}$ and $\bar{\mathrm{II}}$ are called *wide sense* and *proper backward stochastic realizations* respectively. We shall now briefly discuss problem $\bar{\mathrm{I}}$, while problem $\bar{\mathrm{II}}$ will be implicitly solved in the next three sections in view of Theorem 2.5 below.

Problem $\bar{\mathrm{I}}$ is equivalent to the *dual spectral factorization problem* considered by Anderson [3] and Faurre [12] which consists in finding all minimal unstable (i.e., with all the poles outside the unit circle) spectral factors $\bar{W}(z)$ of $\Phi(z)$. It follows from the para-Hermitian property of Φ that this problem is equivalent to the spectral factorization problem for $\Phi(\cdot)'$. Hence all the results on problem I have a natural counterpart in the backward setting via the duality relation $(F, G, H, \Phi(\infty))\rightarrow(F', H', G', \Phi(\infty)')$. In particular all solutions to problem $\bar{\mathrm{I}}$ are characterized by

$$[\bar{A}, \bar{B}, \bar{C}, \bar{D}]=[T^{-1}F'T, T^{-1}(\bar{B}_1, \bar{B}_2)V, G'T, (\bar{R}(\bar{P})^{1/2}, 0)V], \tag{2.11}$$

where T and V are as in (2.3), $\bar{B}_1$ is $n\times m$, $\bar{B}_2$ is $n\times(p-m)$, $\bar{P}$ is $n\times n$, symmetric and positive definite, $\bar{R}(\bar{P})=HF^{-1}G+\Phi(\infty)'-G'\bar{P}G$ and $(\bar{P}, \bar{B}_1, \bar{B}_2)$ solve the system

$$\bar{P}=F'\bar{P}F+\bar{B}_1\bar{B}_1'+\bar{B}_2\bar{B}_2', \tag{2.12}$$

$$H'=F'\bar{P}G+\bar{B}_1\bar{R}(\bar{P})^{1/2}. \tag{2.13}$$

Whenever it is appropriate, we shall restrict ourselves to realizations of the type

$$\bar{x}(t-1)=F'\bar{x}(t)+\bar{B}_1\bar{u}(t)+\bar{B}_2\bar{v}(t), \qquad \bar{w}=\begin{pmatrix}\bar{u}\\ \bar{v}\end{pmatrix}, \tag{2.14}$$

$$z(t)=G'\bar{x}(t)+\bar{R}(\bar{P})^{1/2}\bar{u}(t), \tag{2.15}$$

where $\bar{P}$ is the state covariance. The set $\bar{\mathscr{P}}$ of all symmetric, positive definite solutions to (2.12)–(2.13) and $\bar{\mathscr{Q}}$ of all $\bar{P}\in\bar{\mathscr{P}}$ such that $\bar{R}(\bar{P})$ is singular enjoy the same kind of properties as $\mathscr{P}$ and $\mathscr{Q}$ respectively. In particular there exist $\bar{P}_*$ and $\bar{P}^*$ such that $\bar{P}_*\leqq\bar{P}\leqq\bar{P}^*$ for all $\bar{P}\in\bar{\mathscr{P}}$. It is well known [12], [37] that $\bar{\mathscr{P}}=\{P^{-1}|P\in\mathscr{P}\}$, so that $\bar{P}_*=(P^*)^{-1}$ and $\bar{P}^*=(P_*)^{-1}$. Indeed, the following result holds.

PROPOSITION 2.5. *The quadruplet* $[F, B, H, D]$ *with* $B=(B_1, B_2)$ *and* $D=(R(P)^{1/2}, 0)$ *solves problem* I *if and only if* $[F', \bar{B}, G', \bar{D}]$ *solves problem* $\bar{\mathrm{I}}$ *where*

$$\bar{B}=(\bar{B}_1, \bar{B}_2)=-P^{-1}F^{-1}B(I-B'P^{-1}B)^{1/2}, \tag{2.16}$$

$$\begin{aligned}\bar{D}&=(D-HF^{-1}B)(I-B'P^{-1}B)^{1/2}\\ &=(R(P)^{1/2}-HF^{-1}B_1, -HF^{-1}B_2)(I-B'P^{-1}B)^{1/2}.\end{aligned} \tag{2.17}$$

Proof. The result follows from long but simple calculations using (2.4)–(2.5) and (2.12)–(2.13). □

This proposition exhibits a correspondence between forward and backward wide sense realizations and raises the question whether a result of the same type can be established for proper realizations. We turn to this problem in the beginning of the next section.

2.3. Proper stochastic realizations. Let us consider a proper stochastic realization of z $[F, B, H, D; w]$, with state process x and state covariance P. As is well known, the orthogonal decomposition

$$x(t+1)=\hat{E}\{x(t+1)|H_t^-(x)\}+[x(t+1)-\hat{E}\{x(t+1)|H_t^-(x)\}]$$

yields (2.6). Similarly the expression

$$x(t)=\hat{E}\{x(t)|H_{t+1}^+(x)\}+[x(t)-\hat{E}\{x(t)|H_{t+1}^+(x)\}] \tag{2.18}$$

leads to a backward model. In fact, the process x is Markov in both directions and

$$\begin{aligned}\hat{E}\{x(t)|x(t+1)\}&=E\{x(t)x(t)'F'\}E\{x(t+1)x(t+1)'\}^{-1}x(t+1)\\&=PF'P^{-1}x(t+1),\end{aligned}$$

which gives

$$\begin{aligned}P^{-1}x(t)&=F'P^{-1}x(t+1)-F'P^{-1}B[w(t)-B'(F')^{-1}P^{-1}x(t)]\\&=F'P^{-1}x(t+1)-P^{-1}F^{-1}B[I-B'P^{-1}B][w(t)-B'(F')^{-1}P^{-1}x(t)].\end{aligned}$$

Defining

$$\bar{x}(t)=P^{-1}x(t+1) \tag{2.19}$$

and

$$\bar{w}(t)=(I-B'P^{-1}B)^{1/2}(w(t)-B'(F')^{-1}P^{-1}x(t)), \tag{2.20}$$

we finally obtain

$$\bar{x}(t-1)=F'\bar{x}(t)-P^{-1}F^{-1}B(I-B'P^{-1}B)^{1/2}\bar{w}(t). \tag{2.21}$$

It is not difficult to check that $\bar{w}$ is a normalized white noise such that $\bar{w}(t)$ is orthogonal to $H_t^+(\bar{x})$ for all t. The forward and backward noises are related as follows

$$(I-B'P^{-1}B)^{1/2}\bar{w}(t)=w(t)-\hat{E}\{w(t)|H_{t+1}^+(x)\}. \tag{2.22}$$

We also have

$$\begin{aligned}z(t)=Hx(t)+Dw(t)&=[G'(F')^{-1}P^{-1}-DB'(F')^{-1}P^{-1}]x(t)+Dw(t)\\&=G'P^{-1}x(t+1)+[D-G'P^{-1}B][w(t)-B'(F')^{-1}P^{-1}x(t)]\\&=G'\bar{x}(t)+[D-HF^{-1}B][I-B'P^{-1}B]^{1/2}\bar{w}(t).\end{aligned}$$

Summing up we obtain a strict sense version of Proposition 2.5, analogous to the continuous time result of Lindquist and Picci [19].

THEOREM 2.5. *The quintuplet* $[F, B, H, D; w]$ *is a proper (forward) stochastic realization of* z *with state process* x *and state covariance* P *if and only if the quintuplet* $[F', \bar{B}, G', \bar{D}; \bar{w}]$ *is a proper backward stochastic realization of* z *with state process* $\bar{x}$ *given by* (2.19) *and state covariance* P^{-1}, *where* $\bar{w}$ *is as in* (2.20) *and* $\bar{B}, \bar{D}$ *are given by* (2.16)–(2.17).

Results closely related to this theorem have been presented by Akaike [1, p. 168] and Ruckebusch [33, p. 32]. However, the first deals with realizations without noise in the observations, the second does not derive expressions for $\bar{w}$, $\bar{B}$ and $\bar{D}$ such as (2.20), (2.16) and (2.17).

So far we have said nothing about *existence* of proper stochastic realizations. It is well-known that a necessary and sufficient condition for a purely nondeterministic wide sense stationary process z to admit finite dimensional stochastic realizations is that its spectral density is rational and that in such a case there exists a unique realization of the type (2.6)–(2.7) corresponding to P_* (cf. [33] for example). The minimum variance realization

$$x_*(t+1) = Fx_*(t) + B_*u_*(t), \tag{2.23}$$

$$z(t) = Hx_*(t) + R(P_*)^{1/2}u_*(t) \tag{2.24}$$

is the *steady-state Kalman filter*, with the *steady-state Kalman gain* B_* given by

$$B_* = (F\Sigma H' + BD')(H\Sigma H' + DD')^{-1/2} = (G - FP_*H')R(P_*)^{-1/2}, \tag{2.25}$$

where $[F, B, H, D]$ is any wide sense realization and Σ is the unique nonnegative definite solution to the *algebraic Riccati equation*

$$\Sigma = F\Sigma F' - (F\Sigma H' + BD)(H\Sigma H' + DD')^{-1}(H\Sigma F' + DB') + BB'. \tag{2.26}$$

The noise μ_* is called the *innovation process* and is characterized by the fact that $H_t^-(z) = H_t^-(u_*)$ for all $t \in \mathrm{Z}$. Finally, if x is the state process of any proper realization (2.6)–(2.7), we have

$$x_*(t) = \hat{E}\{x(t)|H_{t-1}^-(z)\}. \tag{2.27}$$

By duality there exists a proper backward stochastic realization corresponding to $\bar{P}_*$, namely the *backward steady-state Kalman filter*

$$\bar{x}_*(t-1) = F'\bar{x}_*(t) + \bar{B}_*\bar{u}_*(t), \tag{2.28}$$

$$z(t) = G'\bar{x}_*(t) + \bar{R}(\bar{P}_*)^{1/2}\bar{u}_*(t). \tag{2.29}$$

Here the *backward steady-state Kalman gain* $\bar{B}_*$ is given by

$$\bar{B}_* = (F'\bar{\Sigma}G + \bar{B}\bar{D}')(G'\bar{\Sigma}G + \bar{D}\bar{D}')^{-1/2} = (H' - F'\bar{P}_*G)\bar{R}(\bar{P}_*)^{-1/2}, \tag{2.30}$$

where $[F', \bar{B}, G', \bar{D}]$ is any backward wide sense realization and $\bar{\Sigma}$ is the unique nonnegative definite solution to the *dual algebraic Riccati equation*

$$\bar{\Sigma} = F'\bar{\Sigma}F - (F'\bar{\Sigma}G + \bar{B}\bar{D}')(G'\Sigma G + \bar{D}\bar{D}')^{-1}(G'\bar{\Sigma}F + \bar{D}\bar{B}') + \bar{B}\bar{B}'. \tag{2.31}$$

The equality $H_t^+(z) = H_t^+(\bar{u}_*)$ for all $t \in \mathrm{Z}$ characterizes the *backward innovation process* $\bar{u}_*$. The backward filter satisfies

$$\bar{x}_*(t) = \hat{E}\{\bar{x}(t)|H_{t+1}^+(z)\}, \tag{2.32}$$

where $\bar{x}$ is the state of any proper backward realization (2.14)–(2.15). By Theorem 2.5 there exists a proper stochastic realization corresponding to (2.28)–(2.29) (which, as it will be apparent in the next section, is unique)

$$x^*(t+1) = Fx^*(t) + B^*u^*(t), \tag{2.33}$$

$$z(t) = Hx^*(t) + R(P^*)^{1/2}u^*(t) \tag{2.34}$$

with state covariance P^*. Then, if x is the state process of any realization,

$$\bar{x}_*(t) = (P^*)^{-1}x^*(t+1) = P^{-1}\hat{E}\{x(t+1)|H^+_{t+1}(z)\} \tag{2.35}$$

and

$$\hat{E}\{x(t)|H^+_t(z)\} = P(P^*)^{-1}x^*(t). \tag{2.36}$$

This justifies our choice of working with $P^{-1}x$ rather than x in the backward setting. In fact (2.36) is *not* invariant over $\mathcal{P}$.

DEFINITION 2.6 ([19], [33]). A proper stochastic realization of z with state process x is said to be *internal* if $H(x) \subseteq H(z)$, *external* otherwise.

Internal realizations are of particular interest since they are the only ones we can construct from the process z. For example, the minimum and maximum variance realizations introduced in this section are internal. It should be noted that the existence of external realizations depends on H. If $H = H(z)$, for instance, all realizations would be internal.

2.4. Characterization of internal realizations. Let us consider the spectral representation of z (see e.g. [31]) given by

$$z(t) = \int_{-\pi}^{\pi} e^{i\omega t}\, d\hat{z}(\omega),$$

where $d\hat{z}$ is an orthogonal stochastic measure such that

$$E\{d\hat{z}(\omega)\, d\hat{z}(\omega)\dagger\} = \frac{\Phi(e^{i\omega})}{2\pi}\, d\omega.$$

(Here † denotes complex conjugation and transposition.) Let $W(z) = H(zI-F)^{-1}B_1 + R(P)^{1/2}$ be a square ($m \times m$) spectral factor of $\Phi(z)$. Then the process u, defined by

$$u(t) = \int_{-\pi}^{\pi} e^{i\omega t}[W(e^{i\omega})]^{-1}\, d\hat{z}(\omega), \tag{2.37}$$

is a normalized white noise such that $u(t) \in H(z)$ for all t [31, p. 41] and consequently $[F, B_1, H, R(P)^{1/2}; u]$ is an internal realization of z. The following result shows that $W(\cdot)$ being a square matrix function is also necessary for a realization to be internal.

THEOREM 2.7 ([19], [33]). *A proper stochastic realization is internal if and only if its transfer function is square.*

It follows from this theorem and Remark 2.4 that internal realizations of the form (2.6)–(2.7) are in one to one correspondence with the real symmetric solutions of the matrix equation $\Lambda(P) = 0$. Hence, to characterize further internal realizations, one could derive the discrete time counterpart of the fundamental results of J. C. Willems [40] on the algebraic Riccati equation. However, a result akin to the classification of the solutions of the algebraic Riccati equation can be obtained directly for the state processes of internal realizations. Notice that once the state $x(t)$ of an internal realization has been determined the input $u(t)$ can be obtained inverting (2.9) as follows:

$$u(t) = -R(P)^{-1/2}Hx(t) + R(P)^{-1/2}z(t).$$

(In the case when $R(P)$ is singular we need to perform an appropriate number of differencing operations on the output in various directions (cf. [7] for example) before we can express u in terms of x and z.)

Therefore we turn to the problem of characterizing the state process of internal realizations.

Let us introduce the feedback matrix

$$\Gamma_* = F - B_* R(P_*)^{-1/2} H.$$

The matrix Γ_* is asymptotically stable due to the minimality of z [13]. It plays a central role in stochastic realization theory, as it is clear from what follows. In particular we have the following important result, whose continuous time counterpart can be found in [19].

THEOREM 2.8 ([33]). *The process x is the state of an internal realization if and only if*

$$x(t) = [I - \pi_s] x_*(t) + \pi_s x_*(t), \tag{2.38}$$

where π_s is the projection onto an invariant subspace S of Γ_* along $(P^* - P_*)S^\perp$. The covariance P of x and π_s are related as follows

$$\pi_s = \pi(P) = (P - P_*)(P^* - P_*)^{-1}. \tag{2.39}$$

We shall give a new proof of this theorem, by means of an approach which allows us to characterize also the external realizations in the same framework. Our derivation hinges on the following simple observation. Let $[F, (B_1, B_2), H, R(P)^{1/2}; w]$ be a proper stochastic realization of z with state process x and state covariance P. Then $[\Gamma_*, (B_1 - B_* R(P_*)^{-1/2} R(P)^{1/2}, B_2), R(P_*)^{-1/2} H, R(P_*)^{-1/2} R(P)^{1/2}; w]$ is a proper (nonminimal) stochastic realization of the innovation process u_* with state process $x - x_*$ and state covariance $\tilde{P} = P - P_*$. This can be seen by inverting the filter (2.23)–(2.24) to get

$$x_*(t+1) = \Gamma_* x_*(t) + B_* R(P_*)^{-1/2} z(t), \tag{2.40}$$

$$u_*(t) = -R(P_*)^{-1/2} H x_*(t) + R(P_*)^{-1/2} z(t) \tag{2.41}$$

and by using (2.6)–(2.7). If we set $\tilde{x}(t) = x(t) - x_*(t)$, we obtain the model

$$\tilde{x}(t+1) = \Gamma_* \tilde{x}(t) + (B_1 - B_* R(P_*)^{-1/2} R(P)^{1/2}) u(t) + B_2 v(t), \qquad w = \begin{pmatrix} u \\ v \end{pmatrix}, \tag{2.42}$$

$$u_*(t) = R(P_*)^{-1/2} H \tilde{x}(t) + R(P_*)^{-1/2} R(P)^{1/2} u(t), \tag{2.43}$$

which is a forward stochastic realization of u_* since $w(t) \perp H_t^-(\tilde{x})$ for all t. The representation (2.42)–(2.43) is not minimal since u_* is a white noise and its minimal realizations have dimension zero. Conversely consider a forward stochastic realization of u_* of the form

$$\xi(t+1) = \Gamma_* \xi(t) + \tilde{B}_1 u(t) + \tilde{B}_2 v(t), \qquad w = \begin{pmatrix} u \\ v \end{pmatrix}, \tag{2.44}$$

$$u_*(t) = R(P_*)^{-1/2} [H \xi(t) + R(\tilde{P})^{1/2} u(t)], \tag{2.45}$$

where w is a normalized white noise and $\tilde{B}_1$ is $n \times m$. Observe that $w(t)$ is orthogonal to $H_t^-(x)$, where $x = \xi + x_*$, since $x_*(t) \in H_{t-1}^-(z) = H_{t-1}^-(u_*)$. We conclude from this that $[F, (\tilde{B}_1 + B_* R(\tilde{P})^{1/2}, \tilde{B}_2), H, R(P_*)^{1/2} R(\tilde{P})^{1/2}; w]$ is a minimal stochastic realization of z. We collect these observations in the following

LEMMA 2.9. *The map which sends the realization $[F, (B_1, B_2), H, R(P)^{1/2}; w]$ to the realization $[\Gamma_*, (B_1 - B_* R(P_*)^{-1/2} R(P)^{1/2}, B_2), R(P_*)^{-1/2} H, R(P_*)^{-1/2} R(P)^{1/2}; w]$ is a one to one correspondence between realizations of z of the form* (2.6)–(2.7) *and realizations of u_* of the form* (2.44)–(2.45).

The map in Lemma 2.9 also induces a correspondence between state covariances which maps $P \in \mathcal{P}$ to $P - P_*$, translating the set $\mathcal{P}$ of the amount $-P_*$. The set

$\tilde{\mathcal{P}} = \mathcal{P} - P_*$ has the zero element as its minimum and the positive definite quantity $P^* - P_*$ as its maximum. Notice that the correspondence established in Lemma 2.9 is simply the correspondence between the two input-output relations

$$z(t) = \int_{-\pi}^{\pi} e^{i\omega t} W(e^{i\omega})\, d\hat{w}(\omega)$$

and

$$u_*(t) = \int_{-\pi}^{\pi} e^{i\omega t} W_*^{-1}(e^{i\omega}) W(e^{i\omega})\, d\hat{w}(\omega),$$

where $W(z) = H(zI - F)^{-1}(B_1, B_2) + R(P)^{1/2}$, $d\hat{w}$ is an orthogonal stochastic measure such that $w(t) = \int_{-\pi}^{\pi} e^{i\omega t}\, d\hat{w}(\omega)$ and $W_*(z) = H(zI - F)^{-1}B_* + R(P_*)^{1/2}$.

From (2.23)–(2.24) and (2.40)–(2.41) we know that $H_t^-(z) = H_t^-(u_*)$ for all t and $H(z) = H(u_*)$. Since u_* is a white noise we have the following orthogonal decomposition for the space $H(z)$

(2.46) $$H(z) = H_{t-1}^-(z) \oplus H_t^+(u_*).$$

Then, if x is the state process of an internal realization, we have

$$x(t) = \hat{E}\{x(t) | H(z)\} = \hat{E}\{x(t) | H_{t-1}^-(z)\} + \hat{E}\{x(t) | H_t^+(u_*)\},$$

which implies

(2.47) $$x(t) = x_*(t) + \hat{E}\{x(t) - x_*(t) | H_t^+(u_*)\}$$

in view of (2.27) and the orthogonality between $x_*(t)$ and $H_t^+(u_*)$. To compute $\hat{E}\{x(t) - x_*(t) | H_t^+(u_*)\}$ observe first that $\tilde{x}(t) = x(t) - x_*(t)$ is the state process of a realization of u_* of the form (2.41)–(2.42). Secondly, notice that u_* is stochastic process enjoying all the properties of z. Therefore we simply derive relation (2.26) with $\tilde{x}$ and u_* in place of x and z respectively. This idea of replacing a stochastic process by its innovations is of course very common in filtering theory and it turns out to be helpful also in our context.

We shall now derive the backward counterpart of a realization of the type (2.42)–(2.43) corresponding to an internal realization. We set $B_2 = 0$ in (2.42)–(2.43) and define $\tilde{P} = P - P_*$. An orthogonal decomposition for $\tilde{x}(t)$ as in (2.18) yields the identity

(2.48) $$\tilde{x}(t) = \tilde{P}\Gamma'_*\tilde{P}^{\#}\tilde{x}(t+1) + [\tilde{x}(t) - \tilde{P}\Gamma'_*\tilde{P}^{\#}\tilde{x}(t+1)].$$

Observe that $\tilde{x}(t) - \tilde{P}\Gamma'_*\tilde{P}^{\#}\tilde{x}(t+1)$ is orthogonal to $H_{t-1}^-(z)$. Also, using (2.42)–(2.43), we see that $\hat{E}\{\tilde{x}(t) - \tilde{P}\Gamma'_*\tilde{P}^{\#}\tilde{x}(t+1) | H_{t+1}^+(u_*)\} = 0$. Hence, using (2.46), we have

$$\begin{aligned}\tilde{x}(t) - \tilde{P}\Gamma'_*\tilde{P}^{\#}\tilde{x}(t+1) &= \hat{E}\{\tilde{x}(t) - \tilde{P}\Gamma'_*\tilde{P}^{\#}\tilde{x}(t+1) | u_*(t)\} \\ &= \hat{E}\{\tilde{x}(t) | u_*(t)\} = \tilde{P}H'R(P_*)^{-1/2}u_*(t)\end{aligned}$$

and (2.48) becomes

(2.49) $$\tilde{x}(t) = \tilde{P}\Gamma'_*\tilde{P}^{\#}\tilde{x}(t+1) + \tilde{P}H'R(P_*)^{-1/2}u_*(t)$$

or

(2.50) $$\tilde{P}^{\#}\tilde{x}(t) = \tilde{P}^{\#}\tilde{P}\Gamma'_*\tilde{P}^{\#}\tilde{x}(t+1) + \tilde{P}^{\#}\tilde{P}H'R(P_*)^{-1/2}u_*(t).$$

The output simply reads

(2.51) $$u_*(t) = 0\tilde{P}^{\#}\tilde{x}(t+1) + u_*(t)$$

where 0 is the $m \times n$ zero matrix. The model (2.50)–(2.51) is the backward counterpart of (2.42)–(2.43). We stress the fact that all backward realizations of the innovations which we obtain in this fashion from realizations (2.42)–(2.43) with $B_2 = 0$ have the same input noise u_*. For $\tilde{x} = x^* - x_*$ we obtain the backward filter

$$\begin{aligned}(P^* - P_*)^{-1}(x^*(t) - x_*(t)) &= \Gamma'_*(P^* - P_*)^{-1}(x^*(t+1) - x_*(t+1)) \\ &\quad + H'R(P_*)^{-1/2}u_*(t).\end{aligned} \tag{2.52}$$

Using alternatively (2.42) and (2.49) to compute $E\{\tilde{x}(t+1)\tilde{x}(t)'\}$ we establish the identity $\Gamma_*\tilde{P} = \tilde{P}\tilde{P}^{\#}\Gamma_*\tilde{P}$ which gives

$$\tilde{P}\Gamma'_* = \tilde{P}\Gamma'_*\tilde{P}^{\#}\tilde{P}. \tag{2.53}$$

Then, using (2.49) and (2.53) we obtain

$$\tilde{x}(t) = \tilde{P}\sum_{i=0}^{\infty}(\Gamma'_*)^i H'R(P_*)^{-1/2}u_*(t+i)$$

which, together with (2.52), yields the desired expression

$$x(t) = x_*(t) + (P - P_*)(P^* - P_*)^{-1}(x^*(t) - x_*(t)). \tag{2.54}$$

Hence $\tilde{x}(t) \in H(x^*(t) - x_*(t))$ and (2.54) can be written

$$\begin{aligned}\tilde{x}(t) &= \hat{E}\{\tilde{x}(t) | x^*(t) - x_*(t)\} \\ &= (P - P_*)(P^* - P_*)^{-1}(x^*(t) - x_*(t))\end{aligned} \tag{2.55}$$

from which it is seen that $\pi(P) = (P - P_*)(P^* - P_*)^{-1}$ is a projection. Rewriting (2.53) in the form

$$\Gamma_*\pi(P) = \pi(P)(P^* - P_*)\tilde{P}^{\#}\Gamma_*\pi(P),$$

we see that $\pi(P)$ projects onto an invariant subspace of Γ_*. Since $\pi(P)(P^* - P_*) = (P^* - P_*)\pi(P)'$ and $\pi(P)'$ projects along $S^{\perp}$ [15, p. 61], we conclude that $\pi(P)$ projects parallel to $(P^* - P_*)S^{\perp}$. Conversely if π projects onto an invariant subspace of Γ_* and $\pi(P^* - P_*) = (P^* - P_*)\pi'$, i.e., π is an *admissible projection* in Ruckebush's language, it is easy to construct first a realization of the innovations and then one (internal) of z along the same lines as in [33]. This completes the proof of Theorem 2.8. □

Remark 2.10. Notice that, given the special form of the realization (2.50)–(2.51), we did not need to invoke any invariance property such as (2.32) of the filter (2.52) to compute $\hat{E}\{\tilde{x}(t) | H_t^+(u_*)\}$. The following interpretation for Theorem 2.8 emerged in the proof. The state process of an internal realization of z is given by the forward filter of z plus a "piece" of the maximum variance error $x^*(t) - x_*(t)$. This piece must be such as to conform with the dynamics of $x^*(t) - x_*(t)$ which is determined by the transition matrix Γ_*, i.e., it must correspond to an invariant subspace of Γ_*.

2.5. External realizations. It is clear that a necessary condition for the existence of external realizations is the presence in H of elements orthogonal to $H(z)$. For the sake of simplicity we assume that $H = H(z) \oplus H(\zeta)$, where ζ is an n-dimensional normalized white noise orthogonal to $H(z)$. As it will be apparent from what follows, this assumption is the minimum one needed to guarantee the existence of a proper stochastic realization corresponding to each wide sense stochastic realization.

Let x be the state process of a realization (2.6)–(2.7) and P its covariance. Then the counterpart of (2.47) is

$$\tilde{x}(t) = x_*(t) + \hat{E}\{\tilde{x}(t) | H_t^+(u_*)\} + \hat{E}\{\tilde{x}(t) | H(\zeta)\} \tag{2.56}$$

and (2.48) corresponds to

$$\tilde{x}(t)=\tilde{P}\Gamma_*'\tilde{P}^{\#}\tilde{x}(t+1)+\tilde{P}H'R(P_*)^{-1/2}u_*(t)+E\{\tilde{x}(t)-\tilde{P}\Gamma_*'\tilde{P}^{\#}\tilde{x}(t+1)|H(\zeta)\}. \tag{2.57}$$

Now let us assume that ζ is chosen in such a way that the condition $H_{t-1}^-(\zeta)\perp H_t^+(\tilde{x})$ holds and ζ and $\tilde{x}$ are stationarily correlated for every realization (2.42)–(2.43). This assumption is introduced to enable us to treat ζ in the same way as the innovations. It will be clear from what follows that indeed this is a natural assumption when trying to model all realizations using a unique exogenous noise. We can now add to (2.42)–(2.43) the output

$$\zeta(t)=M\tilde{x}(t)+[\zeta(t)-M\tilde{x}(t)],$$

where $M=E\{\zeta(t)\tilde{x}(t)'\}\tilde{P}^{\#}$ and an argument very similar to that used for the innovations gives $\hat{E}\{\tilde{x}(t)-\tilde{P}\Gamma_*'\tilde{P}^{\#}\tilde{x}(t+1)|H(\zeta)\}=\tilde{P}M'\zeta(t)$ so that (2.57) becomes

$$\tilde{x}(t)=\tilde{P}\Gamma_*'\tilde{P}^{\#}\tilde{x}(t+1)+\tilde{P}H'R(P_*)^{-1/2}u_*(t)+\tilde{P}M'\zeta(t). \tag{2.58}$$

Note that M must satisfy

$$\tilde{P}=\tilde{P}\Gamma_*'\tilde{P}^{\#}\Gamma_*\tilde{P}+\tilde{P}H'R(P_*)^{-1}H\tilde{P}+\tilde{P}M'M\tilde{P}$$

and that, as in the internal case, the input noise (u_*, ζ) is the same for all realizations.

Let $\tilde{x}_I(t)$ and $\tilde{x}_E(t)$ denote $\hat{E}\{\tilde{x}(t)|H_t^+(u_*)\}$ and $\hat{E}\{\tilde{x}(t)|H(\zeta)\}$ respectively. Then it follows from (2.58) that

$$\tilde{x}_I(t)=(P-P_*)(P^*-P_*)^{-1}(x^*(t)-x_*(t)) \tag{2.59}$$

and

$$\tilde{x}_E(t)=\tilde{P}\Gamma_*'\tilde{P}^{\#}\tilde{x}_E(t+1)+\tilde{P}M'\zeta(t). \tag{2.60}$$

Using (2.53), (2.56), (2.59) and (2.60) we conclude that

$$x(t)=x_*(t)+(P-P_*)(P^*-P_*)^{-1}(x^*(t)-x_*(t))+\sum_{i=0}^{\infty}(\Gamma_*')^iM'\zeta(t+i). \tag{2.61}$$

Conversely, given any matrix M such that $M'M\in\mathscr{C}_n$, let $\tilde{P}$ solve

$$\tilde{P}^{-1}=\Gamma_*'\tilde{P}^{-1}\Gamma_*+H'R(P_*)^{-1}H+M'M.$$

Then, using (2.61), we construct the state of a stochastic realization of z. All the realizations with singular $\tilde{P}$ can be obtained through limiting procedures, using realizations corresponding to unbounded sequences of $M'M$ in the cone $\mathscr{C}_n$.

The derivation of the classification of external realizations presented above is quite similar to the one given in [33, p. 65], but we feel it will give some further insight into the concepts described there. Moreover it provides a clear stochastic meaning for the parametric representation of the set $\mathscr{P}$ derived by Faurre [12, p. 52] in continuous time and by Germain [13, p. 61] in discrete time. Finally the input processes of external realizations can be characterized along the same lines as in [19].

3. Discrete time stochastic realization: The singular case.

3.1. Invariant predictable and smoothable subspaces. Problems I and II are called *singular* when $\Phi(\infty)$ is singular. It follows from Theorems 1.6 and 1.9 that in the singular case there exist nontrivial invariant directions for the Riccati equation (1.5) associated to every solution to problem I. Abusing language we shall say that a vector a is invariant (predictable) for $[A, B, C, D]$ if it is invariant (predictable) for the corresponding equation (1.5).

PROPOSITION 3.1. *The space $\mathcal{I}$ of invariant directions is invariant over all wide sense realizations of z.*

Proof. Immediate from Theorems 1.6 and 1.9. □

The following result describes the singular case in a number of different ways.

THEOREM 3.2. *The following statements are equivalent*:

(i) $\Phi(\infty)$ *is singular.*

(ii) Γ_* *is singular.*

(iii) $R(P^*)$ *is singular.*

(iv) $R(P_*)^{1/2}-B'_*(F')^{-1}H'$ *is singular.*

Proof. Let $\gamma \in \mathrm{R}^m$ be in the null space of $\Phi(\infty)$. Then, recalling that $\Phi(\infty)=DD'-DB'(F')^{-1}H'$ where $[F, B, H, D]$ is any wide sense realization, we obtain from (2.25) $B'_*(F')^{-1}H'\gamma=(H\Sigma H'+DD')^{1/2}\gamma=R(P_*)^{1/2}\gamma$. Hence $\gamma \in \mathcal{N}(R(P_*)^{1/2}-B'_*(F')^{-1}H')$ and $(F')^{-1}H'\gamma \in \mathcal{N}(\Gamma'_*)$. Conversely, if (ii) holds, use the fact that the eigenvalues of Γ_* are equal to the zeros of the determinant of W_* to get (iv) from which (i) follows trivially. The equivalence between (ii) and (iii) has been proven by Ruckebusch [33, p. 70]. □

COROLLARY 3.3. *The set $\mathcal{Q}$ is nonempty if and only if $\Phi(\infty)$ is singular.*

Proof. For any $P \in \mathcal{Q}$ we have $R(P^*) \leqq R(P)$. □

This says that the singular case occurs precisely when some of the wide sense realizations have $R(P)$ singular, in particular when $R(P^*)$ is singular. This contrasts with the continuous time situation where, when the innovation process is full rank, all the input noises have nonsingular intensity.

Let $T_H(i)$, $i=0, 1, \cdots$ be as in Theorem 1.9 so that $\Phi(z)=\sum_{i=0}^{\infty} T_H(i)z^{-i}$ for $|z|$ large enough and $\tilde{T}_*$ be the weighting pattern (1.17) corresponding to the minimum variance realization.

THEOREM 3.4. *The following statements are equivalent*:

(i) *a is an s-invariant direction of the wide sense realization* $[F, B, H, D]$.

(ii) $a=\sum_{i=1}^{s}(F')^{-i}H'\lambda_i$ *with* $\sum_{i=0}^{s-j} T_H(i)\lambda_{j+i}=0, j=1, \cdots, s.$

(iii) $a=\sum_{i=1}^{s}(F')^{-i}H'\lambda_i$ *with* $\sum_{i=0}^{s-j} \tilde{T}_*(i)\lambda_{j+1}=0, j=1, \cdots, s.$

(iv) $a=\sum_{i=1}^{s}(F')^{-i}H'\lambda_i$ *with* $a'x_*(t)=\sum_{i=1}^{s}\lambda'_i z(t-i)$ *for all t.*

(v) *a is a generalized eigenvector of rank s (an eigenvector if $s=1$) of Γ'_* corresponding to the eigenvalue zero.*

Proof. The equivalence of (ii) and (iv) is immediate. The rest follows at once from Theorem 1.6, in view of Proposition 3.1 and the fact that the deterministic and stochastic elements in the minimum variance realization can be obtained as limits of the corresponding quantities in a transient Kalman filter of the form (1.3). □

COROLLARY 3.5. *All the invariant directions of* $[F, B_*, H, R(P_*)^{1/2}]$ *are predictable.*

Proof. It follows directly from Theorem 1.11 and condition (iii) of Theorem 3.4. □

Note that in Theorem 3.4 the space $\mathcal{I}$ appears as the invariant subspace of Γ'_* related to the zero eigenvalue. We now introduce the backward counterpart of the concept of invariant direction. A vector $\bar{a}$ is said to be a *dually* s-invariant direction of the dual transient Riccati equation

$$
(3.1)\quad
\begin{aligned}
&\bar{\Sigma}(t-1)=F'\bar{\Sigma}(t)F-(F'\bar{\Sigma}(t)G+\bar{B}\bar{D}')(G'\bar{\Sigma}(t)G+\bar{D}\bar{D}')^{-1}(G'\bar{\Sigma}(t)F+\bar{D}\bar{B}')+\bar{B}\bar{B}',\\
&\bar{\Sigma}(0)=\bar{P}
\end{aligned}
$$

if $\bar{a}'\bar{\Sigma}(-t; \bar{P})=\bar{a}'\bar{\Sigma}(-s; 0)$ for all $t \geqq s$ and all $\bar{P} \in \mathscr{C}_n$. Also let $\tilde{\bar{T}}$ be given by (1.17) with $[F', \bar{B}, G', \bar{R}(\bar{P})^{1/2}]$ in place of $[A, B, C, D]$. Duality now gives the following result.

COROLLARY 3.6. *The following statements are equivalent*:

(i) $\bar{a}$ *is a dually s-invariant direction of the backward wide sense realization* $[F', \bar{B}, G', \bar{D}]$.

(ii) $\bar{a} = \sum_{i=1}^{s} F^{-i} G\mu_i$ *with* $\sum_{i=0}^{s-j} \underset{\sim}{T}_H(i)\mu_{j+i} = 0, j = 1, \cdots, s$.

(iii) $\bar{a} = \sum_{i=1}^{s} F^{-i} G\mu_i$ *with* $\sum_{i=0}^{s-j} \bar{T}_*(i)\mu_{j+i} = 0, j = 1, \cdots, s$.

(iv) $\bar{a} = \sum_{i=1}^{s} F^{-i} G\mu_i$ *with* $\bar{a}'\bar{x}_*(t) = \sum_{i=1}^{s} \mu_i' z(t+i)$ *for all t*.

(v) $\bar{a}$ *is a generalized eigenvector of rank s (an eigenvector if* $s = 1$*) of* $\bar{\Gamma}_*' = F' - \bar{B}_*\bar{R}(\bar{P}_*)^{-1/2}G'$ *corresponding to the eigenvalue zero.*

Next we define the dual counterpart of predictability.

DEFINITION 3.7. The n-dimensional vector $\bar{a}$ is called an *s-smoothable direction* of (3.1) if

$$\bar{a}'\bar{\Sigma}(-t; \bar{P}) = \bar{a}'\bar{\Sigma}(-s; \bar{P}) = 0 \quad \text{for all } t \geqq s. \tag{3.2}$$

The terminology is motivated by the fact that if $\bar{a}$ satisfies (3.2) then, by property (iv) in Corollary 3.6, we can smooth the state of any proper stochastic realization corresponding to $[F', \bar{B}, G', \bar{D}]$ exactly in direction $P^{-1}\bar{a}$. Clearly all the dually invariant directions of $[F', \bar{B}_*, G', \bar{R}(\bar{P}_*)^{1/2}]$ are smoothable. Let $\bar{\mathcal{I}}$ indicate the space of the invariant directions of (3.1) which, by Proposition 3.1 and duality, is invariant over all backward wide sense realization. Ruckebusch proved that $\bar{\Gamma}_* = (P^*)^{-1}(P^* - P_*)\Gamma_*'(P^* - P_*)^{-1}P^*$[33, p. 53]. Therefore it follows from Corollary 3.6 that $(P^* - P_*)(P^*)^{-1}\bar{\mathcal{I}}$ is the invariant subspace of Γ_* corresponding to the zero eigenvalue. Moreover the dimensions of $\mathcal{I}$ and $\bar{\mathcal{I}}$ are equal. The following theorem characterizes the predictable subspace of an internal realization and the smoothable subspace of the corresponding backward realization. It also shows that the sum of the dimensions of these two subspaces is constant and equal to dim $\mathcal{I}$.

THEOREM 3.8. *Let x be the state process of the internal realization* $[F, B_1, H, R(P)^{1/2}; u]$ *and S the invariant subspace of* Γ_* *associated with x in Theorem* 2.8, *so that* $x(t) = x_*(t) + \pi_s(x^*(t) - x_*(t))$ *with* π_s *given by* (2.39). *Then, if* $a = \sum_{i=1}^{n} (F')^{-i}H'\lambda_i$ *belongs to* $S^\perp \cap \mathcal{I}$ *and* $\bar{a} = \sum_{i=1}^{n} F^{-i}G\mu_i$ *belongs to* $P^*(P^* - P_*)^{-1}S \cap \bar{\mathcal{I}}$ *we have*

$$a'x(t) = \sum_{i=1}^{n} \lambda_i' z(t-i) \tag{3.3}$$

and

$$\bar{a}'(P^*)^{-1}x(t) = \sum_{i=1}^{n} \mu_i' z(t+i-1). \tag{3.4}$$

Moreover dim $(S^\perp \cap \mathcal{I})$ + dim $(P^*(P^* - P_*)^{-1}S \cap \bar{\mathcal{I}})$ = dim $\mathcal{I}$.

Proof. Since $(P^* - P_*)^{-1}\pi_s(P^* - P_*) = \pi_s'$ and π_s projects parallel to $(P^* - P_*)S^\perp$, we have $a'\pi_s = 0$ and $\bar{a}'(P^*)^{-1}\pi_s = \bar{a}'(P^*)^{-1}$. Properties (iv) of Theorem 3.4 and Corollary 3.6 now yield (3.3) and (3.4) respectively. Let k be the smallest positive integer such that $\mathcal{I} = \mathcal{N}((\Gamma_*')^k)$; Theorem 3.4(v) insures the existence of such a k. Then we have the direct decomposition $\mathsf{R}^n = \mathcal{I} \oplus \mathcal{R}((\Gamma_*')^k)$, where $\mathcal{R}((\Gamma_*')^k)$ is the range space of $(\Gamma_*')^k$, cf. [15, p. 166] for example. Consider also the usual orthogonal decomposition $\mathsf{R}^n = \mathcal{N}((\Gamma_*)^k) \oplus \mathcal{R}((\Gamma_*')^k)$, where $\mathcal{N}((\Gamma_*)^k) = (P^* - P_*)(P^*)^{-1}\bar{\mathcal{I}}$. It follows that dim $(S \cap \mathcal{I})$ = dim $(S \cap (P^* - P_*)(P^*)^{-1}\bar{\mathcal{I}})$. To complete the proof, observe that $\mathcal{I} = (\mathcal{I} \cap S) \oplus (\mathcal{I} \cap S^\perp)$ and that dim $(S \cap (P^* - P_*)(P^*)^{-1}\bar{\mathcal{I}})$ = dim $(P^*(P^* - P_*)^{-1}S \cap \bar{\mathcal{I}})$. □

It is worthwhile mentioning that $\bar{a}'(P^*)^{-1}$ in (3.4) has actually the form $\sum_{i=1}^{n} \mu_{n-i}' HF^{i-1}$ with $\sum_{k=0}^{n-j} T(k)'\mu_{n-k} = 0$ for $j = 1, \cdots, n$, as one can readily verify using (2.4)–(2.5) and (2.16)–(2.17) to establish the correspondence between $T(\cdot)'$ and $\tilde{T}(\cdot)$. Conversely such a vector leads to a smoothable direction in the backward setting. Hence a predictable-smoothable direction in the forward setting (i.e., a direction in which the state can be computed from a finite number of observations z) has the form $\sum_{i=-n}^{n-1} (F')^i H'\gamma_i$ with $\gamma \in \mathcal{N}(\mathrm{T})$, where $\gamma' = (\gamma'_{n-1}, \gamma'_{n-2}, \cdots, \gamma'_0, \gamma'_{-n}, \cdots, \gamma'_{-1})$ and T is a block diagonal matrix, the two diagonal blocks being block triangular Toeplitz matrices. The upper one has ith row $[T(i-1)', T(i-2)', \cdots, T(0)', 0, \cdots, 0]$ and the lower one has ith row $[\tilde{T}(i-1), \tilde{T}(i-2), \cdots, \tilde{T}(0), 0, \cdots, 0]$, where $i = 1, \cdots, n$.

The linear hull of the components of $x_*(t)$ and $x^*(t)$ is called the *frame space* [18] and denoted by $H_t^{\square}(z)$. In view of Theorem 2.8, we know that the components of the state at time t of an internal realization belong to $H_t^{\square}(z)$. Let us introduce the subspace $H_{t^+}(z)$ of $H_t^{\square}(z)$, given by the linear hull of elements of the form $a'x_*(t)$ and $\bar{a}'(P^*)^{-1}x^*(t)$, where a varies over $\mathcal{I}$ and $\bar{a}$ over $\bar{\mathcal{I}}$. By analogy to the continuous time case [10], we shall call $H_{t^+}(z)$ the *germ space*, since it contains linear combinations of differences of the type $\Delta_r z(s) = z(s) - z(s-r)$ and of certain other values of the process z that indicate precisely the degree of "smoothness" of the covariance of z in different directions. Then Theorem 3.8 shows that $\dim (X(t) \cap H_{t^+}(z)) = \dim \mathcal{I}$, where $X(t)$ is the space spanned by the components of the state $x(t)$ of an internal realization. Note that in contrast to the continuous time situation [18], the inclusion $H_{t^+}(z) \subset X(t)$ does not hold. From now on let $\dim I = \nu$.

THEOREM 3.9. *Let* $[F, B_1, H, R(P)^{1/2}; u]$ *be an internal realization. Then this realization can be embedded in a chain of internal realizations* $[F, B_1(i), H, R(P_i)^{1/2}; u_i]$ *with state spaces* $X_i(t), i = 0, \cdots, \nu$, *such that* $P_0 \leqq P_1 \leqq \cdots \leqq P_\nu$, $(X_0(t) \cap H_{t^+}(z)) \subset H_{t-1}^-(z)$ *and* $(X_\nu(t) \cap H_{t^+}(z)) \subset H_t^+(z)$.

Proof. Let S be as in Theorem 3.8 and $a_1, \cdots, a_r$ be a basis for $S^\perp \cap \mathcal{I}$. Then we can generate a family S_i of invariant subspaces of Γ_*, $i = 0, \cdots, \nu$, with $\dim (S_i^\perp \cap \mathcal{I}) = \nu - i$, simply eliminating from $S^\perp$, one at a time, the a_i or adding to $S^\perp$ new linearly independent elements of $\mathcal{I}$, both operations being performed taking due care of the rank of the generalized eigenvectors which are dropped or added, so that the resulting subspace is indeed invariant for Γ'_*. This can be done since $\mathcal{I}$ can be decomposed into cyclic subspaces. Clearly this procedure yields a family of internal realizations which differ only on the germ space and such that $S = S_{\nu-r}$. The state covariances are totally ordered since, if $i < j$ and $x_i(t)$, $x_j(t)$ are the corresponding state processes, $x_i(t)$ is equal to $x_*(t)$ in any direction in which it differs from $x_j(t)$. Finally, by construction, $[F, B_1(0), H, R(P_0)^{1/2}; u_0]$ has a full size predictable subspace and the backward realization corresponding to $[F, B_1(\nu), H, R(P_\nu)^{1/2}; u_\nu]$ has a full size smoothable subspace. Thus, the last assertion of the theorem follows. □

Notice that the chain of realizations in Theorem 3.9 is by no means unique. However the minimum and the maximum realizations are uniquely determined. In the case when Γ_* is cyclic, the number of internal realizations is finite and $\leqq 2^n$ [40; Remark 18]. Our work has shown that $2^{n-\nu}(\nu+1)$ is actually an upper bound in the cyclic case. In fact internal realizations are in one-to-tone correspondence with the invariant subspaces of Γ_* and, when Γ_* is cyclic, $\mathcal{I}$ is cyclic and the chain of invariant subspaces constructed in Theorem 3.9 is unique, so that the number of different invariant subspaces of Γ_* is less than or equal to $2^{n-\nu}(\nu+1)$.

Let us consider a proper external realization of the form (2.6)–(2.7) and an invariant direction $a = \sum_{i=1}^{n} (F')^i H'\lambda_i$ for it which is not predictable. Then two cases can occur. Either $\sum_{i=1}^{n-j} (F')^{-i} H'\lambda_{i+j}$ belongs to $\mathcal{N}(B_2)$ for $j = 0, \cdots, n-1$ or it does not. It can be seen that in the first case we are in a situation akin to the one for internal realizations and we can associate to the vector a a smoothable direction in the backward setting. In the second case, which always occurs if $B_2B_2' > 0$, a is invariant but the state cannot be determined exactly from a finite string of observations and we would need to have available the process ζ orthogonal to $H(z)$ and to model external realizations as done in § 2.5 to be able to calculate the state in ν linearly independent directions. For the sake of brevity, we have avoided here going into details about external realizations. However, it should be clear from our discussion that the sum of the dimensions of the predictable and smoothable subspaces associated with an external realization is less than or equal to ν. This fact has the intuitive meaning of indeterminacy introduced by the presence of the orthogonal component ζ.

The presence of nontrivial invariant directions allows, as it should be expected, for a reduction in the dimension of the filtering algorithms available in the literature. For instance, it is a simple exercise to verify that Faurre's algorithms to compute P_* and P^* [12, p. 56] reduce to solving $(n-\nu)\times(n-\nu)$ matrix equations, the values of P_* and $(P^*)^{-1}$ on the subspaces $\mathscr{I}$ and $\bar{\mathscr{I}}$ respectively being known a priori in terms of H, F and G. A similar reduction can be obtained for the fast algorithms which compute the gain (1.4) directly (cf. [17] for example), since it is clear that in an invariant direction the value of the gain can be computed directly in terms of the system matrices.

3.2. Noise free stochastic realization and the singular case. Akaike, in his important paper [1], deals with Markovian representations of the process z without noise in the output and only in his concluding remarks discusses representations with additive noise terms. Indeed, his work was based on some results of Faurre [11] which, starting from a certain factorization of the covariance matrices, were phrased in terms of noise-free realizations. In subsequent work [12] Faurre turned to a different factorization of the covariance matrices which led naturally to realizations with noise in the output. The same choice has, since then, been made by a number of authors [13], [22], [23], [33], but, up to our knowledge, it has never been explained whether the two approaches are equivalent and, if not, what are the shortcomings of either one. We shall now show that, precisely in the singular case, the first approach presents a considerable disadvantage, in that many minimal Markovian realizations are lost. Let us start considering a minimal factorization (Ξ, Θ, Ψ) (i.e., completely controllable and observable) like the one in [11], namely

$$\Delta_j = E\{z(t+j)z(t)'\} = \Psi\Xi^j\Theta, \qquad j=0,1,2,\cdots \tag{3.5}$$

and let dim $\Xi = r$. On the other hand, since Φ is the double side z-transform of Δ, we have

$$\Delta_j = \begin{cases} HF^{j-1}G, & j=1,2,3,\cdots, \\ G'(F')^{-1}H'+\Phi(\infty), & j=0. \end{cases} \tag{3.6}$$

THEOREM 3.10. *Let k be the dimension of $\mathscr{N}(\Phi(\infty))$ and assume, without loss of generality, that $\Phi(\infty) = [R \quad 0]$ where R is $(m-k)\times m$. Then (Ξ, Θ, Ψ) is given, up to a change of basis, by*

$$(\Xi, \Theta, \Psi) = \left(\tilde{F}, \begin{bmatrix} F^{-1}G \\ R \end{bmatrix}, \begin{bmatrix} H & \begin{pmatrix} I \\ 0 \end{pmatrix} \end{bmatrix}\right), \tag{3.7}$$

where

$$\tilde{F} = \begin{bmatrix} F & 0 \\ 0 & 0 \end{bmatrix}$$

the identity matrix is m-k dimensional and $r=n+m-k$.

Proof. It is easy to check that the triplet in (3.7) satisfies (3.5). Also (Ξ, Θ) is controllable and (Ξ, Ψ) is observable. In fact suppose $\alpha = \begin{pmatrix}\alpha_1 \\ \alpha_2\end{pmatrix}$ with $\alpha_1 \in \mathsf{R}^n$ and $\alpha_2 \in \mathsf{R}^{m-k}$ is such that

$$(3.8)\qquad (\alpha_1', \alpha_2')\begin{bmatrix} F^{-1}G & G & FG & \cdots & F^{n+m-k-2}G \\ R & 0 & 0 & \cdots & 0 \end{bmatrix} = 0.$$

Then we see that α_1 must be zero, which forces $\alpha_2' R = 0$ and finally $\alpha_2 = 0$. We conclude that the controllability matrix in (3.8) is full rank. Similarly the observability matrix is seen to have rank $n+m-k$. The conclusion now follows from the uniqueness, up to an equivalence as in (2.1), of the triplet (Ξ, Θ, Ψ). □

Let us assume for the moment that $\Phi(\infty)$ is nonsingular and consider a proper stochastic realization of z $[F, B, H, D, w]$. Then we can associate to it the noise free model

$$(3.9)\qquad \xi(t+1) = \tilde{F}\xi(t) + \begin{bmatrix} F^{-1}B \\ D - HF^{-1}B \end{bmatrix}\eta(t),$$

$$(3.10)\qquad z(t) = [H \quad I]\xi(t),$$

where

$$\xi(t) = \begin{pmatrix} F^{-1}x(t+1) \\ (D - HF^{-1}B)w(t) \end{pmatrix}$$

and $\eta(t) = w(t+1)$. This induces a one-to-one correspondence between wide sense realizations of the form $[F, B, H, D]$ and noise free wide sense realizations of the form $[\tilde{F}, \chi, (HI)]$ which are minimal too in view of Theorem 3.10. If we agree to call realizations $[\tilde{F}, \chi, (H \quad I); \eta]$ with $\chi (n+m) \times m$ internal, then the above correspondence is one-to-one between internal realizations. In particular it maps $[F, B_*, H, R(P_*)^{1/2}; u_*]$ to a realization related to the *steady state pure filter*, i.e., the second innovation representation IR_2 in Gevers terminology [14].

Suppose now that $\Phi(\infty)$ is as in Theorem 3.10 with $k > 0$. Then it is possible to set up a correspondence similar to the one in the nonsingular case only for a rather small subclass of wide sense realizations. More explicitly, let $[F, B, H, D; w]$ be a realization such that $\tilde{T}(0) = D' - B'(F')^{-1}H'$ has rank $m-k$ and V an orthogonal matrix such that $[D - HF^{-1}B]V = \begin{bmatrix} S \\ 0 \end{bmatrix}$ where S is $(m-k) \times p$, p being the number of columns of B. Then we have the $n+m-k$ dimensional noise free model

$$(3.11)\qquad \xi(t+1) = \tilde{F}\xi(t) + \begin{bmatrix} F^{-1}B \\ S \end{bmatrix}\eta(t),$$

$$(3.12)\qquad z(t) = \left[H \quad \begin{pmatrix} I \\ 0 \end{pmatrix}\right]\xi(t),$$

where $\xi(t) = \begin{pmatrix} F^{-1}x(t+1) \\ SV'w(t) \end{pmatrix}$ and $\eta(t) = V'w(t+1)$. The wide sense realization given by (3.11)–(3.12) is minimal. This establishes a one-to-one correspondence between minimal wide sense realizations of z such that $\tilde{T}(0)$ has rank $m-k$ and minimal wide sense realizations of the form $\left[\tilde{F}, \chi, \left(H \quad \begin{pmatrix} I \\ 0 \end{pmatrix}\right)\right]$. It is now apparent that the choice of seeking noise free representation of z can cost us, in the singular case, the loss of a considerable number of realizations. Indeed, it is not hard to see that the subset of $\mathcal{P}$ corresponding to realizations with rank $\tilde{T}(0) = m - k$ lies, as $\mathcal{Q}$, in the relative boundary of $\mathcal{P}$.

This shows that, in discrete time, the factorization (3.6) and the associated choice of $H_{t-1}^{-}(z)$, instead of $H_t^{-}(z)$, as past space at time t, is more convenient, even though it implies the unpleasant fact that white noise processes have zero dimensional minimal realizations.

Acknowledgments. The author would like to thank his dissertation advisor Professor A. Lindquist for his constant encouragement and many valuable suggestions and Professor G. Picci with whom he had countless discussions on the stochastic realization problem which led to a considerable improvement of this paper.

REFERENCES

[1] H. AKAIKE, *Markovian representation of stochastic processes by canonical variables*, this Journal, 13 (1975), pp. 162–173.

[2] ———, *Stochastic theory of minimal realization*, IEEE Trans. Automatic Control, AC-19 (1974), pp. 667–674.

[3] B. D. O. ANDERSON, *Dual form of a positive real lemma*, Proc. IEEE 55 (1967), pp. 1749–1750.

[4] ———, *The inverse problem of stationary convariance generation*, J. Statistical Physics, 1 (1969), pp. 133–147.

[5] B. D. O. ANDERSON AND L. HITZ, *Discrete positive-real functions and their applications to system stability*, Proc. IEEE, 116 (1969), pp. 153–155.

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

[7] A. E. BRYSON AND Y. C. HO, *Applied Optimal Control*, Blaisdell, Waltham, MA, 1969.

[8] R. S. BUCY, D. RAPPAPORT AND L. M. SILVERMAN, *Correlated noise filtering and invariant directions for the Riccati equation*, IEEE Trans. Automatic Control, AC-15 (1970), pp. 535–540.

[9] D. J. CLEMENTS AND B. D. O. ANDERSON, *Linear-quadratic discrete-time control and constant directions*, Automatica—J.IFAC., 13 (1977), pp. 255–264.

[10] H. DYM AND H. P. MCKEAN, *Gaussian Processes, Function Theory, and the Inverse Spectral Problem*, Academic Press, New York, 1976.

[11] P. FAURRE, *Identification par minimisation d'une représentation Markovienne de processus aléatoire*, Symposium on Optimization, Lecture Notes in Mathematics 132, Springer, Berlin, 1970, pp. 83–107.

[12] ———, *Réalisations Markoviennes de processus stationnaires*, Research Rep. 13, March 1973, IRIA(LABORIA), Le Chesnay, France.

[13] F. GERMAIN, *Algoritmes de calcul de réalisations Markoviennes. Cas singuliers et stabilité*, Research Rep. 66, April 1974, IRIA (LABORIA), Le Chesnay, France.

[14] M. R. GEVERS, *Structural properties of realizations of discrete-time Markovian processes*, Tech. Rep. 7050-19, May 1972, Information Systems Laboratory, Stanford University, Stanford, CA.

[15] P. R. HALMOS, *Finite Dimensional Vector Spaces*, Princeton University Press, Princeton, NJ, 1942.

[16] B. L. HO AND R. E. KALMAN, *Effective construction of linear state-variable models from input/output functions*, Proc. Third Allerton Conf. (1965), pp, 449–459.

[17] A. LINDQUIST, *A new algorithm for optimal filtering of discrete-time stationary processes*, this Journal, 12 (1974), pp. 736–746.

[18] A. LINDQUIST AND G. PICCI, *On the structure of minimal splitting subspaces in stochastic realization theory*, Proc. 1977 Decision and Control Conference, New Orleans, pp. 42–48.

[19] ———, *On the stochastic realization problem,*, this Journal, to appear.

[20] ———, *A state space theory for stationary stochastic processes*, Proc. 21st Midwest Symposium on Circuits and Systems, Aimes, August 1978.

[21] ———, *A Hardy space approach to the stochastic realization problem*, Proc. 1978 Decision and Control Conf., San Diego, CA, to appear.

[22] A. LINDQUIST, G. PICCI AND G. RUCKEBUSCH, *On minimal splitting subspaces and Markovian representations*, to appear.

[23] M. PAVON, *Constant directions and singular stochastic realization: The scalar case*, Proc. 1^{re} Colloque AFCET-SMF, Palaiseau, France, September 1978.

[24] M. PAVON AND R. J.-B. WETS, *A stochastic variational approach to the duality between estimation and control: Discrete time*, Proc. the Analysis and Optimization of Stochastic Systems Conf., Oxford, England, September 1978.

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

[26] R. PENROSE, *A generalized inverse for matrices*, Proc. Cambridge Phil. Soc., 51 (1955), pp. 406–413.

[27] G. PICCI, *Stochastic realization of Gaussian processes*, Proc. IEEE, 64 (1976), pp. 112–122.

[28] V. M. POPOV, *Hyperstability of Control Systems*, Springer-Verlag, New York, 1973.

[29] D. RAPPAPORT, *Constant directions of the Riccati equations*, Automatica—J.IFAC., 8 (1972), pp. 175–186.

[30] D. RAPPAPORT AND L. M. SILVERMAN, *Structure and stability of discrete-time optimal systems*, IEEE Trans. Automatic Control, AC-16 (1971), pp. 227–232.

[31] Y. A. ROZANOV, *Stationary Random Processes*, Holden-Day, San Francisco, CA, 1967.

[32] ———, *On Markov extensions of a random process*, Theor. Probability Appl., 22 (1977), pp. 190–195.

[33] G. RUCKEBUSCH, *Représentations Markoviennes de processus Gaussien Stationnaires*, Thèse 3ème cycle, Univ. of Paris VI, May 1975.

[34] ———, *Représentations Markoviennes de processus Gaussien stationnaires et applications statistiques*, Springer-Verlag Lecture Notes in Mathematics 636, Grenoble, June 1977.

[35] ———, *A state space approach to the stochastic realization problem*, Proc. 1978 Intern. Symp. Circuits and Systems, New York.

[36] ———, *Factorizations minimales de densités spectrales et représentations Markoviennes*, Proc. 1^{re} Colloque AFCET-SMF, Palaiseau, France, September 1978.

[37] G. S. SIDHU AND U. B. DESAY, *New smoothing algorithms based on reverse-time lumped models*, IEEE Trans. Automatic Control, AC-21 (1976), pp. 538–541.

[38] L. M. SILVERMAN, *Discrete Riccati equations: Alternative algorithms, asymptotic properties and system theory interpretations*, Control and Dynamic Systems, 12 (1976), pp. 313–386.

[39] L. M. SILVERMAN AND H. E. MEADOWS, *Equivalence and synthesis of time variable linear systems*, Proc. Fourth Allterton Conf., (1966), pp. 776–784.

[40] J. C. WILLEMS, *Least square stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Automatic Control, AC-16 (1971), pp. 621–634.

[41] D. C. YOULA AND P. TISSI, *n-port synthesis via reactance extraction—Part I*, IEEE Intern. Convention Record, 14 (1966), pp. 183–205.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 2, March 1980

0363-0129/80/1802-0006 $01.00/0

ASYMPTOTIC STABILITY OF SYSTEMS: RESULTS INVOLVING THE SYSTEM TOPOLOGY*

R. K. MILLER† AND A. N. MICHEL‡

Abstract. In this paper we answer the following question for a large class of (linear and nonlinear) dynamical systems. Given is a system with dissipation and given is the associated conservative system. Suppose the associated conservative system is stable. What properties of the system topology (system configuration) will ensure that the overall system with dissipation is asymptotically stable?

Both linear and nonlinear (Hamiltonian) systems are treated. For the linear case, necessary and sufficient conditions for asymptotic stability are established, while for the nonlinear case, sufficient conditions and also some necessary and sufficient conditions for asymptotic stability are obtained.

It is emphasized that the application of the present results to specific problems will usually *not* require a search for appropriate Lyapunov functions. Indeed, a stability analysis by the present method involves the following two steps:

(a) given a system with dissipation, the stability of its trivial solution (equilibrium) is ascertained by determining the stability of the associated conservative system, i.e., by determining whether the potential energy is a minimum at the equilibrium; and

(b) attractivity of the equilibrium of the entire system (with dissipation) is determined from the system topology (system configuration).

This approach to stability analysis appears to be new. Furthermore, since the present method involves concepts from control theory (namely, the notion of observability), these results provide further insight into the mechanisms of stability (and stabilization).

To provide motivation and to demonstrate the applicability of the results, some specific examples are considered.

1. Introduction. Consider the linear mechanical mass-spring system of Fig. 1 which is governed by the equations

$$
\begin{aligned}
&m_1\ddot{x}_1+k_1x_1+k(x_1-x_2)=0,\\
&m_2\ddot{x}_2+k_2x_2+k(x_2-x_1)=0,
\end{aligned}
\tag{1}
$$

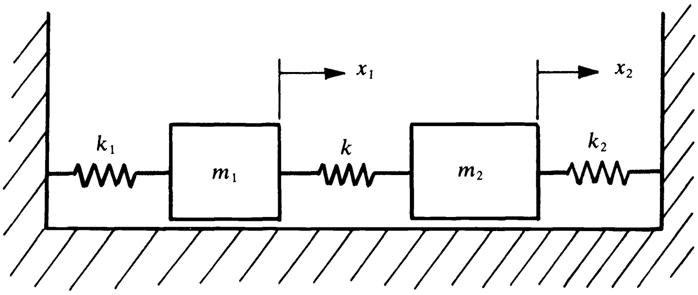


FIG. 1.

where x_i denotes the displacement of mass m_i and k_1, k_2, k denote linear spring constants. When the initial state of this conservative system is displaced from its equilibrium position, the system will remain in motion indefinitely. If linear viscous damping is added at some or all of the masses and springs, as shown in Fig. 2, then the

* Received by the editors November 28, 1978, and in revised form July 17, 1979. This work was supported in part by the National Science Foundation under Grant ENG 77-28446 and by the Engineering Research Institute, Iowa State University.

† Department of Mathematics, Iowa State University, Ames, Iowa 50011.

‡ Department of Electrical Engineering and Engineering Research Institute, Iowa State University, Ames, Iowa 50011.

governing equations become

$$\begin{aligned} m_1\ddot{x}_1+k_1x_1+k(x_1-x_2)+B_1\dot{x}_1+B(\dot{x}_1-\dot{x}_2)&=0,\\ m_2\ddot{x}_2+k_2x_2+k(x_2-x_1)+B_2\dot{x}_2+B(\dot{x}_2-\dot{x}_1)&=0, \end{aligned} \tag{2}$$

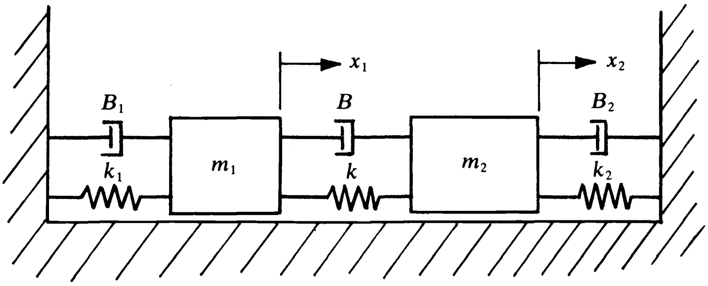


FIG. 2.

where $B_1 \geqq 0, B_2 \geqq 0, B \geqq 0$ and $B_1+B_2+B>0$. At a first glance, it would seem that the indiscriminate or random addition of such damping terms will stabilize the rest position, making system (2) asymptotically stable. Indeed, one could argue that since the addition of a dash pot at even one single location shown in Fig. 2 will reduce the total energy of the system, eventually all of the energy will be dissipated and the motion of the system will tend to its equilibrium.

The preceding argument is simple, appealing but unfortunately wrong. While for most values of the parameters the above conjecture is correct, it is not true when $B_1=B_2=0, B>0, k_1/m_1=k_2/m_2$, for in this case the two masses can be made to move in synchronism. When this happens, x_1-x_2 is constant, the term $B(\dot{x}_1-\dot{x}_2)$ has no effect on the motion and no dissipation of energy will occur.

The conservative system (1) and its corresponding damped system (2) are simple enough to be analyzed by simple inspection. However, in the case of general, highly complex, possibly nonlinear, stable conservative systems, it is far from trivial to decide where damping should be added in order to ensure that the rest position will be asymptotically stable. Similar questions can be asked with respect to adding dissipative terms in electrical systems, electromechanical systems, and so forth.

In the present paper we establish conditions which answer the questions raised above for such systems. Although we give results involving several cases, our main result answers the following question: for a stable and conservative system, what are appropriate conditions which ensure that an associated damped system will be asymptotically stable?

In § 2 we obtain general results for linear systems. These results are applied in § 3 to conservative mechanical systems to obtain a result of Walker and Schmitendorf [9]. The results in § 3 motivate generalizations to nonlinear systems which are presented in § 4. All of our nonlinear results concern conservative mechanical systems to which damping is added. Some related work for nonlinear circuits can be found in Varaiya and Liu [8]. Our results do not overlap those of [8] but have a similar flavor.

2. General linear systems. We will employ the following frequently used definition.

DEFINITION 1. Let U and V be matrices of dimensions $m \times n$ and $n \times n$, respectively. We say that the pair (U, V) is *observable* if and only if the matrix

$$\begin{bmatrix} U \\ UV \\ UV^2 \\ \cdots\cdots \\ UV^{n-1} \end{bmatrix}$$

has full rank.

We consider a linear system of differential equations given by

$$\dot{x} = Ax, \tag{3}$$

where $x \in R^n$, $t \in J = [t_0, \infty)$, $t_0 \geqq 0$, $\dot{x} = dx/dt$, and A is an $n \times n$ matrix. We let $x(t; x_0, t_0)$ denote the solutions of (3) with $x_0 = x(t_0; x_0, t_0)$. We assume that the trivial solution $x \equiv 0$ of (3) is *stable* (in the Lyapunov sense (see Hahn [3])) so that there is a positive definite matrix G(i.e., $G > 0$) such that the matrix

$$B = A^T G + GA$$

is negative semidefinite (i.e., $B \leqq 0$). Thus, there exists a Lyapunov function $v: R^n \to R$ with

$$\begin{aligned} &v(x) = x^T G x, \\ &Dv_{(3)}(x) = x^T B x, \\ &B = A^T G + GA, \end{aligned} \tag{4}$$

where $Dv_{(3)}(x)$ denotes the derivative of v with respect to t along the solutions of (3). Our first result is as follows.

THEOREM 1. *For system* (3) *assume a Lyapunov function* (4) *such that* $G > 0$ *and* $B \leqq 0$. *Then the trivial solution of system* (3) *is asymptotically stable* (*in the Lyapunov sense* (*see* [3])) *if and only if the pair* (B, A) *is observable.*

Proof. Suppose that (B, A) is observable and let

$$N = \{x : Bx = 0\} = \{x : Dv_{(3)}(x) = 0\}.$$

Let N_1 be the largest subset of N which is invariant with respect to system (3). (That is, N_I is the largest subset of N such that $x_0 \in N_I$ implies $x(t; x_0, t_0) \triangleq x(t) \in N_I$ for all $t \in R$.) Now if $x(t) \in N_I$, then $Bx(t) \equiv 0$ so that

$$\begin{aligned} &0 = \frac{d}{dt}(Bx) = B\dot{x} = BAx, \\ &0 = \frac{d}{dt}(BAx) = BA\dot{x} = BA^2 x, \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ &0 = \frac{d}{dt}(BA^{n-2}x) = BA^{n-2}\dot{x} = BA^{n-1}x. \end{aligned}$$

Since (B, A) is observable, we must conclude that any trajectory $x(t)$ in N_I is the rest state, i.e., $N_I = \{0\}$. By the invariance principle (see e.g., [1], [4], [6]) it follows that $x \equiv 0$ is asymptotically stable.

Conversely, assume that (B, A) is not observable. Then there is a trajectory $x(t) = [\exp(At)]x_0$, $x_0 \neq 0$, such that $B[\exp(At)]x_0 \equiv 0$. For this trajectory,

$$\frac{d}{dt}v(x(t)) = x^T(t)Bx(t) = x^T(t) \cdot 0 = 0,$$

and so $v(x(t)) \equiv v(x_0) > 0$ for all $t \geqq 0$. Thus, $x(t)$ can not tend to the origin as $t \to \infty$, i.e., the trivial solution $x = 0$ is not asymptotically stable. □

The method of proof used in Theorem 1 can be modified to establish the following more general result.

THEOREM 2. *For system* (3) *assume that* (4) *is true with* $G > 0$ *and* $B \leqq 0$. *Define*

$$N = \{x : Dv_{(3)}(x) = 0\}.$$

Suppose that there exists a matrix C such that the set

$$N_1 \triangleq \{x : Cx = 0\}$$

equals N, and suppose there exists a matrix D such that

$$N_2 \triangleq \{x : Dx = 0\} \supset N_1.$$

Then the trivial solution of (3) *is asymptotically stable if and only if the pair* $(C, A - D)$ *is observable.*

Proof. First we consider a trajectory $x(t) \in N$ for $-\infty < t < \infty$. Then $Cx(t) \equiv 0$ so that $Dx(t) \equiv 0$ and

$$\begin{aligned} 0 &= \frac{d}{dt}(Cx) = C\dot{x} = CAx = C(A-D)x, \\ 0 &= \frac{d}{dt}[C(A-D)x] = C(A-D)\dot{x} = C(A-D)^2x, \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ 0 &= C(A-D)^{n-1}x. \end{aligned}$$

Since $(C, A - D)$ is observable, it follows that $x(t) \equiv 0$. By the invariance result in [1] it follows that the equilibrium $x \equiv 0$ is asymptotically stable.

Now suppose that $N_2 \supset N_1$ and that $(C, A - D)$ is not observable. Then there exists $x_0 \neq 0$ such that $C\{\exp[(A-D)t]\}x_0 \equiv 0$. Since $N_1 = N$ it follows that

$$\frac{d}{dt}v(x(t)) = x^T(t)Bx(t) \equiv 0, \qquad x(t) \triangleq \exp[(A-D)t]x_0$$

and

$$v(x(t)) \equiv v(x_0) > 0.$$

Thus, $x(t)$ cannot tend to the origin. Also, since $N_2 \supset N_1$, we have

$$\dot{x}(t) = (A-D)x(t) = Ax(t) - 0 = Ax(t),$$

i.e., $x(t)$ solves (3). Thus, the equilibrium of (3) is not asymptotically stable. □

For Theorem 2 there are many possible choices for the matrices C and D. For example, $C = B$ and either $D = 0$ or $D = \pm B$ will do. As another example, since B is symmetric and negative semidefinite, there exists a matrix C such that $C^*C = -B$. For this choice of C we may choose $D = 0$, or $D = \pm C$, or $D = \pm B$. For a third way of choosing the matrices C and D, refer to Theorem 3 in the next section.

We note that David Russell [7] has communicated to the authors a version of Theorem 2 with $C^*C = -B$ and $D = 0$.

3. Linear Hamiltonian system. The Hamiltonian formulation in mechanics is well established (see, e.g., [2], [10]). Given a Hamiltonian function $H(q_1, \cdots, q_n, p_1, \cdots, p_n)$, the Hamiltonian differential equations for conservative systems are

$$\tag{5} \begin{aligned} \dot{q}_i &= \frac{\partial H}{\partial p_i}(q_1, \cdots, q_n, p_1, \cdots, p_n), \\ \dot{p}_i &= \frac{-\partial H}{\partial q_i}(q_1, \cdots, q_n, p_1, \cdots, p_n), \end{aligned}$$

$i = 1, \cdots, n$, or in vector notation,

$$\dot{q} = \frac{\partial H}{\partial p}(q, p), \qquad \dot{p} = \frac{-\partial H}{\partial q}(q, p),$$

where the q_i denote generalized position coordinates, the p_i denote the generalized momentum coordinates and H represents the total energy of a system. (Using appropriate analogies, lossless electrical systems, electro-mechanical systems, etc., can be represented by (5) as well.) The motions of system (5) are always such that the energy is conserved, since

$$\begin{aligned} \frac{d}{dt}H(q, p) &= \sum_{i=1}^{n} \left(\frac{\partial H}{\partial q_i}\dot{q}_i + \frac{\partial H}{\partial p_i}\dot{p}_i \right) \\ &= \sum_{i=1}^{n} \left(\frac{\partial H}{\partial q_i}\frac{\partial H}{\partial p_i} - \frac{\partial H}{\partial p_i}\frac{\partial H}{\partial q_i} \right) = 0. \end{aligned}$$

A linear Hamiltonian differential equation is obtained from a quadratic Hamiltonian of the form

$$\tag{6} H(q, p) = \tfrac{1}{2}q^T H_1 q + \tfrac{1}{2}p^T H_2 p,$$

where we can assume, without loss of generality, that the matrices H_1 and H_2 are symmetric. Since in general $dH/dt \equiv 0$, we can use H as a Lyapunov function for (5) (when the potential energy has a local isolated minimum at the equilibrium $(q^T, p^T) = (0^T, 0^T)$). In particular, in the linear case (6) we ensure stability of the trivial solution with the assumptions that H_1 and H_2 are positive definite.

The linear system of differential equations corresponding to the Hamiltonian (6) is given by

$$\tag{7} \dot{q} = H_1 p, \qquad \dot{p} = -H_2 q.$$

Now if viscous damping is added, then system (7) will be replaced by

$$\tag{8} \dot{q} = H_1 p, \qquad \dot{p} = -H_2 q + Kp,$$

where $B_1 = H_2 K + K^T H_2$ is negative semidefinite. In this case, the derivative of $H(q, p)$ along the solutions of (8) is given by

$$D_{(8)}H = p^T B_1 p \leqq 0.$$

If in particular we specialize (6) to a simple mechanical system consisting of n rigid bodies with masses m_i, $i = 1, \cdots, n$, then (6) will assume the form

$$\tag{9} H(q, p) = \tfrac{1}{2}q^T H q + \tfrac{1}{2}p^T M^{-1} p,$$

where $M = \operatorname{diag}[m_1, m_2, \cdots, m_n]$ and $H = H^T$ (which characterizes the potential energy term) is determined by the system configuration. For this case (8) assumes the form

$$\dot{q}=M^{-1}p, \qquad \dot{p}=-Hq+KM^{-1}p \tag{10}$$

which is equivalent to Newton's law ($p=M\dot{q}, K=K^T$),

$$M\ddot{q}+Hq-K\dot{q}=0.$$

The derivative of (9) along the solutions of (10) is easily computed to be

$$D_{(10)}H=p^TM^{-1}KM^{-1}p.$$

The next result, which is similar to a result reported in [9], is a direct consequence of Theorem 2.

THEOREM 3. *Consider the Hamiltonian* (9) *and the system* (10) *with* $M=\operatorname{diag}[m_1, m_2, \cdots, m_n]>0, H=H^T>0, K=K^T\leqq 0$. *Then the trivial solution of* (10) *is asymptotically stable if and only if the pair* $(K, M^{-1}H)$ *is observable.*

Proof. Applying Theorem 2 with

$$A=\begin{bmatrix}0 & M^{-1}\\ -H & KM^{-1}\end{bmatrix}, \qquad B=\begin{bmatrix}0 & 0\\ 0 & M^{-1}KM^{-1}\end{bmatrix}, \qquad C=D=\begin{bmatrix}0 & 0\\ 0 & KM^{-1}\end{bmatrix}$$

it is easy to see that for $j=0, 1, 2, 3, \cdots$, we have

$$C(A-D)^{2j+1}=\begin{bmatrix}0 & 0\\ K(-M^{-1}H)^{j+1} & 0\end{bmatrix}$$

and

$$C(A-D)^{2j}=\begin{bmatrix}0 & 0\\ 0 & K(-M^{-1}H)^jM^{-1}\end{bmatrix}.$$

Thus, the result follows from Theorem 2. □

On the basis of this theorem, we can formulate the following simple rule for conservative stable systems of the form (10):

Pick a position in the undamped system where it is possible to add damping (e.g., dashpot, resistor, etc.). If it is always possible to detect motion at this position whenever the system is not at rest, then this is a location at which damping, to stabilize the system, can be added. To cover multi-position cases, the above must be modified, using linear combinations of motions at allowable damping points.

The above rule is easily seen to work for the example discussed in the introduction. One can also check the algebra for this example to obtain the results precisely. Indeed, we have

$$K=\begin{bmatrix}(-B_1-B) & B\\ B & (-B_2-B)\end{bmatrix},$$

$$M^{-1}H=\begin{bmatrix}\frac{1}{m_1} & 0\\ 0 & \frac{1}{m_2}\end{bmatrix}\begin{bmatrix}(k_1+k) & -k\\ -k & (k_2+k)\end{bmatrix},$$

and we consider the following possibilities:

Case 1. $\det K\neq 0$;

Case 2a. $\det K=0$ with $B_1=B_2=0$;

Case 2b. $\det K=0$ with $B_1=B=0$;

Case 2c. $\det K=0$ with $B_2=B=0$;

For case 1, the pair $(K, M^{-1}H)$ is observable. For case 2a, we have

$$K(M^{-1}H)=B\begin{bmatrix}\frac{-(k_1+k)}{m_1}-\frac{k}{m_2} & \frac{(k_2+k)}{m_2}+\frac{k}{m_1}\\ \frac{(k_1+k)}{m_1}+\frac{k}{m_2} & \frac{-(k_2+k)}{m_2}-\frac{k}{m_1}\end{bmatrix}.$$

In order that the pair $(K, M^{-1}H)$ *not* be observable, it is necessary and sufficient that the first row of $KM^{-1}H$ be the negative of the second row, i.e.,

$$\frac{k_1+k}{m_1}+\frac{k}{m_2}=\frac{k_2+k}{m_2}+\frac{k}{m_1}$$

or

$$\frac{k_1}{m_1}=\frac{k_2}{m_2}.$$

This is the condition under which it is possible for the two masses to move in synchronism.

For case 2b it is easy to compute that the condition which ensures that the pair $(K, M^{-1}H)$ be observable is $B_2k>0$.

For case 2c the condition which ensures that the pair $(K, M^{-1}H)$ be observable is $B_1k>0$.

Applying Theorem 3 we see that the system of Fig. 2 will be asymptotically stable for all the above cases, except case 2a.

4. Nonlinear Hamiltonian systems. We now extend the results of § 3 to nonlinear systems. To this end we consider a Hamiltonian of the form

$$H(q,p)=\tfrac{1}{2}p^TM^{-1}p+G(q) \tag{11}$$

and the associated system (with damping)

$$\dot q=M^{-1}p,\qquad \dot p=KM^{-1}p-\nabla G(q), \tag{12}$$

where $G: R^n\to R$ is assumed to be continuously differentiable over R^n and where ∇G denotes the gradient of G. We will find it convenient to associate with (12) an *output equation* of the form

$$y=\operatorname{diag}[KM^{-1}, KM^{-1}]\begin{bmatrix}p\\ \nabla G(q)\end{bmatrix}. \tag{13}$$

DEFINITION 2. System (12), (13) is called *distinguishable* (see, e.g., [5, p. 377]) if whenever $(q(t), p(t))$ is a solution of (12) with $(q(0), p(0))\neq(0,0)$, then the output $y(t)\not\equiv 0$. System (12), (13) is called *locally distinguishable* if there is an $\varepsilon>0$ such that when $(q(t), p(t))$ is a solution of (12) and $0<|p(0)|+|q(0)|<\varepsilon$, then the output $y(t)\not\equiv 0$.

We now prove the following result.

THEOREM 4. *Consider Hamiltonian* (11) *and system* (12) *with* $M=\operatorname{diag}[m_1, m_2,\cdots, m_n]>0$, *$G$ positive definite with respect to the origin, and* $K=K^T\leqq 0$. *Then the trivial solution of system* (12) *is asymptotically stable if and only if the system* (12), (13) *is locally distinguishable.*

Proof. Since G is positive definite, we may choose H given by (11) as a Lyapunov function. The derivative of H with respect to t along the solutions of (12) is given by

$$DH_{(12)}(q,p)=p^TM^{-1}KM^{-1}p\leqq 0$$

for all (q, p). Let

$$N = \{(q, p): KM^{-1}p = 0\}$$

be the null set of $DH_{(12)}$ and let N_I be the largest invariant subset of N. If $(q(t), p(t))$ is a solution of (12) with $(q(0), p(0)) \in N_I$ and $|q(0)| + |p(0)|$ sufficiently small, then $KM^{-1}p(t) \equiv 0$ and

$$0 = KM^{-1}\dot{p}(t) = KM^{-1}(KM^{-1}p(t) - \nabla G(q(t))) = -KM^{-1}\nabla G(q(t)).$$

Since system (12), (13) is locally distinguishable, then $p(t) \equiv q(t) \equiv 0$. By the invariance theorem (see [4]) the trivial solution of (12) is asymptotically stable.

Conversely, assume that system (12), (13) is not locally distinguishable. Then in any neighborhood $\mathscr{U}$ of the origin $(0, 0)$ there is a nontrivial solution (q, p) of (12) which starts in $\mathscr{U}$ and for which the output (13) is identically zero. Thus (q, p) will solve the stable, conservative Hamiltonian system given by

$$\dot{q} = M^{-1}p, \qquad \dot{p} = -\nabla G(q).$$

Since G is positive definite and since $M = M^T > 0$, then $H(q(t), p(t)) \equiv H(q(0), p(0)) \triangleq H_0 > 0$ and $(q(t), p(t)) \not\to 0$ as $t \to \infty$. □

Essentially the same proof works for the next result.

THEOREM 5. *Consider Hamiltonian* (11) *and system* (12) *with* $M = \operatorname{diag}[m_1, m_2, \cdots, m_n] > 0$, *with G positive definite* (*with respect to the origin*) *for all q, with $G(q) \to \infty$ as $|q| \to \infty$* (*i.e., G is radially unbounded*), *and with $K = K^T \leqq 0$. Then the trivial solution of* (12) *is asymptotically stable in the large if and only if* (12), (13) *is distinguishable.*

In certain cases the distinguishability of the nonlinear system (12), (13) is easily checked. For example, if G_1 is the linear part of ∇G at $q = 0$ so that

$$\nabla G(q) = G_1 q + o(q), \qquad |q| \to 0,$$

then system (12), (13) can be linearized and we have

$$\begin{aligned} &\dot{q} = M^{-1}p, \\ (14) \qquad &\dot{p} = KM^{-1}p - G_1 q, \\ &y = [KM^{-1}p, KM^{-1}G_1 q]. \end{aligned}$$

In this case we obtain the following result.

COROLLARY 1. *Assume* $M = \operatorname{diag}[m_1, m_2, \cdots, m_n] > 0$, $\nabla G(q) = G_1 q + o(q)$ *near $q = 0$ with $G_1 = G_1^T > 0$ and $K = K^T \leqq 0$. Then the trivial solution of* (12) *is asymptotically stable if $(K, M^{-1}G_1)$ is observable in the sense of Definition* 1 (*see* § 2).

Proof. If $(K, M^{-1}G_1)$ satisfies the criterion of Definition 1, then system (14) is observable (see the proof of Theorem 3 and see, e.g., [5]). If system (14) is observable, then the corresponding nonlinear system (12), (13) must be locally observable (see [5, p. 378]) and hence also distinguishable. Apply now Theorem 4 to complete the proof. □

Remark. Theorems 1–5 can be stated in stronger terms by recalling the facts that (a) asymptotically stable plus autonomous imply uniformly asymptotically stable, and (b) asymptotically stable, autonomous and linear imply global exponential stability.

As a final example, consider the system obtained from Fig. 2 by replacing the linear springs by nonlinear ones. Specifically, replace the linear spring restoring forces $k_1 u$, $k_2 v$, kw by $g_1(u)$, $g_2(v)$ and $g(w)$, respectively, where $g_1: R \to R$, $g_2: R \to R$, and $g: R \to R$ are assumed to be differentiable. Then linear system (2) will be replaced by the nonlinear system

$$\begin{aligned} &m_1\ddot{x}_1+g_1(x_1)+g(x_1-x_2)+B_1\dot{x}_1+B(\dot{x}_1-\dot{x}_2)=0,\\ &m_2\ddot{x}_2+g_2(x_2)-g(x_1-x_2)+B_2\dot{x}_2-B(\dot{x}_1-\dot{x}_2)=0, \end{aligned} \tag{15}$$

where it is assumed that g_1, g_2 and g satisfy the conditions $g_1(0)=g_2(0)=g(0)=0$ and $g_1'(x)>0$, $g_2'(x)>0$, and $g'(x)>0$ for all $x\neq 0$. Once more it will be assumed that $B_1\geqq 0$, $B_2\geqq 0$, $B\geqq 0$, and $B_1+B_2+B>0$.

Next, to (15) we adjoin the outputs

$$\begin{aligned} y_1&=B\dot{x}_2-(B+B_1)\dot{x}_1,\\ y_2&=B\dot{x}_1-(B+B_2)\dot{x}_2,\\ y_3&=[(B+B_1)/m_1]\cdot[g_1(x_1)+g(x_1-x_2)]-(B/m_2)\cdot[g_2(x_2)-g(x_1-x_2)],\\ y_4&=-(B/m_1)\cdot[g_1(x_1)+g(x_1-x_2)]-[(B+B_2)/m_2]\cdot[g_2(x_2)-g(x_1-x_2)]. \end{aligned} \tag{16}$$

In studying the asymptotic stability of the trivial solution of system (15) we check when the system (15), (16) is distinguishable. We accomplish this by considering several cases.

Case 1. $B(B_1+B_2)+B_1B_2\neq 0$. If all $y_i\equiv 0$, then by (16), $\dot{x}_1\equiv\dot{x}_2\equiv 0$. Also, $g_1(x_1)\equiv -g(x_1-x_2)=-g_2(x_2)$. Since $x_1g_1(x_1)>0$ if $x_1\neq 0$ and $x_2g_2(x_2)>0$ if $x_2\neq 0$, then $x_1\equiv x_2\equiv 0$. Thus (15), (16) is distinguishable in this case and system (15) is asymptotically stable in the large.

Case 2. $B_1>0$, $B=B_2=0$. If all $y_i\equiv 0$, then from (16) we see that $\dot{x}_1\equiv 0$ and so $x_1\equiv c_1$ is constant and $g_1(c_1)=-g(c_1-x_2(t))$. Thus $x_2=c_2=-g^{-1}(-g_1(c_1))+c_1$ is constant. The only constant solution of (15) is the trivial one. Thus, system (15), (16) is distinguishable in this case and system (15) is asymptotically stable in the large.

Case 3. $B_2>0$, $B=B_1=0$. Using an identical argument as in Case 2, it follows that system (15), (16) is distinguishable and system (15) is asymptotically stable in the large in this case.

Case 4. $B_1=B_2=0$, $B>0$. This case is more complicated. If all $y_i\equiv 0$, then $\dot{x}_1\equiv\dot{x}_2$ and

$$[g_1(x_1)+g(x_1-x_2)]/m_1=[g_2(x_2)-g(x_1-x_2)]/m_2. \tag{17}$$

If the two masses can be made to move in synchronism, i.e., if

$$g_1(x)/m_1=g_2(x)/m_2$$

in some interval containing the origin, then (17) is possible with nonzero x_1 and $x_2\equiv x_1$. Under such conditions, system (15), (16) is not distinguishable and the trivial solution of (15) is not asymptotically stable.

Conversely, if system (15), (16) is not distinguishable, then $\dot{x}_1\equiv\dot{x}_2$ and (17) is true. Thus $x_1-x_2=c$ is constant. Substitute $x_1=x+c$, $x_2=x$ and $x=0$ into (17). If $c\neq 0$, then one side of (17) is positive and the other side is negative. Thus, c must be zero and $g_1(x)/m_1=g_2(x)/m_2$ for all x in some closed interval I which contains the origin in its interior. If there is no such interval I, then in this case (Case 4) the system (15), (16) is distinguishable and the trivial solution of system (15) is asymptotically stable in the large.

5. Concluding remarks. We re-emphasize that in the present results, the asymptotic stability of the equilibrium of a system is ascertained by (a) determining the stability of the equilibrium of the corresponding conservative system, and (b) determining the attractivity of the equilibrium by examining the topological properties (i.e., an observability condition) of the entire damped system. Step (a) is easily verified for linear as well as for nonlinear systems. Step (b) is also easily verified for linear systems and for certain classes of nonlinear systems (e.g., nearly linear systems); however, in general, the verification of step (b) for nonlinear systems may be quite difficult. In any event, the present results provide added insight into the mechanisms of stability (and stabilization) for a large class of dynamical systems.

REFERENCES

[1] E. A. BARBASIN AND N. N. KRASOVSKII, *Stability of motion in the large*, Dokl. Akad. Nauk SSSR, 86 (1952), p. 3.

[2] H. GOLDSTEIN, *Classical Mechanics*, Addison-Wesley, Reading, MA, 1950.

[3] W. HAHN, *Stability of Motion*, Springer-Verlag, New York, 1967.

[4] J. P. LASALLE, *Asymptotic stability criterion*, Proc. Symposia in Applied Mathematics, Vol. 13, American Mathematical Society, Providence, RI, 1962, pp. 299–307.

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

[6] J. J. LEVIN, *On the global asymptotic behavior of nonlinear systems of differential equations*, Arch. Rational Mech. Anal. 6 (1960), pp. 65–74.

[7] D. L. RUSSELL, *Mathematics of Finite Dimensional Control Systems*, Marcel Dekker, New York, 1979.

[8] P. P. VARAIYA AND R. LIU, *Normal form and stability of a class of coupled nonlinear networks*, IEEE Trans. on Circuit Theory, CT-13, (1966), pp. 413–418.

[9] J. A. WALKER AND W. E. SCHMITENDORF, *A simple test for asymptotic stability in partially dissipative symmetric systems*, Trans. ASME Ser. G. J. Dynamic Systems, Measurement and Control, 95 (1973), pp. 1120–1121.

[10] E. T. WHITTAKER, *A Treatise on Analytical Dynamics of Particles and Rigid Bodies*, 4th ed., Cambridge University Press, London, 1970.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 2, March 1980

THE OPTIMAL STRATEGY IN THE CONTROL PROBLEM ASSOCIATED WITH THE HAMILTON-JACOBI-BELLMAN EQUATION*

AVNER FRIEDMAN† AND PIERRE-LOUIS LIONS‡

Abstract. Consider the Hamilton–Jacobi–Bellman equation $\max_m \{A_m u(x) - f_m(x)\} = 0$ a.e. in R^n, where A_m ($m = 1, 2, \cdots$) are the infinitesimal generators of diffusion processes with constant coefficients and with discount $c_m \geqq \alpha > 0$. It is known that the solution can be represented as the optimal cost functional in which one can switch from one stochastic system to another without penalty. In this paper it is shown that if, for some k, $A_k f_m(x) - A_m f_k(x) \geqq c > 0$ for all $m \neq k$, $|x| > R$, then $A_k u(x) - f_k(x) = 0$ if $|x| > R_1$ for some R_1 sufficiently large; that means that the optimal strategy when $|x| > R_1$ is to stay with the diffusion and cost associated with A_k, f_k.

1. The main result. For each positive integer m, let $\sigma^m = (\sigma^m_{ij})$ be an $n \times n$ matrix of constants, and let $b^m = (b^m_i)$ be an n-vector of constants. Let

$$a^m_{ij} = \frac{1}{2} \sum_{k=1}^{n} \sigma^m_{ik} \sigma^m_{jk}$$

and introduce the (generally degenerate) elliptic operator

$$A_m v \equiv -\sum_{i,j=1}^{n} a^m_{ij} \frac{\partial^2 v}{\partial x_i\, \partial x_j} - \sum_{i=1}^{n} b^m_i \frac{\partial v}{\partial x_i} + c^m v, \tag{1.1}$$

where c^m are constants. We assume that

$$|a^m_{ij}| \leqq C_0, \qquad |b^m_i| \leqq C_0, \qquad c^m \geqq \alpha, \tag{1.2}$$

where C_0, α are positive constants. We also assume that there exist numbers $\theta_l \in (0, 1)$ $(1 \leqq l \leqq n_0)$ and integers $1 \leqq m_1 < m_2 < \cdots < m_{n_0}$ such that

$$\sum_{l=1}^{n_0} \theta_l = 1, \qquad \sum_{l=1}^{n_0} \sum_{i,j=1}^{n} \theta_l a^{m_l}_{ij} \xi_i \xi_j \geqq \nu |\xi|^2 \qquad (\nu > 0) \tag{1.3}$$

for all $\xi \in R^n$ (This assumption is not essential; see § 3, Remark 6).

Let $f_m(x)$ be functions in $W^{2,\infty}(R^n)$, satisfying

$$\|f_m\|_{W^{2,\infty}(R^n)} \leqq C_1; \tag{1.4}$$

the constants C_1 and C_0, α are independent of m.

Consider the Bellman equation

$$\sup_{m \geqq 1} \{A_m u(x) - f_m(x)\} = 0 \quad \text{a.e. in } R^n. \tag{1.5}$$

THEOREM 1.1. *There exists a unique solution $u(x)$ of* (1.5) *in* $W^{2,\infty}(R^n)$.

This theorem is due to P. L. Lions [5]; under more restrictive assumptions it was proved earlier by Krylov [2], [3]; see also [1], [4], [6] for the study of (1.5) in case of the Dirichlet problem.

The solution of (1.5) has the probabilistic interpretation

$$u(x) = \inf_v J(x, v), \tag{1.6}$$

* Received by the editors June 11, 1979.

† Department of Mathematics, Northwestern University, Evanston, Illinois 60201. The work of this author was supported in part by the National Science Foundation under Grant MCS-7817204.

‡ Labortoire d'Analyse Numérique, Université de Paris, 75230 Paris 5e, France.

where $v = v(t)$ is any nonanticipative control function taking values $1, 2, 3, \cdots$,

$$J(x, v) = E\left[\int_0^\infty f_{v(t)}(y_x(t, v)) \exp\left[-c^{v(t)}t\right] dt\right] \tag{1.7}$$

and $y_x(t, v)$ is the stochastic integral defined by

$$dy(t) = \sigma^{v(t)}\, dw(t) + b^{v(t)}\, dt, \qquad y(0) = x,$$

where $w(t)$ is an n-dimensional Brownian motion in the canonical Wiener space.

Thus, in the cost function $J(x, v)$ one may switch from any diffusion process (corresponding to σ^m, b^m) with its corresponding running cost f_m and discount c^m to any other one without any cost for the switching. The question naturally arises: Which is the best diffusion to choose at a particular point x? Analytically, the problem can be formulated as follows:

(1.8) For a specific k, when does the equality
$\sup_{m\geqq 1}\{A_m u(x) - f_m(x)\} = A_k u(x) - f_k(x)$ hold?

The purpose of this paper is to give *a sufficient condition under which the equality in* (1.8) holds. The main result is contained in the following theorem.

THEOREM 1.2. *Suppose there exist constants $R > 0$ and $c > 0$ such that*

$$A_k f_m(x) - A_m f_k(x) \geqq c \quad \textit{for all } m \neq k, \quad |x| > R. \tag{1.9}$$

Then there exists a number $R_1 > R$ such that

$$\sup_{m\geqq 1}\{A_m u(x) - f_m(x)\} = A_k u(x) - f_k(x) \quad \textit{if } |x| > R_1. \tag{1.10}$$

The proof is given in § 2. In § 3 we show that the condition (1.9) is rather sharp, and make some remarks on generalizations of Theorem 1.2.

2. Proof of Theorem 1.2. The proof is divided into two parts: in part I we show that it is sufficient to consider the case where $k = 1$, all the operators are uniformly elliptic, and (1.5) is replaced by a penalized system approximating (1.5). In part II we prove the theorem in this case, applying A_1 to the system and using some arguments reminiscent to some which occur in variational inequalities.

Part I. Without loss of generality we may take $k = 1$. Then (1.9) becomes

$$A_1 f_m - A_m f_1 \geqq c \quad \text{if } |x| > R, \quad m \geqq 2. \tag{2.1}$$

Without loss of generality we may assume that

(2.2) the A_m are nondegenerate elliptic,
with modulus of ellipticity independent of m

Indeed, otherwise we replace each A_m by $A_m^\varepsilon = A_m - \varepsilon\Delta$ (Δ = Laplacian in R^n). In view of (1.4), the condition (2.1) remains true for the A_m^ε (with another $c > 0$) provided ε is sufficiently small. As shown in [5], the corresponding solution u^ε of the Hamilton–Jacobi–Bellman equation satisfies: $u^\varepsilon(x) \to u(x)$ if $\varepsilon \to 0$. If we can prove that

$$A_1^\varepsilon u^\varepsilon - f^1 = 0 \quad \text{for } |x| > R_1 \tag{2.3}$$

(with R_1 independent of ε) then it would follow that also

$$A_1 u - f^1 = 0 \quad \text{if } |x| > R_1.$$

Thus it remains to prove the theorem under the assumption (2.2) (with R_1 independent of the modulus of ellipticity of the A_m).

Let $\beta(t)$ be a C^∞ function satisfying

$$\beta(t)=0 \quad \text{if } t<0, \qquad 0<\beta'(t)\leqq 1 \quad \text{and} \quad \beta''(t)\geqq 0 \quad \text{if } t\geqq 0 \tag{2.4}$$

and set

$$\beta_\varepsilon(t)=\frac{\beta(t)}{\varepsilon}. \tag{2.5}$$

Consider the system of elliptic equations

$$A_m u_m+\beta_\varepsilon(u_m-u_{m+1})=f_m, \qquad 1\leqq m\leqq N, \tag{2.6}$$

in R^n, where N is a fixed positive integer, to be taken arbitrarily large later on.

In [1], [4], [6] such a system is considered (with the assumption (2.2)) in a bounded domain Ω, with boundary conditions $u_m=0$ on $\partial\Omega$, and it is shown that the solution $u_m=u_m^{\varepsilon,N}$ satisfies

$$u_m(x)\to u(x) \quad \text{as } \varepsilon\to 0, \quad N\to\infty, \tag{2.7}$$

where $u(x)$ is the solution of the Hamilton–Jacobi–Bellman equation in Ω with $u=0$ on $\partial\Omega$. The method of proof shows that by taking $\Omega=\{x;|x|<1/k\}$, $k\to\infty$ we obtain the assertion (2.7) uniformly with respect to k; in fact the corresponding solution $u_m^{\varepsilon,N,k}$ satisfies

$$u_m^{\varepsilon,N,k}\to\tilde{u}_m^{\varepsilon,N}, \tag{2.8}$$

where $\tilde{u}_m^{\varepsilon,N}$ is the unique solution in $W^{2,\infty}(R^n)$ of (2.6), and

$$\tilde{u}_m^{\varepsilon,N}\to u \text{ as } \quad \varepsilon\to 0, \quad N\to\infty. \tag{2.9}$$

From now on we denote $\tilde{u}_m^{\varepsilon,N}$ by u_m. In view of (2.9), it suffices to show that

$$A_1u_2-f_1\geqq 0 \quad \text{if } |x|>R_1, \tag{2.10}$$

where R_1 is independent of ε, N.

Part II. We shall suppose that

$$A_1u_2(x^0)-f_1(x^0)<0 \quad \text{for some } x^0, \quad |x^0|>R_1 \tag{2.11}$$

and derive a contradiction for a suitably large R_1.

Consider the functions

$$z_m(x)=[A_1u_m(x)-f_1(x)]+\gamma|x-x^0|^2, \qquad \gamma>0 \tag{2.12}$$

in the set

$$G_\rho=B_\rho(x^0)\cap\{\min_{1\leqq m\leqq N}[A_1u_m(x)-f_1(x)]<0\}, \tag{2.13}$$

where

$$B_\rho(x^0)=\{x;|x-x^0|<\rho\}$$

and ρ is a sufficiently large positive number to be determined later on.

Notice that (since $f\in W^{2,\infty}$) the u_m belong to $W_{\text{loc}}^{4,p}$ for any $p<\infty$, and therefore, G_ρ is an open set. On the part $\partial G_\rho\cap B_\rho(x^0)$ of its boundary,

$$A_1u_m-f_1\geqq 0 \quad \text{for all } 1\leqq m\leqq N;$$

hence

$$z_m > 0 \quad \text{on } \partial G_\rho \cap B_\rho(x^0). \tag{2.14}$$

In view of the convexity of $\beta(t)$,

$$c^1 \beta_\varepsilon(u_m - u_{m+1}) \leqq c^1 \beta'_\varepsilon(u_m - u_{m+1})(u_m - u_{m+1}).$$

It follows that

$$A_1[A_m u_m + \beta_\varepsilon(u_m - u_{m+1})] \leqq A_m A_1 u_m + \beta'_\varepsilon(u_m - u_{m+1}) A_1(u_m - u_{m+1})$$
$$- \beta''_\varepsilon(u_m - u_{m+1}) \sum a^1_{ij} \frac{\partial}{\partial x_i}(u_m - u_{m+1}) \frac{\partial}{\partial x_j}(u_m - u_{m+1}).$$

Applying A_1 to both sides of (2.6) and using the last relation and the inequality $\beta''_\varepsilon(t) \geqq 0$, we find that

$$A_m(A_1 u_m) + \beta_{m,\varepsilon}(A_1 u_m - A_1 u_{m+1}) \geqq A_1 f_m,$$

where

$$\beta_{m,\varepsilon} \equiv \beta'_\varepsilon(u_m - u_{m+1}) = \frac{1}{\varepsilon} \beta'(u_m - u_{m+1}) \geqq 0. \tag{2.15}$$

Hence

$$A_m z_m + \beta_{m,\varepsilon}(z_m - z_{m+1}) \geqq \zeta_m, \tag{2.16}$$

where

$$\zeta_m = A_1 f_m - A_m f_1 + \gamma A_m(|x - x^0|^2).$$

In view of (1.2),

$$A_m(|x - x^0|^2) \geqq \alpha |x - x^0|^2 - C C_0(1 + |x - x^0|).$$

Hence, by (2.1),

$$\zeta_m(x) > \frac{c}{2} \quad \text{if } x \in G_\rho, \quad 2 \leqq m \leqq N, \tag{2.17}$$

$$\zeta_1(x) \geqq -C\gamma \quad \text{if } x \in G_\rho \tag{2.18}$$

provided ρ is such that

$$G_\rho \subset \{x; |x| > R\} \tag{2.19}$$

and provided γ is sufficiently small, say $\gamma \leqq \gamma_0$; γ_0 depends on α and C_0, but is independent of ε, N and the modulus of ellipticity of the A_m.

Since $|A_1 u_m - f_1| \leqq C$, for any γ we can choose $\rho = \rho(\gamma)$ such that

$$z_m > 0 \quad \text{on } \partial B_\rho(x^0); \tag{2.20}$$

from now on we fix $\rho = \rho(\gamma)$.

Let

$$M_i = \min_{x \in \bar{G}_\rho} z_i(x).$$

Now, if (2.11) holds with R_1 sufficiently large, then (2.19) holds; we shall show that this is impossible if $\rho = \rho(\gamma)$ and γ is sufficiently small.

Clearly (2.11) implies that $x^0 \in G_\rho$ and $z_2(x^0) < 0$. Therefore $M_2 < 0$. Let y be a point in $\bar{G}_\rho$ such that $z_2(y) = M_2$. In view of (2.14), (2.20), y belongs to G_ρ.

Consequently, $A_2z_2(y)<0$. From (2.16), (2.17) we then obtain

$$\beta_{2,\varepsilon}(z_2(y)-z_3(y))>\frac{c}{2}.$$

Since $\beta_{2,\varepsilon}\geqq 0$ we deduce that $z_2(y)-z_3(y)>0$ and, using also (2.15) we obtain

$$z_2(y)-z_3(y)\geqq(\beta')^{-1}\left(\frac{c}{2}\varepsilon\right).$$

It follows that

$$M_2-M_3>(\beta')^{-1}\left(\frac{c}{2}\varepsilon\right)$$

and, in particular, $M_3<0$.

Proceeding in this way step by step, we get

$$M_j-M_{j+1}>(\beta')^{-1}\left(\frac{c}{2}\varepsilon\right)\quad \text{if } 2\leqq j\leqq N\quad (M_{N+1}=M_1) \tag{2.21}$$

and, in particular,

$$M_1<M_2-(N-1)(\beta')^{-1}\left(\frac{c}{2}\varepsilon\right).$$

Let $\tilde{y}$ be a point in $\bar{G}_\rho$ such that $z_1(\tilde{y})=M_1$. In view of (2.14), (2.20), $\tilde{y}$ belongs to G_ρ.

Using (2.18) and the inequality $A_1z(\tilde{y})<0$ in the relation (2.16) for $m=1$, we get

$$\frac{1}{\varepsilon}(z_1(\tilde{y})-z_2(\tilde{y}))\geqq -C\gamma. \tag{2.22}$$

Adding the inequalities (2.21), (2.22) we obtain

$$C\gamma\varepsilon>(N-1)(\beta')^{-1}\left(\frac{c}{2}\varepsilon\right)\geqq(N-1)\frac{c}{2}\varepsilon\quad \text{if } \varepsilon\leqq\varepsilon_0$$

since $\beta'(t)\leqq 1$. This is impossible if γ is sufficiently small.

3. Remarks and generalizations.

Remark 1. R_1 in Theorem 1.2 can be determined explicitly in terms of c, R and the $W^{2,\infty}$ norm of f_m.

Remark 2. The proof of Theorem 1.2 remains valid if we replace (2.1) by the weaker condition

$$A_kf_m(x)-A_mf_k(x)\geqq\gamma_m,\qquad \gamma_m>0,\quad m\neq k \tag{3.1}$$

provided

$$\frac{1}{N}\sum_{\substack{m=1\\ m\neq k}}^{N}\gamma_m\geqq c,\qquad c>0. \tag{3.2}$$

Remark 3. If some of the c_m in (2.1) are equal to zero then the assertion of Theorem 1.2 is generally false. Consider for example the case of

$$\max\,(A_1u, A_2u, A_3u-f_3)=0,$$

where $A_2w \equiv A_1w + w$, $f_3 \geqq 0$, $A_1f_3 \geqq c > 0$. From (1.6), (1.7) we see that

$$u > 0. \tag{3.3}$$

Now, if the assertion of the theorem were valid in this case (with $k = 1$) then $A_1u = 0$ if $|x| > R_1$. Since $A_2u \leqq 0$, we deduce that $u = A_2u - A_1u \leqq 0$ if $|x| > R_1$, contradicting (3.3).

Remark 4. For two operators A_1, A_2, Theorem 1.2 asserts that

$$\text{if } A_1f_2 - A_2f_1 \geqq c > 0 \text{ for } |x| > R, \quad \text{then } A_1u = 0 \text{ if } |x| > R_1. \tag{3.4}$$

In the special case of $A_1w \equiv w$, $f_1 \equiv 0$ this gives a well-known result on the support of solutions of variational inequalities, namely, if

$$u \leqq 0, \qquad A_2u - f_2 \leqq 0, \qquad u(A_2u - f_2) = 0 \quad \text{a.e. in } R^n$$

and if $f_2 \geqq c > 0$ for $|x| > R$, then $u = 0$ if $|x| > R_1$. In this special case, the proof of Theorem 1.2 extends to A_1 with variable coefficients.

Remark 5. To motivate the condition (2.1) notice that if the assertion of Theorem 1.2 holds then, for $|x| > R_1$, $v_m = A_mu - f_m$ satisfies

$$A_1v_m = A_1A_mu - A_1f_m = A_mA_1u - A_1f_m = A_mf_1 - A_1f_m. \tag{3.5}$$

Since also $v_m \leqq 0$, the right-hand side of (3.5) cannot be "too positive." In Theorem 1.2 we assume that this right-hand side is uniformly negative.

Remark 6. The proof of Theorem 1.2 is actually local; it shows that (1.10) holds in an open set G provided (1.9) holds in a ρ-neighborhood of G, where ρ is a sufficiently large postive number (independent of G). In particular, if (1.9) holds in a half space $x_n > R$, then (1.10) holds for $x_n > R_1$ provided R_1 is sufficiently large.

Remark 7. If we drop the condition (1.5), then existence and uniqueness of a solution ("suitably" regular) is given in [5]; the proof of Theorem 1.2 for this case remains valid without any changes.

Remark 8. It seems natural to extend (1.8) by asking: When does the relation

$$\sup_{m \geqq 1} \{A_mu(x) - f_m(x)\} = \max \{A_1u(x) - f_1(x), A_2u(x) - f_2(x)\} \tag{3.6}$$

hold? The answer cannot be very simple. This is due to the fact that

$$\max (A_mg, A_mh) - A_m \max (g, h)$$

is a positive distribution, in general.

Let us explain the difficulties in the following one-dimensional example. Consider

$$A_1 = A_2 = A_3 = -\frac{d^2}{dx^2} + \alpha^2, \qquad f_1 = -f, \quad f_2 = 0.$$

Then the Hamilton–Jacobi–Bellman equation for these operators reduces to

$$\max \{A_1u + f^+, A_3u - f_3\} = 0, \tag{3.7}$$

and (3.6) reduces to

$$A_1u + f^+ = 0 \quad \text{if } |x| > R. \tag{3.8}$$

In general,

$$A_3f^+ = -\sum a_n\delta(x - x_n) + g, \qquad g \text{ bounded,}$$

where $a_n > 0$, $\delta(x)$ the Dirac function. Suppose

$$A_1 f_3 + A_3 f^+ = c - \sum a_n \delta(x - x_n) \tag{3.9}$$

and set $w = A_3 u - f_3$. If (3.8) holds then

$$A_1 w = -A_3 f^+ - A_1 f_3. \tag{3.10}$$

Writing w explicitly, in an interval $a < x < \infty$, and using (3.9), we get

$$w(x) = \frac{w(a)}{\alpha^2} e^{-\alpha(x-a)} + \sum_{x_n > a} \frac{a_n}{\alpha^2} e^{-\alpha(x-x_n)^+} - \frac{c}{\alpha^2}. \tag{3.11}$$

Since we must have $w(x) \leqq 0$, we obtain the necessary condition

$$\sum_{x_n > a} a_n e^{-\alpha(x-x_n)^+} \leqq c \quad \text{if } x - a \text{ is sufficiently large.} \tag{3.12}$$

On the other hand, this condition is nearly sufficient. Indeed, we proceed as in the proof of Theorem 1.2, assuming

$$A_1 u(x^0) + f^+(x^0) < 0,$$

and designating by (a', b') the largest interval containing x^0 where the inequality $A_1 u + f^+ < 0$ holds and $a' \geqq R$. Then

$$A_3 u - f_3 = 0 \quad \text{if } a' < x < b',$$

and $\tilde{w} \equiv A_1 u + f^+$ satisfies $A_1 \tilde{w} = A_3 f^+ + A_1 f_3$ in (a', b'). Representing $\tilde{w}$ analogously to (3.11) with $a = a'$ and taking $x = x^0$, the inequality $\tilde{w}(x^0) < 0$ gives

$$\frac{\tilde{w}(a')\, e^{-\alpha(x^0 - a')}}{\alpha^2} - \sum_{a' < x_n < b'} a_n e^{-\alpha(x^0 - x_n)^+} + c < 0;$$

this is impossible if

$$\sum_{a' < x_n < b'} a_n e^{-\alpha(x^0 - x_n)^+} < c - \eta, \tag{3.13}$$

where

$$\eta = \frac{\tilde{w}(a')}{\alpha^2} e^{-\alpha(x^0 - a')}$$

is either equal to zero or else is positive and very small if x^0 is sufficiently large. We have thus shown that if (3.13) holds then $A_1 u + f^+ = 0$ at x^0 (and if $\tilde{w}(a') = 0$ then $A_1 u + f^+ = 0$ in (a', b')).

In conclusion, the location and size of the measures arising in $A_3 \max(A_1 u - f_1, A_2 u - f_2)$ affect the answer to the question (3.6) in the above special case, and similarly also in the general case. If $A_1 \neq A_2$, these measures involve, in addition to f_1, f_2, the function u and its first two derivatives.

REFERENCES

[1] L. C. EVANS AND A. FRIEDMAN, *Optimal switching and the Dirichlet problem for the Bellman equation*, Trans. Amer. Math. Soc., to appear.

[2] N. V. KRYLOV, *Control of a solution of a stochastic integral equation*, Theor. Probability Appl., 17 (1972), pp. 114–130.

[3] ———, *On control of the solution of a stochastic integral equation with degeneration*, Math. USSR-Izv., 6 (1972), pp. 249–262.

[4] P. L. LIONS, *Résolution de problèmes généraux de Bellman–Dirichlet*, C. R. Acad. Sci. Paris Sér A–B, 287 (1978), pp. 747–750. (Detailed paper to appear.)

[5] ———, *Contrôle de diffusions dans* R^N, C. R. Acad. Sci. Paris Sér A–B, 288 (1979), pp. 339–342. (Detailed paper to appear.)

[6] P. L. LIONS AND J. L. MENALDI, *Contrôl d'integrales stochastiques et équations de Bellman*, C. R. Acad. Sci. Paris Sér A–B, 287 (1978), pp. 409–413.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 2, March 1980

0363-0129/80/1802-0008$01.00/0

EXISTENCE DE SOLUTION ET ALGORITHME DE RESOLUTION NUMERIQUE, DE PROBLEME DE CONTROLE OPTIMAL DE DIFFUSION STOCHASTIQUE DEGENEREE OU NON*

JEAN PIERRE QUADRAT†

Motivation et introduction. Donnons des exemples pratiques de contrôle stochastique que nous avons eu à résoudre.

1. *Une gestion de réservoir* Delebecque–Quadrat [8] (Problème posé par EDF).

t le temps.

X_t désigne les apports dans le réservoir. Ils sont modélisés par une diffusion stochastique. $X_t \geqq 0$ cette diffusion sera donc dégénérée.

S_t le stock d'eau dans le réservoir à l'instant t, $S_{\max}$[resp $S_{\min}$] le stock maximum [resp minimum].

u_t le débit turbiné à l'instant t,

$P(S_t, u_t)$ la puissance fournie par les turbines lorsque le débit est u_t, le stock S_t.

$D(t)$ la demande d'électricité en puissance à l'instant t.

La puissance thermique à produire sera alors $D(t)-P(S_t, u_t)$, le coût associé sera:

$$C(D(t)-P(S_t, u_t)).$$

Le problème de contrôle stochastique s'écrit alors:

$$dX_t = b(t, X_t)\,dt + \sigma(t, X_t)\,dw_t \qquad \sigma = 0 \text{ pour } X<0$$

$$(0.1)\qquad dS_t = \begin{cases} -(X_t-u_t)^-\,dt & \text{si } S_t = S_{\max}, \\ (X_t-u_t)\,dt & \text{si } S_{\min} < S_t < S_{\max}, \\ (X_t-u_t)^+\,dt & \text{si } S_t = S_{\min},\ X_t \geqq u_t \end{cases}$$

$$\operatorname*{Min}_u E\int_0^T C(D(t)-P(S_t, u_t))\,dt.$$

2. *Un problème de croissance de firme* Bensoussan–Lesourne [9].

X_t trésorerie à l'instant t,

y_t capital investi à l'instant,

$f(y_t)(\lambda\,dt+dw_t)$ rendement du capital investi à l'instant (w_t est un brownien),

v_t investissement,

u_t dividende versé aux actionnaires,

τ temps de faillite (trésorerie $=0$).

Contraintes:

$$u_t \geqq 0,$$

$$v_t \geqq 0,$$

$$u_t + v_t \leqq \lambda f(y_t) \text{ investissement} + \text{dividende}$$

$$\leqq \text{rendement moyen de l'investissement.}$$

* Received by the editors November 30, 1977, and in revised form March 5, 1979.

† Institut de Recherche d'Informatique et d'Automatique, Rocquencourt, 78150 Le Chesnay, France.

Le problème de contrôle stochastique correspondant:

$$\tag{0.2} \begin{aligned} &\operatorname*{Max}_{v,u} E\int_0^T u_t e^{-it}\,dt,\\ &dX_t = f(y_t)(\lambda\,dt + dw_t) - v\,dt - u\,dt,\\ &dy_t = v\,dt. \end{aligned}$$

Le critère représente la maximisation des dividendes actualisés (i taux d'actualisation) versés aux actionnaires.

3. *Un problème de gestion de portefeuille* Merton [16]. Soit X_t le capital dont on dispose à l'instant t. On a le choix entre acheter des actions à rendement aléatoire, et un placement à rendement fixe. Notons u_t la proportion investie dans les actions. Notons:

$1 + dR_t^1$ le rendement des actions,

$1 + dR_t^2$ le rendement du deuxième placement.

On modélise:

$$\begin{aligned} &dR_t^1 = \alpha_1\,dt + \sigma\,dw_t, \qquad w_t \text{ brownien};\\ &dR_t^2 = \alpha_2\,dt. \end{aligned}$$

Soit C_t la consommation à l'instant du capital à l'instant, $f(C_t)$ la fonction d'utilité de cette consommation.

L'évolution du capital est alors donnée par:

$$\tag{0.3} \begin{aligned} &dX_t = X_t(u\,dR_t^1 + (1-u)\,dR_t^2) - C\,dt,\\ &\quad = X_t u(\alpha_1\,dt + \sigma\,dw_t) + (1-u)X_t\alpha_2\,dt - C\,dt, \qquad 0 \leqq u \leqq 1,\\ &\text{Le critère:}\\ &\operatorname*{Max}_{u,C} E\int_0^T f(C_t) + \phi(X_t). \end{aligned}$$

ϕ représente une fonction de legs, le critère représente la maximisation de l'utilité de la consommation plus le leg en fin de gestion.

Nous constatons que ces trois problèmes sont dégénérés, le terme de diffusion peut s'annuler. Dans le troisème problème le contrôle apparaît dans le terme de diffusion. Dans le premier problème, la deuxième équation d'évolution a un second membre discontinu.

Le but de ce travail est de donner des *théorèmes d'existence pour de tels problèmes*, et de *caractériser la solution optimale de façon à ce que l'on puisse la calculer effectivement.*

Un certain nombre de résultats existent dans la littérature Krylov–Nisio [12], Kushner–Chen-Fu-Yu [14], Fleming–Rishel [10], Sentis [21], Bismut [4], Kushner [13] mais aucun de ces travaux ne donne une réponse à ces trois problèmes.

Ce travail donne une réponse complète aux problèmes 1 et 3 et un théorème d'existence pour le problème 2 (la caractérisation des contrôles optimaux lorsqu'on arrête processus n'étant pas donné dans ce travail).

La méthode de résolution utilise deux techniques:

1. La formulation faible Stroock–Varadhan [22] de diffusion stochastique (problème de martingale).

2. Les techniques utilisées en contrôle déterministe décrivant le système commandé en terme de multiapplication Young [24], Castaing [6], Valadier [23], Ekeland–Temam [9], Sentis [21].

On est donc amené à définir le problème de martingale pour des multiapplications s.c.s. (semi continue superieurement).

On donne un théorème d'existence très général qui contient comme cas particulier des équations différentielles déterministes multivoques pour des multiapplications s.c.s. La méthode employée est celle utilisée dans Stroock–Varadhan [22] dans leur "invariance principle" montrant la convergence de chaîne de Markov vers des diffusions, et d'un lemme abstrait énoncé en 1.2.3.

Une fois l'existence assurée, pour de tels problèmes, on montre que l'ensemble des solutions au problème de martingale multivoque est un ensemble convexe compact de mesure de probabilités sur l'espace des fonctions continues sur $(0, T)$.

L'existence d'une solution au problème de contrôle stochastique en découle alors immédiatement.

La caractérisation du contrôle optimal se fait alors en déterminant une suite de mesures convergeant étroitement vers une solution optimale; cette suite de mesures étant obtenue comme solution de problème ce contrôle de chaîne de Markov. Les problèmes de contrôle de chaîne de Markov sont définis grâce à une technique très proche de celle employée dans Sentis [21] pour la résolution de problème de contrôle déterministe. L'idée d'approcher le problème de contrôle stochastique par un problème de contrôle de chaîne de Markov a été abondamment utilisé par Kushner [13] par example Kushner–Chen-Fu Yu [17]; les techniques employées ici sont différentes et permettent de résoudre complètement le problème, alors que dans Kushner–Chen-Fu Yu [14] (dans un cadre d'hypothèses moins général), le résultat obtenu peut s'énoncer ainsi, on construit un feedback meilleur que tout feedback lipschitzien. On donne en III un contre exemple montrant que ce résultat bien que pratiquement intéressant, est insuffisant. On peut, avec ces feedbacks être très loin du coût optimal (en fait, aussi loin qu'on veut).

On donne enfin une suite de problèmes de contrôle de chaîne de Markov, en temps discret, et à état discret qui converge vers une solution optimale. A chaque étape en temps il faut résoudre un problème de programmation mathématique (minimisation d'une forme linéaire sur un ensemble convexe) en dimension finie. Ce problème peut dans certaines applications être lourd à résoudre. On donne alors des résultats qui permettent d'obtenir un feedback meilleur que tout feedback lipschitzien (résultat du type Kushner–Chen-Fu Yu [14] dans un cadre plus général). Ces résultats peuvent avoir un intérêt pratique avec la restriction énoncée plus haut. Utilisant alors des résultats de Bismut [4] on montre que le feedback meilleur que tout feedback lipschitzien est optimal dans le *cas non dégénéré.* Ce qui permet de trouver des résultats du même type que ceux obtenus dans Goursat–Quadrat [11], Quadrat [19] par une méthode purement probabiliste, dans un cadre plus général, alors que dans cet article la méthode était basée essentiellement sur l'analyse numérique de l'équation de Bellman correspondante.

L'ensemble des résultats obtenus par les techniques des équations aux dérivées partielles est donné dans Bensoussan–Lions [1] bien que ne semblant pas pouvoir atteindre le degré de généralité obtenu par les méthodes probabilistes. Cette première méthode donne des résultats plus précis lorsqu'elle s'applique.

Signalons enfin une technique de semi-discrétisation en espace interprété en terme de processus ponctuel convergeant étroitement vers la diffusion, développé dans Robin [20].

Les résultats de cet article ont été annoncés dans Quadrat [25]. L'extension de ces résultats au cas des processus de diffusion avec sauts sera donné dans Quadrat [26].

PLAN.

1. Le problème de martingale.

1.1. Définition du problème de martingale.

1.1.1. Notations.

Soient

$$\Omega = C(0, T; R^m), \qquad X_t(\omega) = \omega_t,$$

$$F_t = \sigma(X_s, s \leqq t),$$

$F = F_T$ la tribu des boréliens de Ω,

$\mathcal{P}$ la tribu des prévisibles de $(\Omega \times [0, T])$,

la multiapplication

(H$_1$) $C\colon [0, T] \times R^m \to G = R^m \times S^+(m)$ *s.c.s.,*[1] *à valeur convexe dans un compact fixe noté M_C, où S_m^+ désigne le cône convexe des matrices symétriques non négatives.*

Si l'on désigne par p_1 et p_2 les projections

$$\begin{array}{lll} p_1\colon R^m \times S_m^+ \to R^m & & , \\ \quad (x, y) & p_1(x, y) = x & \\ p_2\colon R^m \times S_m^+ \to S_m^+ & & . \\ \quad (x, y) & p_2(x, y) = y & \end{array}$$

on appellera:

A la multiapplication $p_2 \circ C$,

B la multiapplication $p_1 \circ C$.

[1] s.c.s: semi continue supérieurement; s.c.i: semi continue inférieurement.

On notera:

$$M_b = \sup_{s,x} \sup_{b\in B(s,x)} |b|,$$

$$M_a = \sup_{s,x} \sup_{a\in A(s,x)} |a|.$$

Pour

$\varphi \in C_b^{1,2}([0, T]\times R^m)$,

$\hat{c} = (\hat{b}, \hat{a})\colon [0, T]\times \Omega \to R^m \times S_m^+$ processus prévisible à valeur dans C,

on désigne par:

$$L_{\hat{c}}\varphi(s, \omega) = \frac{\partial\varphi}{\partial t}(s, X_s(\omega)) + \sum_{i=1}^m \hat{b}_i(s, \omega)\frac{\partial\varphi}{\partial x_i}(s, X_s(\omega))$$
$$+ \sum_{i,j} \hat{a}_{ij}(s, \omega)\frac{\partial^2\varphi}{\partial x_i \partial x_j}(s, X_s(\omega)).$$

Pour

$$\tilde{c} = (\tilde{b}, \tilde{a})\colon [0, T]\times R^m \to R^m \times S_m^+,$$

on désignera par:

$$L_{\tilde{c}}\varphi(s, x) = \frac{\partial\varphi}{\partial t}(s, x) + \sum_{i=1}^m \tilde{b}_i(s, x)\frac{\partial\varphi}{\partial x_i}(s, x) + \sum_{i,j} \tilde{a}_{ij}(s, x)\frac{\partial^2\varphi}{\partial x_i \partial x_j}(s, x).$$

$\mathcal{M}_b(\Omega)$ désigne l'ensemble des mesures bornées muni de la topologie de la convergence étroite.

$\mathcal{M}_+^1(\Omega)$ désigne le convexe des lois de probabilités sur Ω.

1.1.2. Définition du problème de martingale. *Une mesure de probabilité P sur (Ω, F_t, F) sera appelée solution du problème de martingale pour le doublet (x, C) si:*

(i) $P(X_0 = x) = 1$

(ii) *il existe un processus $c(s, \omega)$ prévisible vérifiant:*

$$\hat{c}(s, \omega) \in C(s, X_s(\omega))$$

$$\varphi(t, X_t(\omega)) - \int_0^t L_{\hat{c}}\varphi(s, \omega)\, ds \quad \textit{est une } (P, F_t) \textit{ martingale.}$$

On désignera par:

$\mathcal{P}(K, C)$ l'ensemble ces *mesures de probabilités* sur Ω, *solution du problème de martingale* (x, C), $x \in K$, *K compact de R^m.*

1.2. Construction d'une mesure solution du problème de martingale.

1.2.1. Une famille de probabilités de transition. On se donne un nombre $n \in N$, on pose $h = T/n$. Pour ρ, α, β, >0;

notons:

$$\prod^{n,C,\rho,\alpha,\beta}(s, x) = \Big\{ \pi \in \mathcal{M}_+^1(R^m):$$

(1.1), (1.2) $$\left(\int (y-x)\pi(dy), \int (y-x)^{\otimes 2}\pi(dy)\right) = (b(s, x)h, a(s, x)h) \in C(s, x)h,$$

(1.3) $$\int |y-x|^\beta \pi(dy) \leq \rho h^\alpha \Big\}.$$

Lemme 1. $\Pi^{n,C,\rho,\alpha,\beta}$: $[0, T]\times R^m \to \mathcal{M}_+^1(R^m)$ *est à valeur relativement compacte.*

Démonstration. Grâce au critère de Prohorov, il suffit de montrer:

$$\forall \varepsilon > 0 \quad \exists M, \qquad \pi_{s,x}(|y-x| \geqq M) \leqq \varepsilon.$$

Or, l'inégalité de Tchebycheff donne:

$$\pi_{s,x}(|y-x| \geqq M) \leqq \frac{M_a^2 h}{M^2}.$$

En prenant $M \geqq M_a/\sqrt{h/\varepsilon}$ on obtient le résultat. □

LEMME 2. *Soit* φ: $R^m \to R$ *continue vérifiant*:

(i) $\exists M_1, M_2 : |z| \geqq M_1 \quad \Rightarrow \quad \varphi(z) \leqq |z|^\beta M_2.$

Soit une suite $\{\pi^n\} \in \mathcal{M}_+^1(R^m)$ *convergent étroitement vers* π *vérifiant* $\exists \beta' > \beta$ *et* ρ:

(ii) $\int |y-x|^{\beta'} \pi^n(dy) \leqq \rho$

alors:

(a) $\int |y-x|^{\beta'} \pi(dy) \leqq \rho$

(b) $\lim_{-n} \int \varphi(y-x_n)\pi^n(dy) = \int \varphi(y-x)\pi(dy)$ *pour toute suite* $x_n \to x$.

Démonstration. (a) Notons:

$$\Psi_M: R \to R$$

$$\Psi_M(x) = \begin{cases} x & \text{si} -M \leqq x \leqq M, \\ M & \text{si } x > M, \\ -M & \text{si } x < -M. \end{cases}$$

On a:

$$\rho \geqq \int |y-x|^{\beta'} \pi^n(dy) \geqq \int \Psi_M(|y-x|^{\beta'})\pi^n(dy) \xrightarrow[n\to\infty]{} \int \Psi_M(|y-x|^{\beta'})\pi(dy)$$

et donc:

$$\int \Psi_M(|y-x|^{\beta'})\pi(dy) \leqq \rho \quad \forall M,$$

$$\rho \geqq \sup_M \int \Psi_M(|y-x|^{\beta'})\pi(dy) = \int \sup_M \Psi_M(|y-x|^{\beta'})\pi(dy) = \int |y-x|^{\beta'}\pi(dy).$$

(1.4) (b)
$$\left|\int \varphi(y-x_n)\pi^n(dy) - \int \varphi(y-x)\pi(dy)\right|$$
$$\leqq \left|\int (\varphi(y-x_n)-\varphi(y-x))\pi^n(dy)\right| + \left|\int \varphi(y-x)(\pi^n-\pi)(dy)\right|,$$

$$\int |\varphi(y-x_n)-\varphi(y-x)|\pi^n(dy)$$
$$\leqq \int_{|y-x|\geqq M_3} |\varphi(y-x_n)-\varphi(y-x)|\pi^n(dy)$$
$$+\int_{|y-x|\leqq M_3} |\varphi(y-x_n)-\varphi(y-x)|\pi^n(dy),$$

$$\int_{|y-x|\leqq M_3} |\varphi(y-x_n)-\varphi(y-x)|\pi^n(dy) \leqq \sup_{|y-x|\leqq M_3} |\varphi(y-x_n)-\varphi(y-x)| \to 0$$
$$n \to \infty.$$

En choisissant M_3 suffisamment grand

$$\int_{|y-x|\geqq M_3} |\varphi(y-x_n)-\varphi(y-x)|\pi^n(dy)\leqq M_4 \int_{|y-x|\geqq M_3} |y-x|^\beta \pi^n(dy)+|x-x^n|^\beta M_6$$

$$\leqq \frac{M_4}{M_3^{\beta'-\beta}}\int |y-x|^{\beta'}\pi^n(dy)$$

$$+|x-x^n|^\beta M_6 \to 0, \qquad n\to\infty,$$

en faisant tendre n et $M_3\to\infty$; d'autre part

$$\left|\int \varphi(y-x)(\pi^n-\pi)(dy)\right|$$

$$\leqq \left|\int \Psi_M\circ\varphi(y-x)(\pi^n-\pi)(dy)\right| \tag{1.5}$$

$$+\frac{1}{M^{(\beta'-\beta)/\beta}}\int_{|\varphi(y-x)|\geqq M} |\varphi(y-x)|^{\beta'/\beta}|\pi^n-\pi|(dy)$$

$\Psi_M\circ\varphi(y-x)$ est continue donc:

$$\int \Psi_M\circ\varphi(y-x)(\pi^n-\pi)\,dy\to 0, \qquad n\to\infty. \tag{1.6}$$

Grâce à (i)

$$\int_{|\varphi(y-x)|\geqq M} |\varphi|^{\beta'/\beta}(y-x)|\pi^n-\pi|(dy)\leqq M_5\int |y-x|^{\beta'}|\pi^n-\pi|(dy)\leqq 2\rho M_5; \tag{1.7}$$

(1.4), (1.5), (1.6), (1.7)$\Rightarrow b$ en faisant tendre $M\to\infty$. □

PROPOSITION 0. *La multiapplication* $\Pi^{n,C,\rho,\alpha,\beta}:[0,T]\times R^m\to\mathcal{M}^1_+(R^m)$ *est s.c.s.*

Démonstration. Montrons qu'elle est de graphe fermé et donc grâce au Lemme 1 à valeur compacte:

Soit:

(s_k, x_k, π_k) une suite de $[0,T]\times R^m\times\mathcal{M}^1_+(R^m)$ convergeant vers (s,x,π). Montrons que $\pi\in\Pi^{n,C,\rho,\alpha,\beta}$

Pour cela il suffit de montrer que π vérifie (1.1), (1.2) et (1.3).

Or, (1.3) résulte du (a) du Lemme 2, tandis que (1.1) et (1.2) résulte de (b) du Lemme 2.

1.2.2. Une famille de probabilités sur Ω. A partir de la famille de probabilité de transition $\Pi^{n,C,\rho,\alpha,\beta}$ construisons une famille de probabilité sur Ω de la façon suivante:

Etant données une section borélienne $\pi_{s,x}$ de $\Pi^{n,C,\rho,\alpha,\beta}$ construisons la mesure notée

$$\begin{aligned}&\bar{P}^n_\pi \text{ sur } (R^{m\times(n+1)},\mathcal{B}) \text{ définie par:}^2\\ &\bar{P}^n_{\pi,y_0}=\pi_{h,y_0}(dy_1)\cdots\pi_{(n-1)h,y_{n-1}}(dy_n).\end{aligned} \tag{1.8}$$

Considérons maintenant la variable aléatoire interpolation linéaire.

$$(R^{m\times(n+1)},\mathcal{B})\xrightarrow{I_n}(\Omega,F_T),$$

$$I_n(y_0,y_1,\cdots y_n)(t)=y_i+\frac{y_{i+1}-y_i}{h}(t-ih), \qquad t\in[ih,(i+1)h].$$

[2] $\mathcal{B}$ désigne la tribu des boréliens.

On désignera par P^n_{π,y_0} la mesure image de $\bar{P}^n_{\pi,y_0}$ par I_n. Notons alors:

(1.9) $\mathcal{P}(n, K, C, \rho, \alpha, \beta) = \{P^n_{\pi, y_0}$: π section borélienne de $\Pi^{n,C,\rho,\alpha,\beta}$, $y_0 \in K\}$

de même:

$$\mathcal{P}(N, K, C, \rho, \alpha, \beta) = \bigcup_{n \in N} \mathcal{P}(n, K, C, \rho, \alpha, \beta).$$

Lemme 3. *$\mathcal{P}(N, K, C, \rho, \alpha, \beta)$ est étroitement relativement compacte pour $\forall \alpha, \beta$ $\alpha > 1$, $4 \geqq \beta > 2$.*

Démonstration. On utilise le critère de relative étroite compacité suivant [3 th. 12.3 Pb 7, p. 102]: (a) $\forall \varepsilon > 0$ $\exists L$ compact de R^m:

$$P(X(0) \in L) \geqq 1 - \varepsilon \quad \forall P \in \mathcal{P}(N, K, C, \rho, \alpha, \beta).\atop{2<\beta\leqq 4}$$

(b) $\exists \gamma \geqq 0$ et $\delta > 1$ et une fonction continue non décroissante F tels que:

$$E_P\{|X(t_2) - X(t)|^\gamma |X(t) - X(t_1)|^\gamma\} \leqq |F(t_2) - F(t_1)|^\delta,$$
$$\forall P \in \mathcal{P}(N, K, C, \rho, \alpha, \beta) \qquad 2 < \beta \leqq 4 \text{ et } \alpha > 1,$$
$$\forall t_1, t, t_2, \quad 0 \leqq t_1 \leqq t \leqq t_2 \leqq T.$$

On obtient (a) en prenant $L = K$.

Démontrons le (b): Soit: $t_1 \leqq t \leqq t_2$, β défini en 1.3., $\gamma = \beta/2$,

$$t_1'' = \left(\left[\frac{t_1}{h}\right] + 1\right) h, \quad t_2' = \left[\frac{t_2}{h}\right] h, \quad t' = \left[\frac{t}{h}\right] h, \quad t'' = \left(\left[\frac{t}{h}\right] + 1\right) h,$$

où $[\cdot]$ désigne la partie entière.

Plaçons nous dans le cas où $t_1'' < t' < t'' < t_2'$, les autres cas conduisant à des démonstrations analogues. On a:

$$\begin{aligned}(1.10)\qquad & E|X_{t_2} - X_t|^\gamma |X_t - X_{t_1}|^\gamma \\ & \leqq M(\beta) E(|X_{t_2} - X_{t_2'}|^\gamma \\ & \qquad + |X_{t_2'} - X_{t''}|^\gamma + |X_{t''} - X_t|^\gamma)(|X_t - X_{t'}|^\gamma + |X_{t'} - X_{t_1''}|^\gamma + |X_{t_1''} - X_{t_1}|^\gamma) \\ & \leqq M(\beta)(S_1 + S_2 + S_3)\end{aligned}$$

avec

$$S_1 = E\{E(|X_{t_2} - X_{t_2'}|^\gamma + |X_{t_2'} - X_{t''}|^\gamma | F_{t''})(|X_t - X_{t'}|^\gamma + |X_{t'} - X_{t_1''}|^\gamma + |X_{t_1''} - X_{t_1}|^\gamma)\}$$
$$S_2 = E\{|X_{t''} - X_t|^\gamma |X_t - X_{t'}|^\gamma\},$$
$$S_3 = E\{E(|X_{t''} - X_t|^\gamma | F_{t'})(|X_{t'} - X_{t''}|^\gamma + |X_{t_1''} - X_{t_1}|^\gamma)\};$$

or:

$$\begin{aligned}& E(|X_{t_2} - X_{t_2'}|^\gamma | F_{t''}) = \frac{|t_2 - t_2'|^\gamma}{h^\gamma} E((X_{t_2'+h} - X_{t_2'})^\gamma | F_{t''}) \\ (1.11)\qquad & \qquad \leqq \frac{|t_2 - t_2'|^\gamma}{h^\gamma} h^{\alpha/2} \leqq M_1 |t_2 - t_2'|^{\inf(\alpha/2, \gamma)}\end{aligned}$$

$$E(|X_{t_2'} - X_{t''}|^\gamma | F_{t''}) \leqq M_2(\gamma) E(|Y_{t_2'} - Y_{t''}|^\gamma | F_{t''}) + E\left(\left|\sum_{t'' \leqq ih < t_2'} hb(ih, X_{ih})\right|^\gamma\right)$$

avec:

$$Y_t = X_t - \sum_{ih<t} b(ih, X_{ih})h \quad \text{où} \quad b(ih, X_{ih}) = E(X_{(i+1)h} - X_{ih}|X_{ih})$$
$$= \int (y - X_{ih})\pi_{ih,X_{ih}}(dy).$$

Y_t est une F_{ih} martingale on a alors:

$$E(|Y_{t_2'} - Y_{t''}|^\gamma|F_{t''}) \leqq E((Y_{t_2'} - Y_{t''})^2|F_{t''})^{\gamma/2} \quad \text{grâce à l'inégalité de Jansen}$$
$$\leqq M_a^{\gamma/2}(t_2' - t'')^{\gamma/2}$$

grâce à la propriété de martingale de Y_t et $\int (y - X_{ih})^2 \pi_{ih,X_{ih}}(dy) \leqq M_a$.

En utilisant de plus le fait que:

$$\sup_{i,\omega} |b(ih, X_{ih})| \leqq M_b$$

on obtient:

(1.12) $$E(|X_{t_2'} - X_{t''}|^\gamma|F_{t''}) \leqq M_2(|t_2' - t''|^{\gamma/2} + |t_2' - t''|^\gamma) \leqq M_3|t_2' - t''|^{\gamma/2}$$

D'autre part:

(1.13) $$E(|X_{t''} - X_t|^\gamma|X_t - X_{t'}|^\gamma) \leqq \left(\frac{t'' - t}{h}\right)^\gamma \left(\frac{t - t'}{h}\right)^\gamma E|X_{t''} - X_{t'}|^\beta$$
$$\leqq (t'' - t)^\gamma (t - t')^\gamma \frac{h^\alpha}{h^{2\gamma}} \leqq |t'' - t'|^{\mathrm{Inf}(2\gamma,\alpha)}.$$

En combinant les majorations du type (1.11), (1.12), (1.13), on obtient:

$$E|X_{t_2} - X_t|^\gamma |X_t - X_{t_1}|^\gamma \leqq M_4|t_2 - t_1|^{\mathrm{Inf}(2\gamma,\alpha)}.$$

Dans le cas où $t_1'' < t' < t'' < t_2'$ n'est pas vérifié, on a à faire des majorations du type (1.11), (1.12), (1.13) pour obtenir le résultat.

LEMME 4 (Stroock–Varadhan [22]). *Soit*

$$\varphi \in C_b^{1,2}([0, T], R^m).$$

Notons:

$$\Delta_\varphi^h(t, x, y) = \varphi(t + h, y) - \varphi(t, x) - h\frac{\partial\varphi}{\partial t}(t, x)$$
$$-(y - x)D\varphi(t, x) - {}^t(y - x)D^2\varphi(t, x)(y - x)$$
$$\bar{\Delta}_{\varphi,\pi}^h(t, x) = \int |\Delta_\varphi^h(t, x, y)|\pi_{t,x}(dy)$$

pour π section borélienne de $\Pi^{n,C,\rho,\alpha,\beta}$.

Alors

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}(n,K,C)} E_P\left\{\sum_i \bar{\Delta}_{\varphi,\pi}^h(ih, X_{ih}(\omega))\right\} = 0.$$

Démonstration. La famille $\mathcal{P}(N, K, C)$ étant étroitement relativement compacte,

$$\forall\varepsilon > 0 \quad \exists M \sup_{P\in\mathcal{P}(N,K,C)} P\left\{\sup_{0\leqq t\leqq T} |X(t)| \geqq M\right\} \geqq \varepsilon,$$
$$\forall\varphi \in C_b^{1,2} \; \exists K(\varphi); |\Delta_\varphi^h(t, x, y)| \leqq K(|x - y|^2 + h),$$

et donc

$$\bar{\Delta}^h_{\varphi,\pi}(t,x)\leqq M_1h.$$

Il reste à montrer que:

$$\sup_{\substack{\pi \text{ section borélienne de}\\ \Pi^{n,C,\rho,\alpha,\beta}}} \bar{\Delta}^h_{\varphi,\pi}(t,x)=o(h).$$

Le théorème de Taylor donne:

$|\Delta^h_\varphi(t,x,y)|=o(h)+o(|x-y|)^2$ uniformément pour $x|x|\leqq M$, c.à.d.:

$\forall\varepsilon>0\ \exists\delta(\varepsilon,\varphi)$: $\quad h\leqq\delta|x-y|\leqq\delta\Rightarrow|\Delta^h_\varphi(t,x,y)|\leqq\varepsilon h+\varepsilon|x-y|^2$

donc:

$$\begin{aligned}\bar{\Delta}^h_{\varphi,\pi}(t,x)&=\int_{|y-x|\leqq\delta}|\Delta^h_\varphi(t,x,y)|\pi_{t,x}(dy)+\int_{|y-x|\geqq\delta}|\Delta^h_\varphi(t,x,y)|\pi_{t,x}(dy)\\&\leqq\varepsilon\int|y-x|^2\pi_{t,x}(dy)+\frac{1}{\delta^{\beta-2}}\int_{|y-x|\geqq\delta}|y-x|^\beta\pi_{t,x}(dy)+o(h)\\&\leqq\varepsilon M_ah+\rho\frac{h^\alpha}{\delta^{\beta-2}}+o(h)\end{aligned}$$

d'où le lemme. □

1.2.3. Un théorème abstrait.

Proposition 1. *Soit sur Ω un espace polonais*

(i) *une suite de mesures de probabilité $\{Q_n\}$ convergeant étroitement vers Q,*

(ii) *une suite $\{C_n\}$ de multiapplication de $\Omega\to R^n$, s.c.s., à valeur convexe dans un compact fixe, convergeant ponctuellement en décroissant, pour la distance de Hausdorff, vers une multiapplication s.c.s. C,*

(iii) *une suite $\{f_n\}$ de variables aléatoires à valeur dans $\{C_n\}$;*

Alors:

(a) *la suite de mesure $\{f_nQ_n\}$ est étroitement relativement compacte,*

(b) *il existe une sous-suite extraite de $\{f_nQ_n\}$ convergeant étroitement vers une mesure bornée, notée Q_f, absolument continue par rapport à Q;*

(c) *si l'on note $f=dQ_f/dQ$ la classe des densités de Q_f par rapport à Q, il existe un représentant borélien de f, appartenant à C.*

Remarque. On a évidemment une proposition analogue en remplaçant mesure de probabilité par mesure positive bornée.

On a un théorème analogue lorsque $C_n\to C$ sans décroître à condition que l'on ait $\sup_\omega d(C_n(\omega),D_n(\omega))\to 0$ avec $D_n\downarrow C$.

LEMME 5. *La suite $\{f_nQ_n\}$ définie par* (i), (ii) *et* (iii) *est étroitement relativement compacte.*

Démonstration. f_n est uniformément borné, notons M_c cette borne.

Q_n est étroitement relativement compacte, $\forall\varepsilon$ il existe donc [3] un compact K_ε.

$$Q_n(CK_\varepsilon)\leqq\frac{\varepsilon}{M}$$

et donc

$$|f_nQ_n|(CK_\varepsilon)\leqq|f_n|Q_n(CK_\varepsilon)\leqq\varepsilon$$

d'où le lemme. □

LEMME 6. *De la suite* $\{f_nQ_n\}$ *définie par* (i), (ii) *et* (iii) *il existe une sous-suite convergeant vers une mesure bornée notée* Q_f *absolument continue par rapport à* Q.

Démonstration. Le lemme 1 montre qu'il existe une sous-suite $\{f_{n'}Q_{n'}\}$ convergeant vers une mesure notée Q_f.

$$\sup_{n'} |f_{n'}| \leqq M \quad \text{donc:} \quad -MQ_{n'} \leqq f_{n'}Q_{n'} \leqq MQ_{n'},$$

et donc la suite de mesure $(M - f_{n'})Q_{n'}$ est positive, et converge donc, vers $MQ - Q_f \geqq 0$; de même

$$Q_f - MQ \geqq 0.$$

On obtient donc:

$$-MQ \leqq Q_f \leqq MQ \quad \text{d'où le résultat.} \quad \square$$

LEMME 7. *Soit sur* Ω *un espace métrisable séparable, une famille de mesure de probabilité* Q_n *convergeant étroitement vers* Q *et une famille* Ψ_n *de variable aléatoire s.c.s. bornée décroissante convergeant ponctuellement vers* Ψ *alors*:

$$\limsup_n \int \Psi_n \, dQ_n \leqq \int \Psi \, dQ.$$

Démonstration.

$$\int \Psi_n \, dQ_n \leqq \int \Psi_N \, dQ_n \quad si\ n \geqq N$$

grâce à la décroissance de Ψ_n.

Ψ_n étant s.c.s. le th. 5.5 De Lacherie–Meyer [7] entraîne:

$$\limsup_n \int \Psi_N \, dQ_n \leqq \int \Psi_N \, dQ.$$

et donc

$$\limsup_n \int \Psi_n \, dQ_n \leqq \limsup_n \int \Psi_N \, dQ_n \leqq \int \Psi_N \, dQ \quad \forall N.$$

Ψ_N étant décroissante en N et convergeant vers Ψ, on a (Neveu [17])

$$\lim_N \int \Psi_N \, dQ = \int \lim_N \Psi_N \, dQ = \int \Psi \, dQ$$

d'où le résultat. $\square$

Démonstration de la proposition: Le (a) résulte du lemme 1; le (b) résulte du lemme 2. Montrons le(c). $f_n \in C_n$ et donc $\forall \varphi \in C_b^0(\Omega, R^m)$

$$\operatorname*{Max}_{f \in C_n(\omega)} f \cdot \varphi(\omega) \geqq f_n \cdot \varphi(\omega) \quad \forall \omega.$$

Q_n étant positive,

$$\int \max_{f \in C_n} f \cdot \varphi \, dQ_n \geqq \int f_n \cdot \varphi \, dQ_n$$

l'application:

$$\omega \to \operatorname*{Max}_{f \in C_n(\omega)} f \cdot \varphi(\omega) \quad \text{est s.c.s.}$$

car C_n est une multiapplication s.c.s. et φ est continue, notons Ψ_n cette fonction.

C_n étant décroissante en n, il en est de même de Ψ_n, $d(C_n(\omega), C(\omega)) \to 0$

$$\Rightarrow \Psi_n(\omega) \to \Psi(\omega).$$

Le lemme 6 $\Rightarrow$

$$\limsup_n \int \operatorname{Max}_{f \in C_n} f \cdot \varphi \, dQ_n \leqq \int \operatorname{Max}_{f \in C} f \cdot \varphi \, dQ$$

comme,

$$\int \operatorname{Max}_{f \in C} f \cdot \varphi \, dQ = \operatorname{Max}_{f \in C} \int f \cdot \varphi \, dQ$$

il vient:

$$\lim_n \int f_n \cdot \varphi \, dQ_n \leqq \operatorname{Max}_{f \in C} \int f \cdot \varphi \, dQ \quad \forall n$$

et donc si l'on désigne par $\tilde{f} = dQ_f / dQ$ on obtient

$$\int \tilde{f} \cdot \varphi \, dQ \leqq \operatorname{Max}_{f \in C} \int f \cdot \varphi \, dQ \quad \forall \varphi \; C_b^0(\Omega, R^m),$$

ce qui peut se réécrire

$$0 \geqq \sup_{\varphi \in C_b^0} \{(\varphi, Q_f) - \sup_{Q_c \in CQ} (\varphi, Q_c)\},$$

avec:
$(\cdot, \cdot)$ désigne la dualité séparante [5 p. 59].

$$\mathcal{M}^b(\Omega \times [0, T]), \qquad C^b(\Omega \times [0, T]),$$

CQ désigne le convexe des mesures de $\mathcal{M}^b$ absolument continues par rapport à Q de densité c ayant un représentant $\in C$.

Le théorème 6.3.7 [15] permet alors d'affirmer que dès que CQ est fermé (CQ fermé résulte du lemme 7'):

$$\sup_{\varphi \in C_b} \{(\varphi, Q_f) - \sup_{Q_c \in CQ} (\varphi, Q_c)\} = \chi_{CQ}(Q_f).$$

et donc

$$0 \geqq \chi_{CQ}(Qf) \quad \Rightarrow \quad Q_f \in CQ \quad Q \text{ p.p.} \quad \Rightarrow \quad \text{la proposition.} \quad \square$$

LEMME 7'. *CQ est un ensemble convexe fermé dans $\mathcal{M}_+^1(\Omega)$*

Démonstration. Convexe est évident.

Fermé: Soit $\{f_n Q\}$ une suite $\in CQ$. f_n reste dans un borné de $L^\infty(\Omega, Q)$ donc converge faiblement $\sigma(L^\infty, L^1)$ vers f, donc il existe un compromis convexe $\sum a_n f_n$ convergeant Q p.p. or $\sum a_n f_n(\omega) \in C(\omega) \Rightarrow f(\omega) \in C(\omega)$ puisque $C(\omega)$ est fermé c.q.f.d.

1.2.4. Le théorème d'existence.

THÉORÈME 1.

$$\mathcal{P}(K, C) \supset \bigcap_N \overline{\bigcup_{n \geqq N} \mathcal{P}(n, K, C, \rho, \alpha, \beta)}$$

(1.14)

$$\exists \rho : \bigcap_N \overline{\bigcup_{n \geqq N} \mathcal{P}(n, K, C, \rho, 2, 3)} \neq \varnothing.$$

Démonstration. (a) $\exists \rho : \bigcap_N \overline{\bigcup_{n \geqq N} \mathcal{P}(n, K, C, \rho, 2, 3)} \neq \varnothing$: En effet, $\mathcal{P}(n, K, C, \rho, 2, 3) \neq \varnothing$ (prendre par example des accroissements gaussiens.).

Donc $\exists\{P_n\}P_n \in \mathscr{P}(n, K, C, \rho, 2, 3)$, la suite est étroitement relativement compacte. Il existe donc une sousesuite $P_{n'} \to P \in \bigcap_N \bigcup_{n \geqq N} \mathscr{P}(n, K, C, \rho, 2, 3)$ et donc $\bigcap_N \bigcup_{n \geqq N} \mathscr{P}(n, K, C, \rho, 2, 3) \neq \varnothing$.

(b) $\mathscr{P}(K, C) \supset \bigcap_N \bigcup_{n \geqq N} \mathscr{P}(n, K, C, \rho, \alpha, \beta)$: Soit $P \in \bigcap_N \bigcup_{n \geqq N} \mathscr{P}(n, K, C, \rho, \alpha, \beta)$. $\exists$ une suite $\{P_{n_i}\}P_{n_i} \in \mathscr{P}(n_i, K, C, \rho, \alpha, \beta), P_{n_i} \xrightarrow[i \to \infty]{} P$. Montrons que $P \in \mathscr{P}(K, C)$.

En effet, considérons la suite d'espace mesuré $\{\Omega \times [0, \mathrm{T}], \mathrm{F} \otimes \mathscr{B}; Q_{n_i}\}$ avec

$$Q_{n_i} = P_{n_i}(d\omega) \otimes \frac{T}{n_i} \sum_{j=0}^{n_i-1} \delta_{jh}(dt). \tag{1.15}$$

Q_{n_i} converge étroitement vers $P(d\omega) \otimes dt$.

$$P_{n_i} \in \mathscr{P}(n_i, K, C, \rho, \alpha, \beta) \quad \exists\{\pi^{n_i}\}\{y^{n_i}\} : \{y_{n_i} \in K\}\{\pi^{n_i} \in \Pi^{n_i, C, \rho, \alpha, \beta}\} \tag{1.16}$$

tels que $P_{n_i} = P^{n_i}_{\pi_{n_i}, y_{n_i}}$.

(b1) *Montrons alors* $\exists y \in K, P(x_0 = y) = 1$. K étant compact il existe $\{n'_i\}$: $y_{n'_i} \to y \in K$. Soit $\varphi \in C_b^0\,(R^m)$

$$E^{P n'_i}\varphi(x_0) \xrightarrow[i \to \infty]{} E^P\varphi(x_0) \quad \text{car } \omega \to \varphi \circ x_0(\omega) \text{ est continue bornée} \tag{1.17}$$

or

$$E^{P_{n'_i}}\varphi(x_0) = \varphi(y_{n'_i}) \xrightarrow[i \to \infty]{} \varphi(y). \tag{1.18}$$

(1.17) et (1.18) entraîne alors le résultat

(b2) *Montrons qu'il existe* $\hat{f}(s, \omega)$ *prévisible vérifiant*

$$\hat{f}(s, \omega) \in C(s, \omega) \tag{1.19}$$

$$\varphi(X_t) - \int_0^t L\hat{f}\varphi(s, \omega)\, ds \text{ est une } (P, F_t) \text{ martingale.} \tag{1.20}$$

$P^{n_i} = P^{n_i}_{\pi_{n_i}, y_{n_i}}$; il exist donc

$$c_{n_i}(s, x) \in C(s, x) \quad \text{avec } c_{n_i}(s, x) = (b_{n_i}(s, x), a_{n_i}(s, x))$$

$$\int (y - x)\pi^{n_i}_{s, x}(dy) = b_{n_i}(s, x)h,$$

$$\int (y - x) \otimes (y - x)\pi^{n_i}_{s, x}(dy) = a_{n_i}(s, x)h.$$

Considérons alors la suite de mesure

$$c_{n_i}(s, X_s(\omega))\, dQ_{n_i}$$

Grâce à la proposition 1 $\exists\{n'_i\}$

$$c_{n'_i}(s, X_s(\omega))\, dQ_{n'_i} \to c(s, \omega)\, dQ \tag{1.21}$$

avec

$$c(s, \omega) \in C(s, X_s(\omega)).$$

Notons $\hat{c}(s, \omega)$ la projection sur la tribu des prévisibles de $c(s, \omega)$. L'inégalité de Jensen montre que: $\hat{c}(s, \omega) \in C(s, X_s(\omega))$.

Montrons que P vérifie (1.20) pour $\hat{f}(s, \omega) = \hat{c}(s, \omega)$.

Notons pour $\varphi \in C_b^{1,2}([0, T] \times R^m)$

$$Z_\varphi^n(u, t, \omega) = \int_u^t h \sum_{j=0}^{n-1} \delta_{jh}(ds) \int \{\varphi((j+1)h, y) - \varphi(jh, X_{jh})\} \pi_{jh, X_{jh}}^n(dy)$$

Φ continue F_u mesurable

$k_n h \to t,$

$l_n h \to u,$

$$0 = E^{P_n}\{[\varphi(k_n h, X_{k_n h}) - \varphi(l_n h, X_{l_n h}) - Z(l_n h, k_n h, \omega)]\Phi\}$$

$$= E^{P_n}\left\{\left[\varphi(k_n h, X_{k_n h}) - \varphi(l_n h, X_{l_n h}) - \int_{l_n h}^{k_n h} L_{c_n}\varphi(s, \omega) h \sum_{j=0}^{n-1} \delta_{jh}(s)\right]\Phi\right\}$$

$$+ E^{P_n}\left\{\left[\int_{l_n h}^{k_n h} L_{c_n}\varphi(s, \omega) h \sum_{j=0}^{n-1} \delta_{jh}(s) - Z_\varphi^n(l_n h, k_n h, \omega)\right]\Phi\right\}.$$

Grâce au lemme 4, on a

$$\lim_{n\to\infty} \sup_{P_n} E^{P_n}\left|\left[\int_{l_n h}^{k_n h} L_{c_n}\varphi(s, \omega) h \sum_{j=0}^{n-1} \delta_{jh}(s) - Z_\varphi^n(l_n h, k_n h, \omega)\right]\Phi\right| = 0.$$

Grâce à la définition de $\hat{c}$, $k_n h \to t$, $l_n h \to u$ on a:

$$\lim_{n\to\infty} E^{P_n}\left\{\Phi\left[\varphi(k_n h, X_{k_n h}) - \varphi(l_n h, X_{l_n h}) - \int_{l_n h}^{k_n h} L_{c_n}\varphi(s, \omega) h \sum_{j=0}^{n-1} \delta_{jh}(s)\right]\right\}$$

$$= E^P\left\{\Phi\left[\varphi(t, X_t) - \varphi(u, X_u) - \int_u^t L_{\hat{c}}\varphi(s, \omega)\, ds\right]\right\} = 0 \qquad \text{c.q.f.d.}$$

1.2.5. Existence d'une solution faible à l'équation de Fokker-Planck. Soit C la multiapplication défini en 1.1 *on dira que μ_t est solution faible de l'équation de Fokker Planck s'il existe $\tilde{c}(s, x)$ section borélienne de $C(s, x)$*:

(1) $\mu_t \in \mathcal{M}_+^1(R^m)$;

(2) $\mu_0 = \delta_{z_0}$;

(3) $\forall \varphi \in C_b^{1,2}(R^m)$ on a

$$\int \varphi(T, x)\mu_T(dx) - \varphi(0, x_0) - \int_0^t \int_{R^m} L_{\tilde{c}}\varphi(s, x)\mu_s(dx)\, ds = 0.$$

On notera $\mu(K, C)$ l'ensemble des solutions faibles de l'équation de Fokker–Planck.

THÉORÈME 2.

$$\mu(K, C) \neq 0.$$

Démonstration. $P \in \mathcal{P}(K, C)$ correspond μ_t par l'application $\omega \to X_t(\omega)$: μ_t vérifie (3); en effet, il suffit de prendre $\tilde{c}(s, x) =$ projection de $\hat{c}(s, \omega)$ défini en (1.21) sur la tribu du présent $\{\sigma(X_s)\}$ on a alors

$$O = E^P\left\{\varphi(T, X_T) - \varphi(0, X_0) - \int_0^T L_{\hat{c}}\varphi(s, \omega)\, ds\right\}$$

$$= \int \varphi(T, X)\mu_T(dx) - \varphi(0, X_0) - \int_0^T \int_{R^m} L_{\tilde{c}}\varphi(s, x)\mu_s(dx)\, ds.$$

1.3. Propriétés de $\mathcal{P}(K, C)$.

THÉORÈME 3. *$\mathcal{P}(y, C)$ est un compact convexe non vide de $\mathcal{M}_+^1(\Omega)$.*

On va démontrer ce théorème grâce à trois lemmes.

LEMME 8. *$\mathcal{P}(K, C)$ est étroitement relativement compact.*

Démonstration. On utilise le critère suivant P. Billingsley th. 12.3 [3]. Il existe une fonction continue non décroissante F et deux nombres γ et α, $\gamma \geqq 0$, $\alpha > 1$, tels que

$$E_P|X(t)-X(s)|^\gamma \leqq |F(t)-F(s)|^\alpha \quad \forall P \in \mathcal{P}(K, C).$$

On l'applique avec $\gamma = 3$, $\alpha = \frac{3}{2}$, $F(t) = t$.

Soit en effet $P \in \mathcal{P}(K, C)$, $\exists c(s, \omega) = (b(s, \omega), a(s, \omega)) \in C(s, X_s(\omega))$ prévisible tel que: $M_t = X(t) - \int_0^t b(s, \omega)\, ds$ soit une martingale de processus croissant $\int_0^t a(s, \omega)\, ds$ et donc $\exists \delta(M_b)$:

$$E^P|X(t)-X(s)|^3 \leqq \delta(|t-s|^3 + E|M_t - M_s|^3).$$

En utilisant la proposition 19 de [18] on a:

$$E|M_t - M_s|^3 \leqq E\left(\int_s^t a(u, \omega)\, du\right)^{3/2} \leqq M_a|t-s|^{3/2}. \qquad \text{c.q.f.d.}$$

LEMME 9. *$\mathcal{P}(K, C)$ est fermé.*

Démonstration. Soit P^n une suite de mesures de probabilité appartenant à $\mathcal{P}(K, C)$ convergeant vers P dans $\mathcal{M}_+^1(\Omega)$ montrons que $P \in \mathcal{P}(K, C)$.

On considère la suite de mesures $Q_n = P^n \otimes dt$ sur $(\Omega \times [0, T]$, $\mathcal{B} \otimes F)$ elle est étroitement convergente vers $Q = P \otimes dt$.

$P_n \in \mathcal{P}(K, C)$ il existe donc $c_n(s, \omega) \in C(s, X_s(\omega))$ prévisibles telles que $\forall \varphi \in C_b^{1,2}(0, T \times R^m)$. $\forall \Phi F_u$ mesurable continue

$$E^{P_n}\left\{\left[\varphi(t, X_t) - \varphi(u, x_u) - \int_u^t L_{c_n}\varphi(s, \omega)\, ds\right]\Phi\right\} = 0$$

de la suite $c_n Q_n$ on peut extraire une sous-suite $c_{n'}Q_{n'}$ convergeant vers cQ, d'après la proposition 1, la projection prévisible $\hat{c}$ vérifie

$$E^P\left\{\left[\varphi(t, X_t) - \varphi(u, X_u) - \int_u^t L_{\hat{c}}\varphi(s, \omega)\, ds\right]\Phi\right\} = 0,$$

$$\hat{c}(s, \omega) \in C(s, X_s(\omega)). \qquad \text{c.q.f.d.}$$

LEMME 10. *$\mathcal{P}(y, C)$ est convexe.*

Démonstration. Soit P_1 et $P_2 \in \mathcal{P}(y, C)$. Il existe alors $c_1(s, \omega)$ et $c_2(s, \omega) \in C(s, \omega)$ prévisibles tels que $\forall \varphi \in C_b^{1,2}([0, T] \times R^m)$

$$(1.22) \qquad \varphi(t, X_t) - \varphi(u, X_u) - \int_u^t L_{c_1}\varphi(s, \omega)\, ds \quad \text{est une } (P_1, F_t) \text{ martingale,}$$

$$(1.23) \qquad \varphi(t, X_t) - \varphi(u, X_u) - \int_u^t L_{c_2}\varphi(s, \omega)\, ds \quad \text{est une } (P_2, F_t) \text{ martingale;}$$

alors notons:

$$c_\lambda(s, \omega) = \frac{\lambda c_1(s, \omega)\, dP_1 + (1-\lambda)c_2(s, \omega)\, dP_2}{\lambda\, dP_1 + (1-\lambda)\, dP_2},$$

$\hat{c}_\lambda(s, \omega)$ la projection prévisible de $c_\lambda(s, \omega)$

alors:

$$(1.24)\qquad \varphi(t, X_t)-\varphi(u, X_u)-\int_u^t L_{\hat{c}_\lambda}\varphi(s, \omega)\, ds \quad \text{est une } (\lambda P_1+(1-\lambda)P_2, F_t).$$

En effet, soit Φ continue F_u mesurable on a:

$$E_{\lambda P_1+(1-\lambda)P_2}\left\{\left[\varphi(t, X_t)-\varphi(u, X_u)-\int_u^t L_{\hat{c}_\lambda}\varphi(s, \omega)\, ds\right]\Phi\right\}$$

$$=E_{\lambda P_1+(1-\lambda)P_2}\left\{\left[\varphi(t, X_t)-\varphi(u, X_u)-\int_u^t L_{c_\lambda}\varphi(s, \omega)\, ds\right]\Phi\right\}$$

$$=\lambda E_{P_1}\left\{\left[\varphi(t, X_t)-\varphi(u, X_u)-\int_u^t L_{c_1}\varphi(s, \omega)\, ds\right]\Phi\right\}$$

$$+(1-\lambda)E_{P_2}\left\{\left[\varphi(t, X_t)-\varphi(u, X_u)-\int_u^t L_{c_2}\varphi(s, \omega)\, ds\right]\Phi\right\}=0.$$

Montrons que $c_\lambda(s, \omega)\in C(s, X_s(\omega))$. Notons

$$Q_1=\frac{1}{T}P_1\otimes dt, \qquad Q_2=\frac{1}{T}P_2\otimes dt;$$

$\forall f\in C_b^0(\Omega\times[0, T], R^m\times R^{m\times m})$

$$E^{\lambda Q_1+(1-\lambda)Q_2}c_\lambda\cdot\varphi=\lambda E^{Q_1}c_1\cdot\varphi+(1-\lambda)E^{Q_2}c_2\cdot\varphi$$

$$\leqq\lambda E^{Q_1}\operatorname*{Max}_{c\in C} c\,.\,\varphi+(1-\lambda)E^{Q_2}\operatorname*{Max}_{c\in C} c\,.\,\varphi$$

$$\leqq E^{\lambda Q_1+(1-\lambda)Q_2}\operatorname*{Max}_{c\in C} c\,.\,\varphi=\operatorname*{Max}_{c\in C} E^{\lambda Q_1+(1-\lambda)Q_2}c\,.\,\varphi.$$

En raisonnant comme dans la démonstration de la proposition 1, on en déduit:

$$c_\lambda(s, \omega)\in C(s, X_s(\omega)) \qquad \lambda Q_1+(1-\lambda)Q_2 \quad \text{p.p.}$$

L'inégalité de Jensen montre alors que

$$\hat{c}_\lambda(s, \omega)\in C(s, X_s(\omega))$$

d'où le lemme. □

Soit $P\in\mathcal{P}(y, C)$ donnons quelques propriétés de la loi de probabilité du vecteur:

$$(X_0, X_h, X_{2h}, \cdots, X_{(n-1)h}, X_{nh}) \quad \text{avec } h=\frac{T}{n}.$$

Cette loi de probabilité peut s'écrire:

$$\delta_y(dx_0)\overset{1}{\pi}_{x_0}(dx_1)\overset{2}{\pi}_{x_0,x_1}(dx_2)\overset{3}{\pi}_{x_0,x_1,x_2}(dx_3)\cdots\overset{n}{\pi}_{x_0,x_1,\cdots x_{n-1}}(dx_n).$$

LEMME 11. *Il existe une constante* $\delta(M_a, M_b, T)$ *telle que*:

$$\sup_i\int\overset{i}{\pi}_{x_0,x_1,\cdots x_i}(dy)|y-x_i|^3\leqq\delta h^{3/2}.$$

Démonstration. La même que celle du lemme 8, utilisant le fait que $E^{X_0,\cdots X_{ih}}(\varphi)=E^{X_0,\cdots X_{ih}}E^{F_{ih}}(\varphi)$ avec φ variable aléatoire sur Ω.

Notons

$C_{h,\varepsilon}(t, x) = \bar{\mathscr{C}}_0 \bigcup\limits_{\substack{t \leqq s \leqq t+h \\ |x-z| \leqq \varepsilon}} C(s, z)$ où $\bar{\mathscr{C}}_0$ désigne l'opération de fermeture convexe,

$K^{\varepsilon}(t, \omega) = \{\omega : \sup\limits_{t \leqq s \leqq t+h} |X_s(\omega) - X_t(\omega)| \leqq \varepsilon\}$,

$\tau^{\varepsilon}_{x,t} = \inf \{s \geqq t | X_s - x| > \varepsilon\}$,

$\tau^{\varepsilon,h}_{x,t} = \tau^{\varepsilon}_{x,t} \wedge (t+h)$

$\alpha : \Omega \to \Omega$

$\alpha : \omega \to \qquad \alpha(\omega), \qquad \alpha\omega(s) = \begin{cases} \omega(s), & s \leqq \tau^{\varepsilon,h}_{x,t}, \\ \omega(\tau^{\varepsilon,h}_{x,t}), & s \geqq \tau^{\varepsilon,h}_{x,t}. \end{cases}$

LEMME 12.

$$\left(\int (y-x)\pi^i(dy), \int (y-x)^{\otimes 2}\pi^i(dy)\right) \in [C_{h,\varepsilon}(t, x)h + \mathscr{V}_0(\beta)] \cap G$$

avec

$$t = (i-1)h, \qquad \beta = kh/\varepsilon^2,$$

où k est une constante ne dépendant que de C; $\mathscr{V}_0(\delta)$ désigne un voisinage de 0 de diamètre δ dans $R^m \times R^{(m\times(m+1))/2}$.

Si l'on note $\tilde{F}$ la tribu engendrée par $(X_0, X_h, \cdots, X_{(i-1)}h)$ $\int (y-x)\pi^i(dy) = E^{\tilde{F}}E^{Ft_1}[\int_{t_1}^{t_2} b(s, \alpha\omega)\, ds + \int_{t_1}^{t_2} (b(s, \omega) - b(s, \alpha\omega))\, ds]$.

L'utilisation de la formule d'Ito et la définition de π^i montre: $\int (y-x)^{\otimes 2}\pi^i(dy) = E^{\tilde{F}}E^{F_{t_1}} \int_{t_1}^{t_2} a(s, \alpha\omega)\, ds + \int_{t_1}^{t_2} (a(s, \omega) - a(s, \alpha\omega))\, ds + \int_{t_1}^{t_2} (X_s - X_{t_1}) \otimes b(s, \omega)\, ds$

avec:

CK_ε l'ensemble complémentaire de K_ε,

$\alpha\omega$ construit à partir du temps d'arrêt $\tau^{\varepsilon,h}_{x,t_1}$,

$$t_1 = (i-1)h, \qquad t_2 = ih,$$

or

$$\left(\int_{t_1}^{t_2} b(s, \alpha\omega)\, ds, \int_{t_1}^{t_2} a(s, \alpha\omega)\, ds\right) \in C_{h,\varepsilon}(t, x)h,$$

$$E^{F_{t_1}} \int_{t_1}^{t_2} |b(s, \omega) - b(s, \alpha\omega)| ds \leqq hM_b P^{F_{t_1}} (CK_\varepsilon),$$

$$E^{F_{t_1}} \int_{t_1}^{t_2} |(X_s - X_{t_1}) \otimes b(s, \omega)| ds \leqq hM_b E^{F_{t_1}} \sup_{t_1 \leqq s \leqq t_2} |X_s - X_{t_1}|,$$

$$E^{F_{t_1}} \int_{t_1}^{t_2} |a(s, \omega) - a(s, \alpha\omega)|\, ds \leqq hM_a P^{F_{t_1}} (CK_\varepsilon).$$

Il reste à montrer que $\exists k$:

$$P^{\tilde{F}}(CK_\varepsilon) \leqq \frac{kh}{\varepsilon^2}$$

$$E^{\tilde{F}} \sup_{t_1 \leqq s \leqq t_2} |X_s - X_{t_1}| \leqq \frac{kh}{\varepsilon^2}$$

or

$$\exists k_1 \quad \text{et} \quad k_2:$$

$$P(CK_\varepsilon) \leqq \sup_t P\left\{\omega: \sup_{t \leqq s \leqq t+h} |X_s - X_t| \geqq \varepsilon\right\}$$

$$\leqq \frac{1}{\varepsilon^2} E \sup_{t \leqq s \leqq t+h} |X_s - X_t|^2$$

$$\leqq \frac{k_1 h^2}{\varepsilon^2} + \frac{k_2 h}{\varepsilon^2}$$

le deuxième terme étant obtenu en utilisant le théorème de Doob $E\{\sup_{s \leqq t} M_s^2\} \leqq EM_t^2$ si M_t est une martingale de carré intégrable.

De même,

$$E^{F_{t_1}} \sup_{t_1 \leqq s \leqq t_2} |X_s - X_{t_1}|^2 \leqq k_1 h^2 + k_2 h$$

d'où le lemme. □

2. Contrôle optimal de problèmes de martingale.

2.1. Définition du problème. Soit Φ: $R^m \rightarrow R$ s.c.i., bornée, une fonction coût.

On se pose le problème de contrôle stochastique suivant:

$$\text{(2.1)} \qquad \operatorname*{Min}_{P \in \mathcal{P}(y,\, C)} E^P \Phi(X_T).$$

2.2. Existence d'une solution.

THÉORÈME 4. *Le problème de contrôle* (2.1) *admet une solution.*

Démonstration. L'application

$$\begin{array}{ll} \Omega \rightarrow R^m & \text{est continue} \\ \omega \quad X_T(\omega) & \end{array}$$

Φ étant s.c.i.

$$\Phi \circ X_T : \Omega \rightarrow R \quad \text{est donc s.c.i.}$$

L'application

$$\begin{array}{l} \mathcal{M}_+^1(\Omega) \rightarrow R \\ P \qquad E^P \Phi \circ X_T \end{array}$$

est donc s.c.i. grâce au théorème 5.5 [7].

Grâce au théorème 3, on sait que $\mathcal{P}(y, C)$ est un compact, convexe, non vide de $\mathcal{M}_+^1(\Omega)$. On obtient donc le théorème.

Remarque. Soit $\tau(\omega)$ le temps de sortie d'un ouvert, alors l'application:

$$\omega \rightarrow \tau(\omega) \quad \text{est s.c.i.}$$

Considérons le problème de contrôle stochastique

$$\operatorname*{Min}_{P \in \mathcal{P}(y,\, C)} E^P X_m(\tau(\omega), \omega);$$

les problèmes de contrôle avec coût seulement intégral peuvent toujours s'écrire de cette façon, de plus, dans ce dernier cas si l'intégrande est positif. $t \rightarrow X_m(t, \omega)$ est P p.s.

croissante $\forall P \in \mathcal{P}(y, C)$ et l'application

$$\omega \to E^P X_m(\tau(\omega), \omega) \quad \text{sera alors s.c.i.}$$

On aura encore existence du contrôle optimal dans ce cas.

Pour des problèmes arrêtés plus généraux, coûts sur l'état final il faudra pour pouvoir être assuré de l'existence, montrer que le temps de sortie de l'ouvert et du fermé sont P p.s. les mêmes $\forall P \in \mathcal{P}(y, C)$. Ce qui est vrai au moins dans le cas non dégénéré [22], et dans certains cas dégénérés également [22].

2.3. Caractérisation d'un contrôle optimal (discrétisation en temps).

2.3.1. Définition d'un problème de contrôle optimal approché. On considère la multiapplication

$$(s, x) \to C_{h,\varepsilon}(s, x) \quad \text{avec} \quad C_{h,\varepsilon}(s, x) = \bar{\mathscr{C}}_0\left\{ \bigcup_{\substack{\varepsilon \leqq t \leqq s+h \\ |y-x| \leqq \varepsilon}} C(t, y)\right\}.$$

LEMME 14. *$C_{h,\varepsilon}(s, x)$ est une multiapplication s.c.s. convergeant (au sens de la distance de Hausdorff) ponctuellement en décroissant vers la multiapplication $C(s, x)$ lorsque $h \downarrow 0$ $\varepsilon \downarrow 0$.*

Démonstration. $C_{h,\varepsilon}(s, x)$ est s.c.s. Il suffit de montrer que $C_{h,\varepsilon}(s, x)$ est de graphe fermé c.a.d. soit

$$(s_n, x_n, c_n) \xrightarrow[n \to \infty]{} (s, x, c) \quad c_n \in C_{h,\varepsilon}(s_n, x_n)$$

Montrons que $c \in C_{h,\varepsilon}(s, x)$

$$c_n \in C_{h,\varepsilon}(s_n, x_n) \quad \Rightarrow \quad \text{il existe:}$$

$$\begin{array}{lll} & d_{n,i}, & i = 1, \cdots, \dfrac{m(m+1)}{2} + m + 1; \\ & \lambda_{n,i}, & \lambda_{n,i} \geqq 0, \quad \sum_i \lambda_{n,i} = 1,; \\ (2.2) & s_{n,i}, & 0 \leqq s_{n,i} - s_n \leqq h;; \\ & x_{n,i}, & |x_{n,i} - x_n| \leqq \varepsilon, \\ & d_{n,i} \in C(s_{n,i}, x_{n,i}), & C_n = \sum_i \lambda_{n,i} d_{n,i}. \end{array}$$

$(s_{n,i}, x_{n,i}, \lambda_{n,i}, d_{n,i})$ appartenant à un compact, il existe une sous-suite:

$$\begin{array}{lll} & s_{n',i} \xrightarrow[nn' \to \infty]{} s_i, & 0 \leqq s_i - s \leqq h, \\ & x_{n',i} \xrightarrow[n' \to \infty]{} x_i & |x_i - x| \leqq \varepsilon, \\ (2.3) & & \\ & \lambda_{n',i} \xrightarrow[n' \to \infty]{} \lambda_i & \lambda_i \geqq 0, \quad \sum \lambda_i = 1, \\ & d_{n',i} \xrightarrow[n' \to \infty]{} d_i & d_i \in C(s_i, x_i) \text{ car } C \text{ est s.c.s.} \end{array}$$

et donc

$$\sum \lambda_i d_i \in \bar{\mathscr{C}}_0 \Big\{ \bigcup_{\substack{|y-x| \leqq \varepsilon \\ 0 \leqq t-s \leqq h}} C(t, y) \Big\}$$

d 'où le résultat.

(2.4) $$\lim_{\substack{h\downarrow 0 \\ \varepsilon \downarrow 0}} d(C_{h,\varepsilon}(s, x), C(s, x)) = 0.$$

En effet, $C_{h,\varepsilon}(s, x)$ étant compact

$$\overline{\bigcap_{h,\varepsilon} C_{h,\varepsilon}(s, x)} = \{c \mid \exists c_n \in C_{h_n,\varepsilon_n}(s, x),\ c_n \xrightarrow[n\to\infty]{} c,\ h_n \downarrow 0,\ \varepsilon_n \downarrow 0\}$$

$$\subset \Big\{ c \mid \exists (s_{n,i}, x_{n,i}, \lambda_{n,i}, d_{n,i})$$

$$s_{n,i} \xrightarrow[n\to\infty]{} s_i,\ x_{n,i} \xrightarrow[n\to\infty]{} x_i,\ \lambda_{n,i} \to \lambda_i \ \sum_i \lambda_{n,i} = 1,\ \lambda_{n,i} \geqq 0$$

$$d_{n,i} \in C(s_{n,i}, x_{n,i}),\ c_n = \sum_i \lambda_{n,i} d_{n,i} \Big\}$$

$$\subset \Big\{ c \mid c \in C(s, x) \Big\} \text{ car } c_n \to c = \sum \lambda_i d_i,$$

$$\lambda_i = \lim_{n'} \lambda_{n',i},\ d_{n',i} \xrightarrow[n'\to\infty]{} d_i,\ \lambda_i \geqq 0,\ \sum \lambda_i = 1,$$

$d_i \in C(s, x)$ car C est de graphe fermé, et $c \in C(s, x)$ car C est à valeur convexe, et donc

(2.5) $$\overline{\bigcap_{h,\varepsilon} C_{h,\varepsilon}(s, x)} = C(s, x)$$

l'inclusion dans l'autre sens étant évidente.

Supposons que (2.4) soit fausse, la décroissance de $C_{h,\varepsilon}$ lorsque $h \downarrow 0$ et $\varepsilon \downarrow 0 \Rightarrow$ $\exists h_n,\ \varepsilon_n,\ y_n \in C_{h_n,\varepsilon_n}(s, x),\ d(y_n, C(s, x)) > \varepsilon$;

$$y_n \in \text{Compact},\ \exists n',\ y_{n'} \to y \text{ et } d(y, C(s, x)) \geqq \varepsilon$$

or

$$y \in \overline{\bigcap_{h_n,\varepsilon_n} C_{h_n,\varepsilon_n}(s, x)} = C(s, x)$$

d'où la contradiction. □

Considérons la multiapplication $C_n = [C_{1/n,(1/n)^\gamma} + \mathscr{V}^*(0, kn^{2\gamma}/n)] \cap G$ avec $0 < \gamma < \frac{1}{2}$,[3] k défini au lemme 12. On a $C_n(s, x) \searrow C(s, x)$ pour la topologie de Hausdorff grâce au lemme 14. Considérons la multiapplication

$$\prod^n : (s, x) \to \prod_{(s,x)}^{n, C_n, \delta, 3/2, 3} \text{ (que nous noterons } \prod^n)$$

avec δ défini au lemme 11.

[3] $\mathscr{V}^*(0, \rho)$ désigne la boule fermé de centre 0 et de rayon ρ dans $R^{m\times(m+1)/2}$. R^m.

C_n étant s.c.s. Π^n est s.c.s. (proposition 0). Considérons le problème de Programmation dynamique

$$V_n(T, x) = \Phi(x) \quad \text{avec } \Phi(T, \cdot) \text{ s.c.i.},$$

$$\begin{aligned} V_n((n-1)h, x) &= \min_{\pi \in \Pi^n((n-1)h, x)} \int V_n(T, y)\pi(dy), \\ V_n(ih, x) &= \min_{\pi \in \Pi^n(ih, x)} \int (V_n(i+1)h, y)\pi(dy), \\ V_n(0, x) &= \min_{\pi \in \Pi^n(0, x)} \int V_n(h, y)\pi(dy). \end{aligned} \tag{2.6}$$

THÉORÈME 5. *Le problème* (2.6) *admet une solution de plus* $V(ih, x)$ *est s.c.i.* $\forall i$.

Démonstration. On le démontre par récurrence. L'application:

$$\pi \to \int V_n(T, y)\pi(dy)$$

est s.c.i. car Φ est s.c.i. $\Pi^n((n-1)h, x)$ est à valeur compacte car elle est s.c.s.

Et donc $\min_{\pi \in \Pi^n((n-1)h, x)} \int V_n(T, y)\pi(dy)$ existe et $V(T-h, x)$ est s.c.i. grâce au théorème du maximum [2].

Il existe $\pi^*(ih, x)$, borélienne en x, réalisant le minimum. A π^* associons $P^n_{\pi^*}$ par la méthode exposée en § 1.2.2. $\{P^n_{\pi^*}\}$ est étroitement relativement compacte grâce au lemme 3, car les C_n sont décroissants, et C_1 vérifie les hypothèses du lemme 3.

Par la méthode exposée dans le théorème d'existence, en remplaçant partout C par C_n, on obtient que toute sous-suite convergente de $P^{n'}_{\pi^*} \xrightarrow[n' \to \infty]{} P \in \mathcal{P}(y, C)$

THÉORÈME 6. (*Caractérisation d'un contrôle optimal*). *Si* $\{P^n_{\pi^*}\}$ *désigne la suite de mesure sur* Ω *définie, par interpolation linéaire, sur la chaîne de Markov, solution du problème de contrôle discrétisé, de probabilité de transition* π^*_n; $\{P^n_{\pi^*}\}$ *est étroitement relativement compacte, et toute sous-suite convergente converge vers un élément de* $P^* \in \mathcal{P}(y, C)$ *solution du problème de contrôle optimal* (2.1).

Démonstration. On a démontré l'admissibilité de P^* démontrons son optimalité. Pour cela, supposons que $\tilde{P}$ [resp $\tilde{V}$] désigne un contrôle optimal [resp le coût optimal] du problème (2.1) et considérons la loi du vecteur $(X_0, X_h, X_{2h}, \ldots, X_{nh})$. Elle peut s'écrire:

$$\delta_y(dx_0)\pi^0_{x_0}(dx_1)\pi^1_{x_0,x_1}(dx_2) \cdots \pi^{n-1}_{x_0,x_1\cdots,x_{n-1}}(dx_n)$$

Grâce au lemmes 11, 12 et la définition de C_n, on vérifie que $\pi^i_{x_0,\cdots x_i} \in \Pi^n(ih, x_i)$ et donc

$$V_n(0, x_0) \leqq \tilde{V} \quad \forall n$$

or,

$$V_n(0, x_0) = E_{P^n_{\pi^*}}\Phi(X_T), \ \Phi \text{ étant s.c.i.}$$

$$\tilde{V} \geqq \liminf_{n \to \infty} V_n \geqq E_{P^*}\Phi(X_T) \text{ c.q.f.d.}$$

2.4. Caractérisation d'un contrôle optimal (discrétisation en temps et en espace).

2.4.1. Définition d'un problème de contrôle optimal de chaîne de Markov à état discret. Soit la multiapplication $C_{h,\varepsilon,k}(s, x)$ dont le graphe est défini par:

$$\overline{\left\{(s, x, c) \middle| c \in \mathscr{C}_0\left\{ \bigcup_{\substack{-\varepsilon+ik \leqq y \leqq (i+1)k+\varepsilon \\ s \leqq t \leqq s+h}} C(s, y), x \in [ik, (i+1)k] \right\}\right.}.$$

$C_{h,\varepsilon,k}$ est donc une multiapplication s.c.s., constante par morceaux en x, à valeur convexe.

Remarque.

$$C_{h,\varepsilon,0}(s,x)=\bar{\mathscr{C}}_0\{C_{h,\varepsilon}(s,x)\}.$$

LEMME 15. *$C_{h,\varepsilon,k}(s,x)$ est une multiapplication s.c.s. à valeur convexe convergeant ponctuellement, en décroissant, au sens de la distance de Hausdorff, vers la multiapplication $C(s,x)$ lorsque h, ε, k tendent vers 0 en décroissant.*

Démonstration.

$$\lim_{\substack{h\downarrow 0\\ \varepsilon\downarrow 0\\ k\downarrow 0}} d(C_{h,\varepsilon,k}(s,x),C(s,x))=0. \tag{2.7}$$

En effet, $C_{h,\varepsilon,k}(s,x)$ *étant compact*:

$$\overline{\bigcap_{h,\varepsilon,k} C_{h,\varepsilon,k}(s,x)}\subset\Big\{c\,|\,\exists(c_n,s_n,x_n)s_n\to s,\ x_n\to x,\ c_n\to c$$

$$c_n\in\mathscr{C}_0\Big\{\bigcup_{\substack{s_n\leqq t\leqq s_n+h_n\\ x_n-2k_n-\varepsilon_n\leqq y\leqq x_n+2k_n+\varepsilon_n}} C(t,y)\Big\},\ h_n,\varepsilon_n,k_n\downarrow 0\Big\};$$

c.a.d.

$$c_n=\sum_i d_{n,i}\lambda_{n,i},\quad \lambda_{n,i}\geqq 0,\quad \sum\lambda_{n,i}=1,\quad d_{n,i}\in C(t_{n,i},y_{n,i}),\ i=1,\cdots,m+m(m+1)/2+1,$$

$$|x-y_{n,i}|\leqq 2k_n+\varepsilon_n+|x_n-x|,\qquad |s-t_{n,i}|\leqq h_n+|s_n-s|,$$

et donc, grâce à la s.c.s. de $C(s,x)$

$$d_{n,i}\xrightarrow[n\to\infty]{} d_i\in C(s,x)$$

$$\lambda_{n,i}\xrightarrow[n\to\infty]{}\lambda_i,\ \lambda_i\geqq 0,\qquad \sum\lambda_i=1$$

$$c_n\to c=\sum\lambda_i c_i$$

et en utilisant la convexité de C, $c\in C(s,x)$; et donc

$$\overline{\bigcap_{h,\varepsilon,k} C_{h,\varepsilon,k}(s,x)}=C(s,x). \tag{2.8}$$

Supposons que (2.7) soit fausse, en utilisant la décroissance de $C_{h,\varepsilon,k}(h,\varepsilon,k)\downarrow 0$ on a:

$$\exists(h_n,\varepsilon_n,k_n)\downarrow 0,\quad y_n\in C_{h_n,\varepsilon_n,k_n}(s,x)\quad\text{et}\quad d(y_n,C(s,x))>\varepsilon.$$

$y_n\in$ compact, $\exists$ une sous-suite convergeant vers y et donc

$$d(y,C(s,x))\geqq\varepsilon \text{ or } y\in\overline{\bigcap_{h,\varepsilon,k} C_{h,\varepsilon,k}(s,x)}=\bigcap_{h,\varepsilon,k} C_{h,\varepsilon,k}(s,x)=C(s,x)$$

d'où la contradiction. □

Notons

$$r_k y=\left[\frac{y}{k}\right]k+\frac{k}{2}.$$

On a:

LEMME 16. $\exists M(\rho, k)$ et $\rho_1(\rho, k)$

$$\left|\int (y-x)\pi(dy) - \int [r_k(y)-r_k(x)]\pi(dy)\right| \leqq k, \tag{2.9}$$

$$\left|\int (y-x)^{\otimes 2}\pi(dy) - \int [r_k(y)-r_k(x)]^{\otimes 2}\pi(dy)\right| \leqq Mk, \tag{2.10}$$

$$\int |r_k(y)-r_k(x)|^3\pi(dy) \leqq \rho_1(\rho, k), \quad \forall \pi \in \mathcal{M}_+^1(R^m): \int \pi(dy)(y-x)^3 \leqq \rho. \tag{2.11}$$

Démonstration. (2.9) est évident. (2.10) résulte de

$$\begin{aligned} &|(y-x)^{\otimes 2} - (r_k(y)-r_k(x))^{\otimes 2}| \\ &\qquad \leqq M_1|(y-x)-(r_k(y)-r_k(x))|\,|(y-x)+r_k(y)-r_k(x)| \\ &\qquad \leqq M_1 k \int (2|y-x|+k)\pi(dy). \end{aligned}$$

L'inégalité de Holder donne alors (2.10).

(2.11) résulte de $|r_k(y)-r_k(x)| \leqq |y-x|+k$.

Considérons les multiapplications:

$$C_n = \left[C_{1/n,(1/n)^\gamma} + \mathcal{V}\left(\frac{Mn^{2\gamma}}{n}\right)\right] \cap G$$

$$\bar{C}_n = \left[C_{1/n,(1/n)^\gamma,(1/2^n)} + \mathcal{V}\left(\frac{Mn^{2\gamma}}{n} + M\left(\frac{1}{2^n}\right)\right)\right] \cap G, \qquad 0<\gamma<\tfrac{1}{2},$$

M constante suffisamment grande.

On a C_n est une multiapplication s.c.s., constante par morceaux en x, convergeant ponctuellement pour la topologie de Hausdorff grâce au lemme 15.

Considérons la multiapplication $(s, x) \rightarrow \bar{\Pi}^{n,\bar{C}_n,\delta,3/2,3,(1/2^n)^\theta}$ que nous dénoterons $\bar{\Pi}^n$ avec

$$\bar{\Pi}^{n,\bar{C},\rho,\alpha,\beta,k} = \Big\{\pi \in \mathcal{M}_+^1(R^m), \Big(\int [r_k(y)-r_k(x)]\pi(dy), \int [r_k(y)-r_k(x)]^{\otimes 2}\pi(dy)\Big) \in \bar{C}(s, x)h$$
$$\int |r_k(y)-r_k(x)|^\beta \pi(dy) \leqq \rho h^\alpha\Big\}.$$

Grâce au lemme 16, on a:

$$\pi \in \Pi^{n,C_n,\delta,3/2,3}(s, x) \quad \Rightarrow \quad \pi \in \bar{\Pi}^{n,\bar{C}_n,\delta,3/2,3,(1/n)^\theta}(s, x)$$

Soiţ Φ_n s.c.i. constante par morceaux $\Phi_n =$ régularisée s.c.i. $\sum \Phi_n^i \chi_{[ik,(i+1)k]}$, $\Phi_n \nearrow_n \Phi$ ponctuellement, $\Phi_n^i = \operatorname{Inf} \Phi(x)_{x \in [ik,(i+1)k]}$, Φ s.c.i., $k = 1/2^n$.

Considérons le problème de programmation dynamique

$$\begin{aligned} &\bar{V}_n(T, x) = \Phi_n(x), \\ &\bar{V}_n((n-1)h, x) = \min_{\pi \in \bar{\Pi}^n((n-1)h,x)} \int \bar{V}_n(T, y)\pi(dy), \\ &\bar{V}_n(ih, x) = \min_{\pi \in \bar{\Pi}^n(ih, x)} \int (\bar{V}_n(i+1)h, y)\pi(dy), \\ &\bar{V}_n(0, x) = \min_{\pi \in \bar{\Pi}^n(0, x)} \int \bar{V}_n(h, y)\pi(dy). \end{aligned} \tag{2.12}$$

THÉORÈME 7. *Le problème* (2.12) *admet une solution, de plus,* $\bar{V}(ih, x)$ *est s.c.i. constante par morceaux.*

Démonstration. On le démontre par récurrence. L'application: $\pi \to \int \bar{V}_n(T, y)\pi(dy)$ est s.c.i. car Φ_n est s.c.i.

$\bar{\Pi}^n((n-1)h, x)$ est à valeur compacte car elle est s.c.s. (démonstration analogue à la s.c.s. de Π^n) et donc $\min_{\pi\in\bar{\Pi}^n((n-1)h,\, x)} \int V_n(T, y)\pi(dy)$ existe et $\bar{V}_n(T-h, x)$ est s.c.i. grâce au théorème du maximum [2]. $\bar{V}_n(T-h, x)$ est constante par morceaux car $x \to \bar{\Pi}^n(T-h, x)$ l'est. □

Il existe donc $\bar{\pi}_n^*(ih, x)$, borélienne en x, constante par morceaux réalisant le minimum.

$A\bar{\pi}_n^*$ associons $P^n_{\pi_n^*}$ par la méthode exposée en 1.2.2.

$\{P^n_{\bar{\pi}_n^*}\}$ est étroitement relativement compacte grâce aux lemmes 3 et 16.

Par la méthode exposée dans le théorème d'existence, en remplaçant C par $\bar{C}_n$ on obtient que toute sous-suite convergente de $P^{n'}_{\bar{\pi}_n^*} \to P \in \mathscr{P}(k, C)$ grâce au lemme 16.

THÉORÈME 8. (*Caractérisation d'un contrôle optimal, discrétisation en temps et en espace*). *Si* $\{P^n_{\bar{\pi}_{4n}^*}\}$ *désigne la suite de mesure sur* Ω *définie par interpolation linéaire, sur la chaîne de markov, solution du problème discrétisé* (2.12) *de probabilité de transition* $\bar{\pi}_n^*$; $\{P^n_{\bar{\pi}_n^*}\}$ *est étroitement relativement compacte et tout point adhérent est solution du problème de contrôle optimal* (2.1).

Démonstration. On a démontré l'admissibilité de P^*, démontrons son optimalité; pour cela, il suffit de remarquer que: si $\tilde{P}$ désigne un contrôle optimal, $\tilde{V}$ le coût optimal du problème 2.1, la loi du vecteur; $(X_0, X_h, X_{2h}, \cdots, X_{nh})$ peut s'écrire $\pi^0_{x_0}(dx_1) \times \pi^1_{x_0,x_1}(dx_2) \cdots \pi^{n-1}_{x_0,\cdots,x_{n-1}}(dx_n)\delta_y(dx_0)$ et grâce au lemmes 11 et 12, à la définition de $\bar{C}_n$ et au lemme 14, que

$$\pi^i_{x_0,\cdots x_i} \in \bar{\Pi}^n(ih, x)$$

et donc

$$E^{P^n_{\bar{\pi}_n^*}}(\Phi_n) = \bar{V}_n(0, x_0) \leqq E^{\tilde{P}}\Phi_n \leqq E^{\tilde{P}}(\Phi)$$

car $\Phi_n \leqq \Phi$.

Grâce au lemme 7, on a:

$$E^{P^*}\Phi \leqq \liminf_n E^{P^n_{\bar{\pi}_n^*}}\Phi_n \leqq E^{\bar{P}}(\Phi) \qquad \text{c.q.f.d.}$$

Sur la résolution numérique du problème (2.12). Nous avons vu que le $\overset{*n}{\pi}_{s,x}(dy)$ est constant par morceaux ainsi que V^n.

Notons alors $P^{*i}_{j,j'} = \bar{P}^n_{\bar{\pi}_n^*}(x_{(i+1)h} \in A_j | x_{ih} \in A_i)$ où, $\bar{P}^n_{\bar{\pi}_n^*}$ désigne la loi de probabilité construite sur $R^{(n+1)\times m}$ à partir de $\bar{\pi}_n^*$ comme en (1.8), $A_j = [j/2^n, (j+1)/2^n] + j/2^{(n+1)}$

Notons $\bar{V}^i_{n,j}$ le coût optimal sur le $j$ème morceau à l'instant ih (2.12) se réecrit:

$$\bar{V}^n_{n,j} = \Phi_{n,j}$$

$$\text{(2.13)} \qquad \bar{V}^{i-1}_{n,j} = \underset{p_{j'}}{\text{Min}} \sum_{j'} p_{j'}\bar{V}^i_{n,j'} \qquad \Big(\sum_{j'} p_{j'}(j'-j), \sum_{j'} p_{j'}(j'-j)^{\otimes 2}\Big) \in \bar{C}^n_{i,j},$$

$$\Big(\sum_{j'} p_{j'}|j'-j|^3 \leqq \bar{\rho}h^{3/2}\Big)$$

$$\bar{V}^0_{n,j} = \underset{p_{j'}}{\text{Min}} \sum_{j'} p_{j'}\bar{V}^1_{n,j'}.$$

Le résultat de l'optimisation sera $\overset{*i}{p}_{j,j'}$.

Chaque étape de la récurrence de (2.13) est un problème de programmation mathématique dans R^q, minimisation d'une forme linéaire sous des contraintes convexes, $q = \text{Card}\,\{j\}$.

3. Quelques resultats particuliers. Soit $\{C_n\}$ une suite de multiapplication s.c.s. à valeur convexe dans un compact fixe $C_n \underset{n\nearrow\infty}{\downarrow} C$ ponctuellement; $(s, x) \to \tilde{\Pi}^{n,C_n,\rho,\alpha,\beta}(s, x)$ une multiapplication à valeur dans $\mathcal{M}_+^1(R^m)$ vérifiant

$$(3.1)\quad \begin{aligned} &\forall \pi \in \tilde{\Pi}^{n,C_n,\rho,\alpha,\beta}(s, x) \quad \Rightarrow \quad \left(\int (y-x)\pi(dy), \int (y-x)^{\otimes 2}\pi(dy)\right) \in C_n(s, x)h,\\ &\int |y-x|^{\beta}\pi(dy) \leqq \rho h^{\alpha}; \end{aligned}$$

$$(3.2)\quad \begin{aligned} &\forall c \in C(s, x) \exists \pi \in \tilde{\Pi}^{n,C_n,\rho,\alpha,\beta}(s, x)\left(\int (y-x)\pi(dy), \int (y-x)^{\otimes 2}\pi(dy)\right) \in ch,\\ &\int |y-x|^{\beta}\pi(dy) \leqq \rho h^{\alpha}, \qquad \beta > 2, \quad \alpha > 1. \end{aligned}$$

On suppose de plus que $\tilde{\Pi}^{n,C_n,\rho,\alpha,\beta}$ est s.c.s.

Remarque. Ici, $\tilde{\Pi}^{n,C_n}$ ne représente plus toutes les lois de probabilités ayant le couple (dérive, diffusion) dans C_n, mais seulement une famille de probabilités donnant tous les $c \in C_n$ possibles. Pratiquement, il peut être intéressant de se limiter à une telle famille. Par exemple, on fixe le support de la probabilité de transition (chaîne de markov particulière à état discret).

On construit un problème approché (2.6) en remplaçant Π^n par $\tilde{\Pi}^{n,C_n,\rho,\alpha,\beta}$. Ce problème admet alors une solution optimale que l'on notera $\pi^{*n}_{s,x}$.

Si l'on note P_n la loi de probabilité définie sur les fonctions continues par π^{*n} on a la proposition suivante:

PROPOSITION 2. *Si P désigne un point adhérent à la suite $\{P_n\}$, Φ une fonction continue*

$$E^P\Phi(X_T) \leqq \inf_{\substack{K',C'\\ K'\subset K\\ C'\subset C}} \sup_{P'\in\mathcal{P}(K',C')} \Phi(X_T).$$

Démonstration. $\exists$ une sous-suite encore, notée n:

$$P_n \xrightarrow[n\to\infty]{} P \in \mathcal{P}(K, C).$$

Soit $c'_n \in C'_n$ il lui correspond $\pi'_n \in \tilde{\Pi}^{n,C'_n}$ et donc P'_n mesure sur les fonctions continues, et donc $\exists$ s.s. encore noté n (avec $C'_n \subset C_n$, $C'_n \downarrow C'$):

$$P'_n \xrightarrow[n\to\infty]{} P' \in \mathcal{P}(K', C')$$

or, on a $E^{P_n}\Phi(X_T) \leqq E^{P'_n}\Phi(X_T)$ et donc, en passant à la limite:

$$E^P\Phi(X_T) \leqq \sup_{P'\in\mathcal{P}(K',C')} E^{P'}\Phi(X_T)$$

d'où la proposition. □

De cette proposition, on déduit un corollaire lorsque le couple (K', C') est tel qu'il y ait unicité au problème de martingale.

COROLLAIRE. Sous les hypothèses de la proposition 2:

$$E^P\Phi(X_T) \leqq \operatorname*{Inf}_{\substack{\left.\begin{array}{l}a(t,x)\\ b(t,x)\end{array}\right\}\text{lipschitz en } x \text{ continue en } t\\ (a,b)\in C\\ x\in K}} E^{P^{x,a,b}}\Phi(X_t)$$

où $P^{x,a,b}$ désigne la solution du problème de martingale au sens classique [22].

Remarque. Un résultat de ce type dans un cadre moins général est démontré dans [14].

Remarque. Le contre exemple suivant montre que dans le cas où a dégénère, on ne peut espérer mieux.

Considérons:

$$\dot{x} = g(x), \qquad g = 2\operatorname{sgn}(x)\sqrt{|x|}, \qquad x(0) = 0.$$

Cette équation admet au moins trois solutions:

$$x = 0,$$
$$x = \pm t^2.$$

Supposons que l'on veuille maximiser $Ef_M(X_T)$ sur $\mathcal{P}(0, (0, g))$ avec

$$f_M(x) = \begin{cases} x^2 & \text{si } x \leqq M, \\ M^2 & \text{si } x \geqq M. \end{cases}$$

La solution de ce problème sera la mesure de Dirac sur la solution $x = t^2$ or,

$$\Pi^{n,(0,g)}(s, 0) = \delta_0$$

et donc, P^n sera la masse de Dirac sur la trajectoire $x = 0$ qui n'est pas la bonne solution.

Remarque. On a toujours supposé que C_n est à valeur convexe. Notons $\bar{C}_n$ le convexifié de C_n. Alors $\Pi^{n,\bar{C}_n}$ est le convexifié de Π^{n,C_n} et donc les coûts V_n et $\bar{V}_n$ associés sont les mêmes, $\exists \pi^n$ optimal pour $\Pi^{n,\bar{C}_n}$ appartenant à Π^{n,C_n}. Cette remarque montre que dans le cas où C n'est pas convexe, on peut trouver une suite $c_n \in C$ optimal tel que la conclusion de la proposition 2 soit vérifiée.

Un cas particulier important.

$$C(s, x) = B(s, x) \otimes a(s, x),$$

B multiapplication s.c.s. à valeur dans R^{m+1}, $a(s, x)$ fonction continue > 0.

On considère le problème

$$\operatorname*{Min}_{(b_0,P)\in\tilde{\mathcal{P}}(y,C)} E^P \int_0^t b_0(s, x_s)\, ds + \Phi(X_T)$$

où $\tilde{\mathcal{P}}(y, c)$ désigne l'ensemble couples (fonction borélienne, mesure de probabilités, solution du problème de martingale $(b_1, \cdots, b_m; a)$) tels que $(b_0, b_1, \ldots, b_m)\otimes a \in C$, $(b_0, \cdots, b_m)(s, \omega)$ prévisible.

THÉORÈME 9.

$$\operatorname*{Min}_{(b_0,P)\in\tilde{\mathcal{P}}(y,C)} E^P \int_0^T b_0(s, X_s) ds + \Phi(X_T) = \operatorname*{Inf}_{\substack{b\\ \text{section lipschitz en } x,\\ \text{continue en } t \text{ de } B}} E^{P^{y,b,a}} \int_0^T b_0(s, X_s)\, ds + \Phi(X_T).$$

Démonstration. On sait [4, th.IV.6]

$$\underset{\substack{b(s,x)\in B(s,x)\\ b \text{ borélienne}}}{\text{Min}} E^{P y,b,a} \int_0^t b_0(s, x_s)\, ds + \Phi(X_T)$$

$$= \underset{\substack{b(s,\omega)\in B(s,X_s(\omega))\\ b \text{ progressivement}\\ \text{mesurable.}}}{\text{Min}} E^{P y,b,a} \int_0^t b_0(s, \omega)\, ds + \Phi(X_T)$$

et on sait que les minimums existent. À $(b_1(s, x), \cdots, b_n(s, x)) \to P$ unique [22] et donc minimiser par rapport à $b \in B$ ou a $(b_0, P) \in \tilde{\mathcal{P}}(y, C)$ donne le même coût optimal.

Soit donc $b^*(s, x) \in \underset{b\in B}{\text{Arg min}}\, E \int_0^T b_0(s, X_s)\, ds + \Phi(X_T)$. $\exists \{b_n\}$ lipschitz en x, continu en t

$$b_n \xrightarrow[n\to\infty]{} b^* \text{ dans } \sigma(L^\infty([0, T]\times R^m), L^1([0, T]\times R^m))$$

et le th IV.3 de J. M. Bismut [4] montre que

$$E^{P(y,b_1^n\cdots,b_m^n,a)} \int_0^T b_n^0(s, X_s)\, ds + \Phi(X_T) \xrightarrow[n\to\infty]{} E^{P(y,b_1^*\cdots b_m^*,a)} \int_0^T b^0(s, X_s)\, ds + \Phi(X_T),$$

et donc P construit au début du § 3 est optimal dans ce cas particulier.

REFERENCES

[0] A. BENSOUSSAN AND J. LESOURNE, *An unreverisble investment*, Cahier des Mathématiques de la Décision, Paris IX, 1976.

[1] A. BENSOUSSAN AND J. L. LIONS, *Applications des inéquations variationnelles en Contrôle Stochastique*, Dunod, Paris 1978.

[2] C. BERGE, *Espace topologique, fonctions multivoques*, Dunod, Paris, 1959.

[3] P. BILLINGSLEY, *Convergence of probability measures*, J. Wiley, New York, 1968.

[4] J. M. BISMUT, *Th. Prob. du contrôle des diffusions*, Mémoires American Mathematical Society, Vol. 4, n° 167, 1976.

[5] N. BOURBAKI, *Integration*, Ch. IX, Herman, Paris.

[6] C. CASTAING, *Sur les multiapplications mesurables*, RIRO, n° 1 (1967), pp 91–126.

[7] C. DELACHERIE AND P. A. MEYER, *Probabilités et potentiel*, Herman, Paris, 1975.

[8] F. DELEBECQUE AND J. P. QUADRAT, *Contribution of stochastic control, singular perturbation, averaging and team theories to an example of large scale systems: Management of hydropower production*, IEEE Trans. Automatic Control, AC-23 (1978), pp. 209–222.

[9] I. EKELAND AND R. TEMAM, *Analyse convexe et problèmes variationnelles*, Dunod, Paris, 1974.

[10] W. H. FLEMING AND W. RISHEL, *Optimal Deterministic and Stochastic Control*, Springer-Verlag, Berlin, 1975.

[11a] M. GOURSAT AND J. P. QUADRAT, *Analyse numérique d'inégalités variationnelles elliptique associées à des problèmes de temps d'arrêt optimal en contrôle stochastique.* Rapport IRIA-Laboria n° 154, Rocquencourt, France 1976.

[11b] ———, *Analyse numérique d'inégalités quasi-variationnelles elliptique associées à des problèmes de contrôle impulsionnel*, Rapport IRIA-Laboria n° 186, Rocquencourt, France, 1976.

[12] N. V. KRILOV, *On control of the solution of a stochastic integral equation with degeneration*, Izv. Akad. Nauk USSR, Ser. Math., 36 (1972), pp. 249–262.

[13] H. KUSHNER, *A survey of some applications of probability and stochastic control theory to finite diffusion methods for degenerate elliptic and parabolic equations*, SIAM Review, 18 (1976), pp. 545–577.

[14] H. KUSHNER AND CHEN-FU-YU, *Approximation, existence and numerical procedures, for optimal stochastic control*, J. Math. Appl., 45 (1974), pp. 563–587.

[15] P. J. LAURENT, *Approximation et optimisation*, Herman, Paris, 1972.

[16] R. C. MERTON, *Lifetime Portefolio selection underuncertainty, the continuous case.* Rev. Econ. Statist., LI, pp. 247–257.

[17] J. NEVEU, *Bases mathématiques du calcul des probabilités*, Masson, Paris, 1964.

[18] P. PRIOURET, *C.R. Ecole d'Eté de St. Flour*, Lectures Notes in Mathematics, Springer-Verlag, Berlin, 1974.

[19] J. P. QUADRAT, *Analyse numérique de l'équation de Bellman stochastique*, Rapport de Recherche IRIA-Laboria n° 140, Rocquencourt, France, 1975.

[20] M. ROBIN, *Contrôle impulsionnel des processus de Markov*, Thèse, Paris IX, 1978.

[21] R. SENTIS, *Discrétisation élémentaire d'un problème de commande optimal*, Cahier de Mathématiques de la Décision n° 7712, Paris IX, 1977.

[22a] D. W. STROOCK AND S. R. S. VARADHAN, *Diffusion with continuous coefficients I and II.* Comm. Pure Appl. Math., 22 (1967) pp. 479–530.

[22b] ———, *Diffusion with boundary conditions*, Ibid., 24 (1971), pp. 142–225.

[23] M. VALADIER, *Multiapplications mesurables à valeurs convexes compactes*, J. Math. Pures et Appl., 50 (1971) pp. 265–297.

[24] L. C. YOUNG, *Lectures on the Calculus of Variations ad Optimal Control Theory*, Saunders, Philadelphia, 1969.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 2, March 1980

0363-0129/80/1802-0009 $01.00/0

BOUNDARY CONTROL PROBLEMS WITH CONVEX COST CRITERION*

VIOREL BARBU†

Abstract. A class of boundary-distributed linear control systems in Banach spaces is studied. A maximum principle for a convex control problem associated with such systems is obtained.

1. Introduction. In the past decade or so have been developed several abstract settings to describe the distributed control systems on a domain Ω in which the control is exercised through the boundary Γ. We mention in this context the Hilbert theory of Lions (see [13]) and the semigroup approach of boundary control system developed by Fattorini [11]. In all these approaches we encounter the same difficulty: for existence of a sufficiently regular solution, say continuous in t, to state system, the control must be taken in a space of sufficiently smooth functions on $[0, T]$. Starting from Fattorini's model we introduce here a general class of boundary control systems in a Banach space which are well posed whenever the boundary control is a summable function in t. A related class of linear control processes in Banach space has been considered by Curtain and Prichard [8] and by Balakrishnan [1]. These systems are described in § 2. In § 3 we show that the parabolic boundary control systems governed by Dirichlet and Neumann problems are covered by the preceding theory. For such systems we consider in § 4 the control problem with convex cost criterion and formulate the maximum principle in the subdifferential form. This is the main result of this paper which will be proved in § 5 paralleling the treatment in [3], [5] (see also [6, Chap. IV]). A duality result is given in § 6.

After this work was submitted, we learned of two other related works: (i) In [2] Balakrishnan shows that the solution to a parabolic boundary control equation with L^2-inputs can be expressed as a "mild" solution to an operator equation of the form (2.8) below. We shall see in § 3 that this fact is also implied by our theory. (ii) In [12] I. Lasiecka studies the regularity of optimal boundary controls for parabolic equations with quadratic cost criterion. In particular an estimate of the form (2.9) is obtained for the Dirichlet boundary value problem.

Notation. Given a Banach space X with norm $\|\cdot\|$ and a real interval $[0, T]$ we denote by $C(0, T; X)$ the Banach space of all continuous functions $x: [0, T] \to X$ endowed with the standard norm. For each $1 \leqq p \leqq \infty$ denote by $L^p(0, T; X)$ the space of all p-summable functions on $[0, T]$ with values in X.

Given another Banach space Y we denote by $L(X, Y)$ the algebra of linear continuous operators from X to Y endowed with the usual norm $\|\cdot\|_{L(X,Y)}$. If A is a densely defined linear operator on X denote by $D(A)$ its domain provided with the graph norm.

Let $\varphi: X \to \bar{R} =]-\infty, +\infty]$ be a lower semicontinuous convex function. The subdifferential $\partial\varphi: X \to X^*$ is defined by

$$\partial\varphi(x_0) = \{x_0^* \in X^*; \varphi(x_0) - \varphi(x) \leqq (x_0^*, x_0 - x) \text{ for all } x \in X\}.$$

Here X^* is the dual space of X (which is assumed real) and $(\cdot, \cdot)$ denotes the pairing between X and X^*. For each $\lambda > 0$ denote by $\varphi_\lambda: X \to R$ the function defined by

$$\varphi_\lambda(x) = \inf\left\{\frac{1}{2\lambda}\|x - y\|^2 + \varphi(y); y \in X\right\}, \qquad x \in X. \tag{1.1}$$

* Received by the editors June 12, 1978.

† Faculty of Mathematics, University of Iasi, Iasi, Romania.

If X and X^* are strictly convex and reflexive then φ_λ is convex, Gâteaux differentiable on X and (see [6], p. 107).

$$\partial\varphi_\lambda(x) = (\partial\varphi)_\lambda x \quad \text{for all } \lambda > 0, x \in X \tag{1.2}$$

$$\varphi_\lambda(x) = \frac{\lambda}{2}\|\partial\varphi_\lambda(x)\|^2 + \varphi(J_\lambda x); \qquad \lambda > 0, x \in X, \tag{1.3}$$

where $J_\lambda x = x_\lambda$ is the solution to

$$\phi(x_\lambda - x) + \lambda\,\partial\varphi(x_\lambda) \ni 0 \tag{1.4}$$

and

$$(\partial\varphi)_\lambda x = -\lambda^{-1}\phi(x_\lambda - x). \tag{1.5}$$

Here $\phi: X \to X^*$ is the duality mapping of X. For other notions and results of convex analysis relevant to this paper we refer the reader to [6], [10], [16] and [17].

2. Boundary control systems. To begin with let us briefly describe Fattorini's theory of boundary-distributed control system (see [11]).

Let E be a (real or complex) Banach space and let σ be a closed, linear densely defined operator in E. Let τ be a linear operator (the boundary operator) with domain in E and range in some Banach space X. Finally, let U_1 and U_2 be two Banach spaces which in the sequel will be referred to as the control spaces of the system.

The control system we shall consider is

$$y'(t) = \sigma y(t) + B_1 u_1(t) + f(t), \qquad \tau y(t) = B_2 u_2(t) \text{ over } [0, T] \tag{2.1}$$

with initial value condition

$$y(0) = y^0 \tag{2.2}$$

where B_1: $U_1 \to E$ and B_2: $U_2 \to X$ are linear continuous operators and $[0, T]$ is a fixed interval. The controllers $u_1(\cdot)$ and $u_2(\cdot)$ are summable functions on $[0, T]$ with values in U_1 and U_2, respectively. We shall call u_1, u_2 the distributed and boundary control, respectively. Here f is a given E-valued summable function.

In applications the state space E is a space of functions on some domain Ω of the Euclidean space R^n, σ is a partial differential operator on Ω and τ a partial differential operator acting on the boundary Γ of Ω.

Assumption I. $D(\sigma) \subset D(\tau)$ and the restriction of τ to $D(\sigma)$ is continuous relative to graph norm of $D(\sigma)$.

Let $A: E \to E$ be the linear operator defined by

$$D(A) = \{y \in D(\sigma);\ \tau y = 0\}, \qquad Ay = \sigma y \quad \text{for } y \in D(A). \tag{2.3}$$

Assumption II. The operator A is the infinitesimal generator of a strongly continuous semigroup $\{S(t);\ t \geqq 0\}$ on E.

Assumption III. There exists a linear continuous operator B: $U_2 \to E$ such that

$$\sigma B \in L(U_2, E), \qquad \tau(Bu) = B_2 u \quad \text{for all } u \in U_2 \tag{2.4}$$

$$\|Bu\|_E \leqq C\|B_2 u\|_X \quad \text{for all } u \in U_2 \tag{2.5}$$

where C is some positive constant.

In terms of A and B system (2.1) can be written as

$$\begin{aligned} y' &= Az + B_1 u_1 + \sigma B u_2 + f, \qquad 0 \leqq t \leqq T, \\ y &= z + Bu_2. \end{aligned} \tag{2.6}$$

If $u_2(\cdot)$ is continuously differentiable on $[0, T]$ then z can be defined as a "mild" solution to the Cauchy problem

$$z' = Az + B_1u_1 + \sigma Bu_2 - Bu_2' + f,$$
$$z(0) = y^0 - Bu_2(0).$$

In this way one can define the solution y to system (2.1), (2.2) by the variation of constant formula

$$y(t) = S(t)(y^0 - Bu_2(0)) + Bu_2(t) + \int_0^t S(t-s)(B_1u_1(s) + \sigma Bu_2(s) - Bu_2'(s) + f(s))\,ds. \tag{2.7}$$

Since the differentiability of controller u_2 represents an unrealistic and severe requirement, we are led to extend the concept of solution to (2.1), (2.2) for general $u_2 \in L^1(0, T; U_2)$.

Integrating (formally) by parts in (2.7) we get

$$y(t) = S(t)y^0 - \int_0^t AS(t-s)Bu_2(s)\,ds + \int_0^t S(t-s)(B_1u_1(s) + \sigma Bu_2(s) + f(s))\,ds. \tag{2.8}$$

In general, unless we impose further assumptions on $S(t)$ and B, the right-hand side of (2.8) is not well defined.

Assumption IV. For each $t \in]0, T]$ and $u \in U_2$, $S(t)Bu \in D(A)$. There exists a positive function $\gamma \in L^1(0, T)$ such that

$$\|AS(t)B\|_{L(U_2,E)} \leqq \gamma(t) \quad \text{a.e. } t \in]0, T[. \tag{2.9}$$

Since $S(t)Bu \in D(A)$ for all $u \in U_2$, by the closed graph theorem we deduce that the operator $AS(t)B$ is continuous from U_2 to E so that (2.9) makes sense.

Assumptions IV implies that for every $u_2 \in L^1(0, T; U_2)$, the function $t \to \int_0^t AS(t-s)Bu_2(s)\,ds$ is well defined as an element of $L^1(0, T; E)$. By definition, for each $y_0 \in E$, $f \in L^1(0, T; E)$, $u_1 \in L^1(0, T; U_1)$ and $u_2 \in L^1(0, T; U_2)$, the function $y \in L^1(0, T; E)$ defined by (2.8) is the solution of *distributed-boundary* control system (2.1), (2.2).

Since the function $t \to \int_0^t S(t-s)Bu_2(s)\,ds$ belongs to $L^1(0, T; D(A))$, $y(\cdot)$ may be expressed in the following equivalent form

$$y(t) = S(t)y^0 - A\int_0^t S(t-s)Bu_2(s)\,ds + \int_0^t S(t-s)(B_1u_1(s) + \sigma Bu_2(s) + f(s))\,ds \quad \text{a.e. } t \in]0, T[.$$

Let χ_0 be a fixed number in $\rho(A)$ (the resolvent of A) and let $\Pi = A - \chi_0 I$ (I is the identity operator). Thus y may be regarded as solution to *distributed* control system

$$w' = Aw + D_1u_1 + D_2u_2 + \Pi^{-1}f, \qquad y = \Pi w \tag{2.10}$$

where

$$D_1 = (A - \chi_0 I)^{-1}B_1, \qquad D_2 = (A - \chi_0 I)^{-1}(\sigma B - \chi_0 B) - B. \tag{2.11}$$

Denote by U the product space $U_1 \times U_2$ and by Λ: $U \to E$ the linear continuous operator given by

$$\Lambda(u_1, u_2) = D_1u_1 + D_2u_2, \quad \text{for } u_1 \in U_1,\ u_2 \in U_2. \tag{2.12}$$

Then we may rewrite system (2.10) as

$$\begin{aligned} w' &= Aw + \Lambda u + \Pi^{-1}f, \qquad 0 \leqq t \leqq T, \\ y &= \Pi w. \end{aligned} \tag{2.13}$$

Thus, we are led to interpret the solution y to (2.1) as the *observed* value of a control system of the form (2.13) with unbounded *observation operator* Π (we refer the reader to [9] for definition and theory of observation for infinite dimensional systems).

Remark 1. If $u_2 \in L^p(0, T; U_2)$ then $y \in L^p(0, T; E)$. If $\gamma \in L^{p'}(0, T)$; $1/p + 1/p' = 1$ then we see by (2.8) and (2.9) that $y \in C(0, T; E)$.

3. Examples. It should be observed that Assumption IV has some severe implications for system (2.1). In particular if the range $R(B)$ of B is, say, all of E then the semigroup $S(\cdot)$ must be analytic (see e.g. [18, p. 254]). However, this conditions is less restrictive than it might at first appear to be. We shall see here that it is satisfied in some notable cases.

Mixed Dirichlet problem. Let Ω be a bounded and open subset of R^n with a sufficiently smooth boundary Γ.

Consider the boundary control system

$$\begin{aligned} &\frac{\partial y}{\partial t} - \Delta y = f && \text{in } Q = \Omega \times]0, T[, \\ &y|_\Gamma = u && \text{for } t \in [0, T], \\ &y(x, 0) = y_0(x) && \text{for } x \in \Omega, \end{aligned} \tag{3.1}$$

where $u \in L^2(\Sigma)(\Sigma = \Gamma \times]0, T[)$, $y_0 \in L^2(\Omega)$ and $f \in L^2(Q)$.

To formulate this as a boundary control system of the form (2.1) we define $E = U_1 = L^2(\Omega)$, $X = H^{-1/2}(\Gamma)$, $U_2 = L^2(\Gamma)$, $B_1 \equiv 0$, $B_2 \equiv I$ and

$$D(\sigma) = \{y \in L^2(\Omega);\ \Delta y \in L^2(\Omega)\}, \qquad \sigma = \Delta. \tag{3.2}$$

($H^k(\Omega)$,$H^\alpha(\Gamma)$ are usual Sobolev spaces on Ω and Γ.)

The operator τ is the "trace" operator $\gamma_0 y$ which is well defined and belongs to $H^{-1/2}(\Gamma)$ for each $y \in D(\sigma)$ (see Lions–Magenes [15, vol. 1]). The operator A is given by

$$A = \Delta, \qquad D(A) = H_0^1(\Omega) \cap H^2(\Omega). \tag{3.3}$$

Clearly Assumptions I and II are satisfied. To verify III and IV we define the linear operator B: $U_2 = L^2(\Omega) \to L^2(\Gamma)$ by $Bu = w_u$ where $w_u \in L^2(\Omega)$ is the unique (generalized) solution to the Dirichlet boundary-value problem

$$\begin{aligned} \Delta w_u &= 0 \quad \text{in } \Omega, \\ w_u|_\Gamma &= u. \end{aligned} \tag{3.4}$$

In other words,

$$\int_\Omega w_u \Delta\psi \, dx = \int_\Gamma u \frac{\partial\psi}{\partial n} \, d\sigma \quad \text{for all } \psi \in H_0^1(\Omega) \cap H^2(\Omega). \tag{3.5}$$

Here $\partial\psi/\partial n$ denotes the outward normal derivative of ψ which is well defined as an element of $H^{1/2}(\Gamma)$. We need the following lemma.

LEMMA 1. *For every $u \in H^{-1/2}(\Gamma)$ problem* (3.4) *has a unique solution $w_u \in L^2(\Omega)$ satisfying*

$$\|w_u\|_{L^2(\Omega)} \leqq C_1\|u\|_{H^{-1/2}(\Gamma)}. \tag{3.6}$$

If $u \in H^{1/2}(\Gamma)$ then $w_u \in H^1(\Omega)$ and

$$\|w_u\|_{H^1(\Omega)} \leqq C_2|u\|_{H^{1/2}(\Gamma)}. \tag{3.7}$$

Here C_i, $i = 1, 2$ are positive constants independent of u.

Proof. Let $u \in H^{-1/2}(\Gamma)$. The existence of w_u satisfying (3.5) is well-known (see e.g. [13, p. 72]). It follows from the fact that the operator Δ with domain $H_0^1(\Omega) \cap H^2(\Omega)$ is onto on $L^2(\Omega)$ and the functional $\psi \to \int_\Gamma u\, \partial\psi/\partial n\, d\sigma$ is continuous on $H_0^1(\Omega) \cap H^2(\Omega)$. Also the uniqueness of such w_u is immediate. On the other hand by "trace" inequality

$$\left\|\frac{\partial\psi}{\partial n}\right\|_{H^{1/2}(\Gamma)} \leqq C\|\psi\|_{H_0^1(\Omega)\cap H^2(\Omega)} \tag{3.8}$$

and by (3.5) we see that

$$|w_u(\varphi)| \leqq C\|u\|_{H^{-1/2}(\Gamma)}\|A^{-1}\varphi\|_{H^2(\Omega)} \quad \text{for all } \varphi \in L^2(\Omega)$$

thereby proving (3.6).

Suppose now that $u \in H^{1/2}(\Gamma)$. Then again by the "trace" theorem there is $y_u \in H^1(\Omega)$ such that $\tau y_u = u$ and

$$\|y_u\|_{H^1(\Omega)} \leqq C\|u\|_{H^{1/2}(\Gamma)} \tag{3.9}$$

where C is independent of u. On the other hand, it follows from (3.5) and Green's formula that the function $z = y_u - w_u$ satisfies the equation

$$\int_\Omega z\,\Delta\psi\, dx = \int_\Omega \operatorname{grad} y_u \operatorname{grad} \psi\, dx$$

for all $\psi \in H_0^1(\Omega) \cap H^2(\Omega)$. Hence

$$|z(\varphi)| \leqq C\|y_u\|_{H^1(\Omega)}\|\varphi\|_{H^{-1}(\Omega)} \quad \text{for all } \varphi \in L^2(\Omega). \tag{3.10}$$

This shows that $z \in H_0^1(\Omega)$. Hence $w_u \in H^1(\Omega)$. Furthermore, by (3.9) and (3.10) we get (3.7) as claimed. This completes the proof of the lemma.

In particular, Lemma 1 shows that Assumption III is satisfied. As regards Assumption IV, we observe first that by (3.6) and (3.7) it follows that $Bu \in (L^2(\Omega), H^1(\Omega))_{1/2}$ for all $u \in L^2(\Gamma)$ and

$$\|Bu\|_{H^{1/2}(\Omega)} \leqq C\|u\|_{L^2(\Gamma)} \quad \text{for all } u \in L^2(\Gamma). \tag{3.11}$$

Here $(L^2(\Omega), H^1(\Omega))_{1/2}$ denotes the interpolation space $\{y(x, 0);\ y \in L^2(R^+; H^1(\Omega)),\ \partial y/\partial t \in L^2(R^+; L^2(\Omega))\}$ (see e.g. [14]).

Inasmuch as the semigroup $S(\cdot)$ generated by A is analytic (see e.g. [15, vol. II, p. 22]) and [19, p. 254] we have

$$\|AS(t)y_0\|_{L^2(\Omega)} \leqq Ct^{-1}\|y_0\|_{L^2(\Omega)} \tag{3.12}$$

for all $y_0 \in L^2(\Omega)$ and $t > 0$.

On the other hand according to an interpolation result due to Lions [14], we have for each $y_0 \in H^1(\Omega)$,

$$\|AS(t)y_0\|_{L^2(\Omega)} \leqq Ct^{-3/4}\|y_0\|_{H^1(\Omega)} \quad \text{for } t>0. \tag{3.13}$$

Thus interpolating between spaces $H^1(\Omega)$ and $L^2(\Omega)$ we see by (3.11), (3.12) and (3.13) that

$$\|AS(t)Bu\|_{L^2(\Omega)} \leqq Ct^{-7/8}\|u\|_{L^2(\Gamma)} \tag{3.14}$$

for all $u \in L^2(\Gamma)$ and $t>0$. Here C is a positive constant independent of u and t. In other words, inequality (2.9) holds with $\gamma(t)=Ct^{-7/8}$. In particular it follows that for each $p \in [1, \infty[$ the operator $y \to A\int_0^t S(t-s)Bu(s)\,ds$ is continuous from $L^p(0, T; L^2(\Gamma))$ to $L^p(0, T; L^2(\Omega))$.

Thus we have shown that *system* (3.3) *satisfies Assumptions* I *up to* IV *with A and B defined by* (3.3) *and* (3.4) *respectively.*

Remark. In [12] it has been shown by a different approach that estimate (2.9) holds with $\gamma(t)=Ct^{-\alpha}$ where $\alpha>\frac{3}{4}$.

Mixed Neumann problem. Consider the boundary control system

$$\begin{aligned} &\frac{\partial y}{\partial t}-\Delta y=f \quad \text{in } Q=\Omega\times]0, T[,\\ &\frac{\partial y}{\partial n}+\alpha y=u \quad \text{in } \Sigma=\Gamma\times]0, T[,\\ &y(0, x)=y_0(x), \qquad x\in\Omega \end{aligned} \tag{3.15}$$

where $y_0 \in L^2(\Omega)$, $f \in L^2(Q)$ and u (the boundary control) belongs to $L^2(\Gamma)$. Here α is a nonnegative constant.

Define $E=L^2(\Omega)$, $U_2=X=L^2(\Gamma)$, $B_1\equiv 0$, $B_2\equiv I$, $\sigma y=\Delta y$, $D(\sigma)=H^2(\Omega)$ and $\tau y=\alpha y+\partial y/\partial n$. The operator A is given by

$$Ay=\Delta y \quad \text{on} \quad D(A)=\left\{y\in H^2(\Omega); \alpha y+\frac{\partial y}{\partial n}=0\right\}. \tag{3.16}$$

Define the operator $B: L^2(\Gamma)\to L^2(\Omega)$ by $Bu=z_u$ where $z_u \in H^1(\Omega)$ is the solution to boundary-value problem

$$\begin{aligned} &z_u-\Delta z_u=0 \quad \text{in } \Omega,\\ &\alpha z_u+\frac{\partial z_u}{\partial n}=u \quad \text{in } \Gamma. \end{aligned} \tag{3.17}$$

Consider on the product space $H^1(\Omega)\times H^1(\Omega)$ the bilinear functional

$$a(y, \varphi)=\int_\Omega (y\varphi+\text{grad } y \text{ grad } \varphi)\,dx-\int_\Gamma (u-\alpha y)\varphi\,d\sigma \tag{3.18}$$

where $u \in H^{-1/2}(\Gamma)$ (the integral $\int_\Gamma u\varphi\,d\sigma$ must be regarded as the value of u at $\gamma_0\varphi \in H^{1/2}(\Gamma)$). Since a is coercive, there is $z_u \in H^1(\Omega)$ satisfying $a(z_u, \varphi)=0$ for all $\varphi \in H^1(\Omega)$. In other words, $z_u=Bu$ is the solution to (3.17). From (3.18) we see also that

$$\|z_u\|_{H^1(\Omega)} \leqq C\|u\|_{H^{-1/2}(\Gamma)}. \tag{3.19}$$

In particular, we have shown that Assumption III holds. To verify Assumption IV we notice that since the operator $-A$ is self-adjoint and positive, we have

$$(3.20)\qquad \int_0^T \|AS(t)y_0\|^2_{L^2(\Omega)}\,dt \leqq C\|y_0\|^2_{D((-A)^{1/2})}$$

for all $y_0 \in D((-A)^{1/2}) = H^1(\Omega)$. Let δ be the scalar function defined by

$$\delta(t) = \liminf_{n\to\infty} \|A_n S(t)\|_{L(H^1(\Omega),L^2(\Omega))}, \qquad t \in [0, T],$$

where $A_n = A(I + n^{-1}A)^{-1}$ for $n = 1, 2, \cdots$. Obviously,

$$(3.21)\qquad \|AS(t)\|_{L(H^1(\Omega),L^2(\Omega))} \leqq \delta(t) \quad \text{for } t \in]0, T]$$

while inequality (3.20) implies that

$$\int_0^T \|A_n S(t)\|^2_{L(H^1(\Omega),L^2(\Omega))}\,dt \leqq C \quad \text{for all } n.$$

Therefore by Fatou's lemma it follows that $\delta \in L^2(0, T)$ so that inequality (3.21) along with (3.19) yields

$$(3.22)\qquad \|AS(t)Bu\|_{L^2(\Omega)} \leqq C\delta(t)\|u\|_{L^2(\Gamma)}, \qquad t \in]0, T[,$$

for all $u \in L^2(\Gamma)$. Thus Assumption IV is satisfied with some function $\gamma = C\delta \in L^2(0, T)$.

Remark 2. Estimates (3.14) and (3.21) can be impoved if we take the boundary control u to be *more regular*. For instance in the example of (3.1) if $U_2 = H^{1/2}(\Gamma)$ then Assumption IV is satisfied with $\gamma(t) = Ct^{-3/4}$. It should be also emphasized that in preceding examples the Laplacian Δ can be replaced by any second order symmetric elliptic operator A_0 on Ω on the form

$$A_0 y = \sum_{i,j=1}^n (a_{ij}(x)y_{x_i})_{x_j} - a_0(x)y$$

where $a_{ij} \in C^1(\bar{\Omega})$ and $a_0 \in L^\infty(\Omega)$.

4. Integral convex cost criteria. In this section we consider the following unconstrained boundary-distributed control problem: minimize

$$(\mathrm{P})\qquad \int_0^T L_0(t, y, u_1, u_2)\,dt + l(y(0), y(T))$$

in $y \in C(0, T; E)$, $u_i \in L^p(0, T; U_i)$; $i = 1, 2$, subject to state equation (2.1).

Here $2 \leqq p < \infty$ and L_0: $[0, T] \times E \times U_1 \times U_2 \to \bar{R} =]-\infty, +\infty]$, l: $E \times E \to \bar{R}$ are given functions which will be made precise later.

From now on we shall assume that the spaces E, U_1 and U_2 *are reflexive and strictly convex* together with their duals E^*, U_1^* and U_2^*.

We denote by U the product space $U_1 \times U_2$ and denote by $|\cdot|$ (resp. $\|\cdot\|$) the norm in E (resp. U). The pairing between E, E^* and U, U^* will be denoted by $(\cdot, \cdot)$ and $\langle \cdot, \cdot \rangle$, respectively.

Finally we denote by F: $E \to E^*$ and Ξ: $U \to U^*$ the duality mapping of E and U, respectively. It should be recalled (see e.g. [4, p. 13]) that under our assumptions F and Ξ are single valued, injective and demicontinuous (i.e. strongly-weakly continuous).

We shall assume also that Assumptions I–IV are satisfied and the function γ in condition (2.9) belongs to $L^{p'}(0, T)$ where

$$p' = p(p-1)^{-1}.$$

This condition on control spaces $L^p(0, T; U_i)$ was imposed in order to assure the continuity of the "mild" solution y to system (2.1) and so to give a meaning to functional $l(y(0), y(T))$. If the final value $y(T)$ is not present in problem (P) then p can be chosen arbitrary in the interval $]1, +\infty[$.

As seen in § 2, the state system (2.1) can be brought into the form (2.13) where $u=(u_1, u_2)\in U_1\times U_2=U$, $\Pi=A-\chi_0 I(\chi_0\in\rho(A))$ and Λ is given by (2.12). Let L: $[0, T]\times E\times U\to\bar{R}$ be defined by

$$L(t, y, u)=L_0(t, y, u_1, u_2) \quad \text{for } u=(u_1, u_2).$$

Then problem (P) can be equivalently expressed as: minimize

$$\text{(P)} \qquad \int_0^T L(t, y, u)\, dt+l(y(0), y(T))$$

in $y\in C(0, T; E)$ and $u\in L^p(0, T; U)$ subject to

$$\text{(4.1)} \qquad \begin{aligned} w' &= Aw+\Lambda u+\Pi^{-1}f,\\ y &= \Pi w. \end{aligned}$$

Here $f\in L^1(0, T; E)$ is a given function. The solution of (4.1) must be understood in the sense of (2.8), i.e.,

$$\text{(4.1')} \qquad y(t)=S(t)y(0)+\int_0^t \Pi S(t-s)\Lambda u(s)\, ds+\int_0^t S(t-s)f(s)\, ds, \qquad 0\leqq t\leqq T.$$

We notice that Assumption IV and the condition imposed on p imply that

$$\text{(4.2)} \qquad \|\Pi S(t)\Lambda\|_{L(U,E)}\leqq\zeta(t) \quad \text{for } t\in]0, T[$$

where $\zeta\in L^{p'}(0, T)$; $1/p+1/p'=1$.

Beside the above assumptions on E, U, A and B further hypotheses on L_0 and l must be imposed.

(A) For each t the functions $L(t)$ and l are lower semicontinuous and convex on $E\times U$ and $E\times E$, respectively. Furthermore, the following conditions hold.

(a) For each $(y, v)\in E\times U$ the functions $L(t, y, v)$: $[0, T]\to\bar{R}$ and $J_\nu^L(t, y, v)$: $[0, T]\to E\times U$ are Lebesgue measurable $(\nu>0)$.

(b) There exist $r_0\in L^2(0, T; E^*)$, $s_0\in L^\infty(0, T; U^*)$ and $g_0\in L^1(0, T)$ such that for all $(y, u)\in E\times U$,

$$L(t, y, u)\geqq(y, r_0(t))+\langle u, s_0(t)\rangle+g_0(t), \quad \text{a.e. } t\in]0, T[.$$

(c) For each $y_0\in E$ there are a neighborhood $\mathcal{O}$ of y_0, the functions $\alpha\in L^{p'}(0, T)$, $\varkappa\in L^p(0, T)$ and a mapping $\sum$: $[0, T]\times\mathcal{O}\to U$ such that

$$\text{(4.3)} \qquad L(t, y, \Sigma(t, y))\leqq\alpha(t) \quad \text{a.e. } t\in]0, T[,$$

$$\text{(4.4)} \qquad \|\Sigma(t, y)\|\leqq\varkappa(t) \quad \text{a.e. } t\in]0, T[$$

for all $y\in\mathcal{O}$.

Here $J_\nu^L(t, y, u)=(y_\nu, u_\nu)$ denotes the solution to equation (see e.g. [4], p. 41)

$$\text{(4.5)} \qquad \{F(y_\nu-y), \Xi(u_\nu-u)\}+\nu\,\partial L(t, y_\nu, u_\nu)\ni 0$$

where $\partial L(t)$: $E\times U\to E^*\times U^*$ is the subdifferential of $L(t)$.

We notice that condition (a) implies that $L(t, y(t), u(t))$ is a Lebesgue measurable function of t whenever $y(\cdot)$ and $u(\cdot)$ are Lebesgue measurable functions (this may be seen from formula (5.4) below). It can be shown that if spaces E and U are separable then (a) is satisfied iff L is a convex normal integrand in the sense of Rockafellar (see [18]). If L is independent of t then conditions (a) and (b) automatically hold.

As regards (c) it is satisfied in particular if the *Hamiltonian* associated with L is finite on $E\times U$ (other situations are discussed in [6, p. 217]).

An end point pair $(y_1, y_2)\in E\times E$ is called *attainable* for problem (P) is there exist functions $y\in C(0, T; E)$ and $u\in L^p(0, T; U)$ satisfying system (4.1) and such that

$$L(t, y, u)\in L^1(0, T);\qquad y(0)=y_1, y(T)=y_2. \tag{4.6}$$

The set of all attainable pairs will be denoted by C_L.

Denote also by $D(l)=\{(y_1, y_2)\in E\times E;\, l(y_1, y_2)<+\infty\}$ the effective domain of l. Our next assumption is

(B) There is $(y_1, y_2)\in C_L\cap D(l)$ such that one of the following two conditions hold

$$y_2\in \operatorname{int}\{x\in E;\, (y_1, x)\in D(l)\}, \tag{4.7}$$

$$y_2\in \operatorname{int}\{x\in E;\, (y_1, x)\in C_L\}. \tag{4.8}$$

It might be noticed that in general (4.8) fails for infinite dimensional systems because it requires the complete controllability.

The main result of this paper, Theorem 1 below may be regarded as a maximum principle for our boundary-distributed control problem.

THEOREM 1. *Suppose that all above hypotheses on system* (2.1) *and functions L, l are satisfied. Then a given pair (y_0, u_0) is optimal in problem (P) if and only if there exist the functions $p_0\in C(0, T; E^*)\cap L^{p'}(0, T; D(\Lambda^*\Pi^*))$ and $q_0\in L^1(0, T; E^*)$ which satisfy along with y_0 and u_0 the system*

$$\begin{aligned} &w_0'=Aw_0+\Lambda u_0+\Pi^{-1}f,\quad \text{on } [0, T],\\ &y_0=\Pi w_0\end{aligned} \tag{4.9}$$

$$p_0'=-A^*p_0+q_0\quad \text{on } [0, T] \tag{4.10}$$

$$(q_0(t), \Lambda^*\Pi^*p_0(t))\in\partial L(t, y_0(t), u_0(t))\quad a.e.\ t\in\,]0, T[\tag{4.11}$$

and transversality conditions

$$(p_0(0), -p_0(T))\in\partial l(y_0(0), y_0(T)). \tag{4.12}$$

Here $\partial L(t)\colon E\times U\to E^*\times U^*$ and $\partial l\colon E\times E\to E^*\times E^*$ stand for subdifferentials of $L(t)$ and l, respectively.

Of course (4.9) must be considered in the sense of (4.1'), i.e.,

$$y_0(t)=S(t)y_0(0)+\int_0^t \Pi S(t-s)\Lambda u_0(s)\,ds+\int_0^t S(t-s)f(s)\,ds,$$

while (4.10) is taken in the "mild" sense, i.e.,

$$p_0(t)=S^*(T-t)p_0(T)-\int_t^T S^*(s-t)q_0(s)\,ds,\qquad 0\leqq t\leqq T, \tag{4.13}$$

where $S^*(\cdot)$ is the semigroup generated by the adjoint A^* of A. By Λ^*, Π^* we have denoted the adjoints of Λ and Π respectively.

Some insight into the problem and Theorem can be gained from the following simple example. Minimize

$$\int_Q g(x, y(x, t))\,dx\,dt+\int_\Sigma h(u(\sigma, t))\,d\sigma\,dt \tag{4.14}$$

in $y \in C(0, T; L^2(\Omega))$ and $u \in L^p(0, T; L^2(\Omega))$ subject to

$$\frac{\partial y}{\partial t} - \Delta y = 0 \qquad \text{in } Q = \Omega \times [0, T], \tag{4.15}$$

$$y = u \qquad \text{in } \Sigma = \Gamma \times [0, T], \tag{4.16}$$

$$y(x, 0) = y^0(x) \quad x \in \Omega, \tag{4.17}$$

where $y^0 \in L^2(\Omega)$. The function $g: \Omega \times R \to R$ is continuous and convex in y, measurable in x and satisfies

$$|g(x, y)| \leqq C|y|^2 + \zeta(x) \quad \text{a.e. } x \in \Omega, \ y \in R$$

where $\zeta \in L^2(\Omega)$. As regards the function $h: R \to \bar{R}$ it will be assumed convex, lower semicontinuous and *cofinite* i.e.,

$$\lim_{|u| \to \infty} h(u)/|u| = +\infty. \tag{4.18}$$

In particular we may take function h as

$$h(u) = \begin{cases} h_0(u) & u \in U_0, \\ +\infty & \text{otherwise} \end{cases}$$

where h_0 is a continuous convex function on real axis and U_0 is a bounded and closed interval.

Clearly Assumptions (A) and (B) are satisfied where $E = L^2(\Omega)$, $U = U_2 = L^2(\Gamma)$

$$L(t, y, u) = \int_\Omega g(x, y(x))\, dx + \int_\Gamma h(u(\sigma))\, d\sigma$$

and l is defined by

$$l(y_1, y_2) = 0 \quad \text{if } y_1 = y_1^0 \quad \text{and} \quad = +\infty \quad \text{if } y_1 \neq y_1^0.$$

In this case we have also $A = \Pi = \Delta$, $D(A) = H_0^1(\Omega) \cap H^2(\Omega)$ and $D = -B$ where $B: L^2(\Gamma) \to L^2(\Omega)$ is defined by (3.4). Then as easily seen by (3.5) the adjoint B^* is given by

$$B^* y = \frac{\partial}{\partial n} (\Delta)^{-1} y \quad \text{for all } y \in L^2(\Omega)$$

so that system (4.10), (4.11) becomes

$$\frac{\partial p_0}{\partial t} + \Delta p_0 = q_0 \qquad \text{in } Q, \tag{4.19}$$

$$q_0 \in \partial_y g(x, y_0) \qquad \text{in } Q, \tag{4.20}$$

$$\frac{\partial p_0}{\partial n} \in -\partial h(u_0) \qquad \text{in } \Sigma \tag{4.21}$$

while transversality conditions (4.12) reduce to

$$y_0(x, 0) = y^0(x), \qquad p_0(x, T) = 0 \quad \text{a.e. } x \in \Omega. \tag{4.22}$$

Since condition (2.9) holds with a function $\gamma \in L^r(0, T)$ where $1 \leqq r < \frac{8}{7}$, we must choose p in the control space $L^p(0, T; L^2(\Gamma))$ where $p > 8$. Of course in the light of Remark 1, problem (4.14) can be considered over the class of all "mild" solutions

$y \in L^2(0, T; L^2(\Omega)) = L^2(Q)$ to system (4.15) ~ (4.16) and we may choose any $L^p(0, T; L^2(\Gamma)$; $1 < p < \infty$, as space of controls.

Thus by Theorem 1, a given pair (y_0, u_0) is optimal in problem (4.14) iff there exist $p_0 \in C(0, T; L^2(\Omega)) \cap L^{p'}(0, T; H_0^1(\Omega) \cap H^2(\Omega))$ and $q_0 \in L^1(0, T; L^2(\Omega))$ satisfying (4.19) ~ (4.22).

5. Proof of Theorem 1. Since the *sufficiency* is straightforward we confine ourselves to prove the *necessity* of conditions (4.9) ~ (4.12).

Let L_μ and l_μ, $\mu > 0$, denote the functions (see (1.1) and (1.3))

$$\begin{aligned} L_\mu(t, y, u) &= \inf\left\{\frac{1}{2\mu}(|y - \tilde{y}|^2 + \|u - \tilde{u}\|^2) + L(t, \tilde{y}, \tilde{u}); (\tilde{y}, \tilde{u}) \in E \times U\right\} \\ &= L(t, J_\mu^L(y, u)) + \frac{\mu}{2}\|\partial L_\mu(t, y, u)\|^2 \end{aligned} \tag{5.1}$$

$$\begin{aligned} l_\mu(y_1, y_2) &= \inf\left\{\frac{1}{2\mu}(|y_1 - \tilde{y}_1|^2 + |y_2 - \tilde{y}_2|^2) + l(\tilde{y}_1, \tilde{y}_2); (\tilde{y}_1, \tilde{y}_2) \in E \times E\right\} \\ &= l(J_\mu^l(y_1, y_2)) + \frac{\mu}{2}\|\partial l_\mu(y_1, y_2)\|^2. \end{aligned} \tag{5.2}$$

$L_\mu(t)$ and l_μ are Gâteaux differentiable on $E \times U$ and $E \times E$ and their differential $\partial L_\mu(t)$ and ∂l_μ are given by (see (1.4) and (1.5))

$$\partial L_\mu(t, y, u) = \mu^{-1}(G_1(y, u) - J_\mu^L(t, y, u)), \tag{5.3}$$

$$\partial l_\mu(y_1, y_2) = \mu^{-1}(G_2(y_1, y_2) - J_\mu^l(y_1, y_2)) \tag{5.4}$$

where $G_1 = (F, \Xi)$ and $G_2 = (F, F)$ are the duality mappings of $E \times U$ and $E \times E$, respectively. By virtue of Assumption (A) $L_\mu(t, y(t), u(t))$ is a Lebesgue measurable function of t whenever $y(\cdot)$ and $u(\cdot)$ are Lebesgue measurable.

Let (y_0, u_0) be any optimal pair of problem (P) and let $w_0 = \Pi^{-1}y_0$. Consider the approximating problem

$$\inf\left\{\int_0^T (L_\mu(t, y, u) + p^{-1}\|u - u_0\|^p)\,dt + l_\mu(y(0), y(T)) + \tfrac{1}{2}|y(0) - y_0(0)|^2\right\} \tag{5.5}$$

where the infimum is taken over all $u \in L^p(0, T; U)$ and $y \in C(0, T; E)$ satisfying (4.1'). By condition (b) in Assumption (A) and by (5.1), (5.3) we see that for all $y \in E$ and $u \in U$, we have

$$L_\mu(t, y, u) \geqq (r_0, y) + \langle s_0, u\rangle + g_0 \quad \text{a.e. } t \in]0, T[\tag{5.6}$$

where $g_0 \in L^1(0, T)$ is independent of μ. In particular it follows by (5.6) that problem (5.5) has for each $\mu > 0$ a solution (y_μ, u_μ) (unique because U is strictly convex). Remembering that $L_\mu(t)$, l_μ and the norms of U and E are Gâteaux differentiable, we infer that (y_μ, u_μ) satisfy

$$\begin{aligned} &\int_0^T ((\partial_y L_\mu(t, y_\mu, u_\mu), y) \\ &\quad + \langle \partial_u L_\mu(t, y_\mu, u_\mu) + \|u_\mu - u_0\|^{p-2}\Xi(u_\mu - u_0), u\rangle)\,dt + (\partial_1 l_\mu(y_\mu(0), y_\mu(T)) \\ &\qquad + F(y_\mu(0) - y_0(0)), y(0)) + (\partial_2 l_\mu(y_\mu(0), y_\mu(T)), y(T)) = 0 \end{aligned} \tag{5.7}$$

for all $u \in L^p(0, T; U)$ and y satisfying (4.1') with $f = 0$. We see by (5.3) and (5.6) that

$\partial_y L_\mu(t, y_\mu, u_\mu) \in L^p(0, T; E^*)$ and $\partial_u L_\mu(t, y_\mu, u_\mu) \in L^p(0, T; U^*)$. Let $p_\mu \in C(0, T; E^*)$ be defined by

$$p_\mu(t) = S^*(T-t)p_\mu(T) - \int_t^T S^*(s-t)\partial_y L_\mu(s, y_\mu(s), u_\mu(s))\, ds \tag{5.8}$$

where

$$-p_\mu(T) = \partial_2 l_\mu(y_\mu(0), y_\mu(T)). \tag{5.9}$$

($\partial_2 l$ denotes the differential relative to second argument.) We observe that by (4.2) the operator $\Lambda^* S^*(t)\Pi^*$ is continuous from E^* to $U^*$for each $t \in [0, T]$ and

$$\|\Lambda^* S^*(t)\Pi^*\|_{L(E^*, U^*)} \leqq \zeta(t) \quad \text{for } t \in [0, T].$$

Inasmuch as $S^*(t)\Pi^* = \Pi^* S^*(t)$ on $D(A^*)$ we may infer that

$$\|\Lambda^*\Pi^* S^*(t)\|_{L(E^*, U^*)} \leqq \zeta(t), \qquad t \in [0, T]. \tag{5.10}$$

In particular, it follows that $\Lambda^*\Pi^* p_\mu \in L^{p'}(0, T; U^*)$. Thus substituting y by (4.1′) in (5.7) we get after some calculations involving Fubini's theorem that

$$\Lambda^*\Pi^* p_\mu + \|u_0 - u_\mu\|^{p-2}\Xi(u_0 - u_\mu) = \partial_u L_\mu(t, y_\mu, u_\mu), \quad \text{a.e. } t \in]0, T[\tag{5.11}$$

and using (5.9) we find the transversality equations

$$\{p_\mu(0) + F(y_0(0) - y_\mu(0)), -p_\mu(T)\} = \partial l_\mu(y_\mu(0), y_\mu(T)). \tag{5.12}$$

By (5.5) we have

$$\begin{aligned} &\int_0^T (L_\mu(t, y_\mu, u_\mu) + p^{-1}\|u_\mu - u_0\|^p)\, dt + l_\mu(y_\mu(0), y_\mu(T)) + \tfrac{1}{2}|y_\mu(0) - y_0(0)|^2 \\ &\leqq \int_0^T L(t, y_0, u_0)\, dt + l(y_0(0), y_0(T)) \end{aligned} \tag{5.13}$$

because $L_\mu \leqq L$ and $l_\mu \leqq l$ for all $\mu > 0$. In particular, it follows that $\{u_\mu\}$ is bounded in $L^p(0, T; U)$ and by (4.1′) this implies that $\{y_\mu\}$ is bounded in $C(0, T; E)$. Thus extracting, a subsequence if necessary, we may assume that

$$\begin{aligned} &u_\mu \to \tilde{u} \quad \text{weakly in } L^p(0, T; U), \\ &y_\mu \to \tilde{y} \quad \text{weak-star in } L^\infty(0, T; E), \\ &y_\mu(t) \to \tilde{y}(t) \quad \text{weakly in } E \text{ for each } t \in [0, T]. \end{aligned} \tag{5.14}$$

Clearly $(\tilde{y}, \tilde{u}) \in C(0, T; E) \times L^p(0, T; U)$ satisfy (4.1′). On the other hand, we have

$$\int_0^T L(t, \tilde{y}, \tilde{u})\, dt + l(\tilde{y}(0), \tilde{y}(T)) \geqq \int_0^T L(t, y_0, u_0)\, dt + l(y_0(0), y_0(T))$$

and

$$\liminf_{\mu \to 0} \int_0^T L_\mu(t, y_\mu, u_\mu)\, dt \geqq \int_0^T L(t, \tilde{y}, \tilde{u})\, dt \tag{5.15}$$

$$\liminf_{\mu \to 0} l_\mu(y_\mu(0), y_\mu(T)) \geqq l(\tilde{y}(0), \tilde{y}(T)) \tag{5.16}$$

which in conjunction with (5.13) and (5.14) imply

$$u_\mu \to u_0 \quad \text{strongly in } L^p(0, T; U), \tag{5.17}$$

$$y_\mu \to y_0 \quad \text{strongly in } C(0, T; E). \tag{5.18}$$

The justification of inequalities (5.15), (5.16) is seen by invoking relations (5.1) ~ (5.4) and the weak lower semicontinuity of l on $E\times E$ and of convex integrand $\int_0^T L(t, y, u)\,dt$ on $L^p(0, T; E)\times L^p(0, T; U)$.

We have in mind to pass to limit in equations (5.8), (5.11) and (5.12). To this purpose some a priori estimates on p_μ are neceesary. The first is given by

LEMMA 2. *$\{p_\mu(T);\ 0<\mu\leqq 1\}$ is a bounded subset of E^*.*

Proof. Since the proof is essentially the same as that of Lemma 2 in [5] (see also [6, p. 230]) it will be outlined only.

First we assume that condition (4.7) holds i.e., there exist $y\in C(0, T; E)$ and $u\in L^p(0, T; U)$ satisfying (4.1') and such that

$$L(t, y, u)\in L^1(0, T), \qquad y(T)\in \operatorname{int}\{x\in E;\ (y(0), x)\in D(l)\}.$$

Therefore, there is $\rho>0$ and $C>0$ such that

$$l(y(0), y(T)+\rho h)\leqq C \quad \text{for all } h\in E, |h|\leqq 1.$$

Next by (5.12), we have

$$\begin{aligned}&(p_\mu(0), y_\mu(0)-y(0))\\ &\quad -(p_\mu(T), y_\mu(T)-y(T)-\rho h)+(F(y_0(0)-y_\mu(0)), y_\mu(0)-y(0))\\ &\quad \geqq l_\mu(y_\mu(0), y_\mu(T))-l_\mu(y(0), y(T)+\rho h)\end{aligned} \tag{5.19}$$

while by (4.1'), (5.8) and (5.11) we see that

$$\begin{aligned}&-(p_\mu(0), y_\mu(0)-y(0))+(p_\mu(T), y_\mu(T)-y(T))\\ &\quad \geqq \int_0^T (L_\mu(t, y_\mu, u_\mu)+p^{-1}\|u_\mu-u_0\|^p)\,dt-\int_0^T (L(t, y, u)+p^{-1}\|u-u_0\|^p)\,dt.\end{aligned}$$

Combining the latter with (5.19) we get

$$|p_\mu(T)|\leqq C \quad \text{for all } \mu>0 \tag{5.20}$$

as claimed. (In the sequel we shall denote by C several positive constants independent of μ.)

Let $\varphi: E\times E\to \bar{R}$ be the convex function defined by

$$\varphi(h_1, h_2)=\inf\left\{\int_0^T (L(t, y, u)+p^{-1}\|u\|^p)\,dt;\ y(0)=h_1, y(T)=h_2;\ (y, u) \text{ satisfies } (4.1')\right\}.$$

Clearly φ is lower semicontinuous and its effective domain is the very set C_L. If condition (4.8) holds, then there exist $y\in C(0, T; E)$ and $u\in L^p(0, T; U)$ satisfying (4.1') and such that

$$\varphi(y(0), y(T)+\rho h)\leqq C \quad \text{for all } |h|\leqq 1.$$

Then proceeding as in [5] we find that $\{|p_\mu(T)|\}$ is bounded.

We continue the proof of the theorem by noticing that by virtue of Assumption (A) there exist $\alpha \in L^p(0, T)$, $\varkappa \in L^p(0, T)$, $\rho > 0$ and $v_h : [0, T] \to U$ such that $\|v_h(t)\| \leqq \varkappa(t)$ a.e. $t \in]0, T[$ and for all $h \in E$, $|h| \leqq 1$,

$$L(t, y_0(t) + \rho h, v_h(t)) \leqq \alpha(t) \quad \text{a.e. } t \in]0, T[.$$

Next by (5.11) and definition of ∂L_μ we have,

$$\begin{aligned}
&(\partial_y L_\mu(t, y_\mu, u_\mu), y_\mu - y_0 - \rho h) \\
&\qquad + \langle \Lambda^* \Pi^* p_\mu + \Xi(u_0 - u_\mu)\|u_0 - u_\mu\|^{p-2}, u_\mu - v_h \rangle \\
&\geqq L_\mu(t, y_\mu, u_\mu) - \alpha(t), \quad \text{a.e. } t \in]0, T[.
\end{aligned}$$

Invoking (5.6) and (5.18) we find the latter yields for a sufficiently small μ,

$$\begin{aligned}
|\partial_y L_\mu(t, y_\mu, u_\mu)| \leqq C(\|u_\mu\| + \varkappa)(\|u_0 - u_\mu\|^{p-1} \\
+ \|\Lambda^* \Pi^* p_\mu\|) + \delta(t), \quad \text{a.e. } t \in]0, T[
\end{aligned} \tag{5.21}$$

where $\delta \in L^1(0, T)$. We set $q_\mu = \partial_y L_\mu(t, y_\mu, u_\mu)$. Now taking into account Lemma 1 and (5.8), (5.10), (5.21) we obtain

$$\begin{aligned}
\|\Lambda^* \Pi^* p_\mu(t)\| \leqq C\Big(\zeta(T - t) + \int_t^T \zeta(s - t)(\varkappa + \|u_\mu(s)\|), \\
(\|\Lambda^* \Pi^* p_\mu(s)\| + \|u_0(s) - u_\mu(s)\|^{p-1})\, ds + 1\Big) \quad \text{for all } t \in [0, T].
\end{aligned} \tag{5.22}$$

Next by Young's inequality we have

$$\begin{aligned}
&\left(\int_\vartheta^T \left(\int_t^T \zeta(s - t)\|u_\mu(s)\|\,\|\Lambda^* \Pi^* p_\mu(s)\|\, ds\right)^{p'} dt\right)^{1/p'} \\
&\qquad \leqq \left(\int_0^{T-\vartheta} |\zeta(t)|^{p'}\, dt\right)^{1/p'} \int_\vartheta^T \|u_\mu(s)\|\,\|\Lambda^* \Pi^* p_\mu(s)\|\, ds \\
&\qquad \leqq \eta(T - \vartheta)\left(\int_\vartheta^T \|\Lambda^* \Pi^* p_\mu(t)\|^{p'}\, dt\right)^{1/p'} \quad \text{for } 0 \leqq \vartheta \leqq T
\end{aligned}$$

where $\lim_{t \to 0} \eta(t) = 0$. We may therefore conclude from (5.22) that $\{\int_{T-\nu}^T \|\Lambda^* \Pi^* p_\mu\|^{p'}\, dt\}$ is bounded where ν is some positive constant. By (5.8) we see that $\{|p_\mu(t)|\}$ are uniformly bounded on $[T - \nu, T]$. Now reasoning as above with T replaced by $T - \nu$ we get after some steps that $\{\Lambda^* \Pi^* p_\mu\}$ is bounded in $L^{p'}(0, T; U^*)$ and

$$|p_\mu(t)| \leqq C \quad \text{for } t \in [0, T]. \tag{5.23}$$

It should be observed that (5.21) also implies that $\{q_\mu\} \subset L^1(0, T; E^*)$ is equibounded and the measures $\{\vartheta(\Theta) = \int_\Theta q_\mu(t)\, dt;\ \Theta$ measurable subset of $[0, T]\}$ are uniformly absolutely continuous. Then according to Dunford–Pettis criterion in Banach spaces (see [7]), the set $\{q_\mu\}$ is weakly compact in $L^1(0, T; E^*)$. Hence there exists a subsequence (again denoted q_μ) such that for $\mu \to 0$,

$$q_\mu \to q_0 \quad \text{weakly in } L^1(0, T; E^*). \tag{5.24}$$

Extracting further subsequences, we may also assume that

$$p_\mu(T) \to p_T \quad \text{weakly in } E^*, \tag{5.25}$$

$$p_\mu \to p_0 \quad \text{weak-star in } L^\infty(0, T; E^*), \tag{5.26}$$

$$\Lambda^* \Pi^* p_\mu \to A^* \Pi^* p_0 \quad \text{weakly in } L^{p'}(0, T; U^*). \tag{5.27}$$

It follows from (5.25) and (5.8) that for each $t \in [0, T]$,

$$p_\mu(t) \to p_0(t) = S^*(T-t)p_T - \int_t^T S^*(s-t)q_0(s)\,ds$$

in the weak topology of E^*.

Since $y_\mu(t) \to y_0(t)$ uniformly on $[0, T]$ and F is demicontinuous on E we may pass to limit in (5.12) to obtain

$$(5.28) \qquad (p_0(0), -p_0(T)) \in \partial l(y_0(0), y_0(T)).$$

The justification of the final assertion is seen by recalling that $\partial l_\mu(y_\mu(0), y_\mu(T)) \in \partial l(\mathcal{T}^l_\mu(y_\mu(0), y_\mu(T)))$ and the fact that ∂l is demiclosed in $E \times E$ (see e.g. [6, Chapter II]). To conclude the proof it remains to verify (4.11).

By (5.11) and definition of ∂L_μ we have

$$L_\mu(t, y_\mu, u_\mu) \leqq L_\mu(t, y, u) + (q_\mu, y_\mu - y) + \langle \Lambda^*\Pi^* p_\mu + \|u_0 - u_\mu\|^{p-2}\Xi(u_0 - u_\mu), u_\mu - u \rangle$$

for all $u \in L^p(0, T; U)$ and all $y \in C(0, T; E)$.

Integrating over $[0, T]$ and letting μ tend to zero we obtain

$$(5.29) \quad \int_0^T L(t, y_0, u_0)\,dt \leqq \int_0^T L(t, y, u)\,dt + \int_0^T ((q_0, y_0 - y) + \langle \Lambda^*\Pi^* p_0, u_0 - u \rangle)\,dt.$$

(Here we have used relations (5.15)~(5.18), (5.27) and the demicontinuity of the duality mapping Ξ.) By (5.29) we may conclude by a standard argument that

$$(q_0(t), \Lambda^*\Pi^* p_0(t)) \in \partial L(t, y_0(t), u_0(t)) \quad \text{a.e. } t \in]0, T[$$

thereby completing the proof.

6. Duality and boundary observation. Given functions $L(t)$ and l, define

$$(6.1) \quad M(t, p, q) = \sup\{\langle p, v\rangle + (q, y) - L(t, y, v);\ y \in E, v \in U\} \quad \text{for } p \in U^*, q \in E^*$$

and

$$(6.2) \quad m(p_1, p_2) = \sup\{(p_1, y_1) - (p_2, y_2) - l(y_1, y_2);\ y_1, y_2 \in E\} \quad \text{for } p_1, p_2 \in E^*.$$

It is well known that $M(t)$ and m are convex and lower semicontinuous on $U^* \times E^*$ and $E^* \times E^*$, respectively.

We define the dual of (P) (with $f \equiv 0$) to be the following control problem: Minimize

$$(\mathrm{P}^*) \qquad \int_0^T M(t, \Lambda^*\Pi^* p(t), v(t))\,dt + m(p(0), p(T))$$

over all $v \in L^1(0, T; E^*)$ and $p \in C(0, T; E^*) \cap L^r(0, T; D(\Lambda^*\Pi^*))$ subject to

$$(6.3) \qquad p' + A^*p = v \quad \text{on }]0, T[.$$

Here $r \in]1, +\infty[$ is a fixed number.

To make the formulation rigorous, one needs, besides Assumption (A), to assume

(A′) $M(t, q(t), v(t))$ is a measurable function of t whenever $q(t)$: $]0, T[\to U^*$ and $v(t)$: $]0, T[\to E^*$ are measurable in t. There exists $\eta \in L^1(0, T)$ such that

$$(6.4) \qquad M(t, q(t), v(t)) \geqq \eta(t) \quad \text{a.e. } t \in]0, T[.$$

Assumption (A′) implies that the cost expression (P^*) is well defined (either a real number or $+\infty$) for every pair $(q, v) \in L^r(0, T; U^*) \times L^1(0, T; E^*)$.

The general theory of duality for convex optimization problems leads us to expect that the *coextremal arc* p_0 arising in Theorem 1 is a solution to dual problem (P^*).

PROPOSITION 1. *The functions* $p_0 \in C(0, T; E^*) \cap L^r(0, T; D(\Lambda^*\Pi^*))$ *and* $q_0 \in L^1(0, T; E^*)$ *satisfy along with* y_0 *and* u_0 *the optimality system* (4.9) ~ (4.12) *if and only if* (p_0, q_0) *is an optimal pair in problem* (P^*) *and*

$$\min P^* + \min P = 0. \tag{6.5}$$

Proof. By the conjugacy formulas (6.1) and (6.2) we see that (4.11) and (4.12) are satisfied if and only if

$$\begin{aligned} L(t, y_0(t), u_0(t)) &+ M(t, \Lambda^*\Pi^* p_0(t), q_0(t)) \\ &= (y_0(t), q_0(t)) + \langle u_0(t), \Lambda^*\Pi^* p_0(t)\rangle, \end{aligned}$$

respectively

$$l(y_0(0), y_0(T)) + m(p_0(0), p_0(T)) = (y_0(0), p_0(0)) - (y_0(T), p_0(T))$$

while for arbitrary $y \in C(0, T; E)$, $u \in L^p(0, T; U)$, $p \in C(0, T; E^*) \cap L^r(0, T; D(\Lambda^*\Pi^*))$ and $v \in L^1(0, T; E^*)$; it would be true that

$$\begin{aligned} L(t, y(t), u(t)) &+ M(t, \Lambda^*\Pi^* p(t), v(t)) \\ &\geqq (v(t), y(t)) + \langle u(t), \Lambda^*\Pi^* p(t)\rangle \end{aligned} \tag{6.6}$$

respectively

$$l(y(0), y(T)) + m(p(0), p(T)) \geqq (y(0), p(0)) - (y(T), p(T)). \tag{6.7}$$

Integrating both sides of inequality (6.6) and adding (6.7) we get

$$\int_0^T L(t, y(t), u(t))\, dt + \int_0^T M(t, \Lambda^*\Pi^* p(t), v(t))\, dt \geqq 0$$

for all (y, u) and (p, v) satisfying systems (4.1) and (6.3), respectively, with equality if and only if y, u, p and v satisfy system (4.9) ~ (4.12). This completes the proof.

It should be remarked that like primal problem (P), the dual problem (P^*) involves *unbounded observation.* In fact (6.3) with the observing operator $\Lambda^*\Pi^*$ may be regarded as a distributed control system with *boundary observation.* To be more explicit let us consider the special case $B_1 \equiv 0$ and $U_1 = \{0\}$. Then the observing operator $\Lambda^*\Pi^*$ expressed as

$$\Lambda^*\Pi^* = (\sigma B - \chi_0 B)^* - B^*(A - \chi_0 I)^* \tag{6.8}$$

is defined from E^* to the boundary control space U_2^*.

Coming back to the example considered in § 4 we see that

$$M(t, p, q) = \int_\Omega g^*(x, q(x))\, dx + \int_\Gamma h^*(p(\sigma))\, d\sigma \tag{6.9}$$

and

$$m(p_1, p_2) = (p_1, y^0) \quad \text{if } p_2 = 0; \qquad m(p_1, p_2) = +\infty \quad \text{if } p_2 \neq 0. \tag{6.10}$$

Here $g^*(x, \cdot)$ and $h^*(\cdot)$ are the conjugates of $g(x, \cdot)$ and $h(\cdot)$, respectively. On the other hand, by (6.8) we see that

$$\Lambda^*\Pi^* = -\partial/\partial n.$$

Thus the dual problem to (4.14) ~ (4.17) is: Minimize

$$\int_Q g^*(x, v(x, t))\,dx\,dt + \int_\Sigma h^*(-\partial p(t, \sigma)/\partial n)\,dt\,d\sigma + \int_\Omega p(0, x)y^0(x)\,dx \tag{6.11}$$

over all $p \in C(0, T; L^2(\Omega))$ and $v \in L^1(0, T; L^2(\Omega))$ subject to

$$\begin{aligned} &\frac{\partial p}{\partial t} + \Delta p = v && \text{on } Q, \\ &p = 0 && \text{on } \Sigma, \\ &p(T, x) = 0 && \text{for } x \in \Omega. \end{aligned} \tag{6.12}$$

This is a distributed control problem with boundary observation $p \to -\partial p/\partial n$. It is well known (see e.g. [15, vol. II, p. 22]) that if $v \in L^2(Q)$ then the solution p to (6.12) belongs to $L^2(0, T; H^2(\Omega) \cap H_0^1(\Omega))$ and therefore $\partial p/\partial n \in L^2(0, T; H^{1/2}(\Gamma)) \subset L^2(\Sigma)$.

REFERENCES

[1] A. V. Balakrishnan, *Applied Functional Analysis*, Springer-Verlag, Berlin-Heidelberg-New York, 1976.

[2] ———, *Boundary control of parabolic equations*, Proceedings of Theory of Nonlinear Operators, Berlin, September 1977, Akademie Verlag, Berlin.

[3] V. Barbu, *Convex control problem of Bolza in Hilbert space*, this Journal, 13 (1975), pp. 754–771.

[4] ———, *Nonlinear Semigroups and Differential Equations in Banach Spaces*, Noordhoff International Publishing & Publishing House of Romanian Academy, Leyden-Bucharest, 1976.

[5] ———, *Constrained control problems with convex cost in Hilbert space*, J. Math. Anal. Appl., 56 (1976), pp. 502–528.

[6] V. Barbu and Th. Precupanu, *Convexity and Optimization in Banach Spaces*, Sijthoff & Noordhoff, Publishing House of Romainian Academy, 1978.

[7] J. K. Brooks and N. Dinculeanu, *Weak compactness in spaces of Bochner integrable functions and applications*, Advances in Math., 24 (1977), pp. 172–188.

[8] R. F. Curtain and A. J. Prichard, *An abstract theory for unbounded control action for distributed parameter systems*, this Journal, 15 (1977), pp. 566–611.

[9] S. Dolecki and L. Russel, *A general theory of observations and control*, this Journal, 15 (1977), pp. 185–220.

[10] L. Ekeland and R. Temam, *Analyse convexe et problèmes variationnels*, Dunode, Gauthier-Villars, 1974.

[11] H. O. Fattorini, *Boundary control systems*, this Journal, 6 (1968), pp. 349–384.

[12] I. Lasiecka, *Boundary control of parabolic systems: regularity of solutions*, Appl. Math. Optimization, 4 (1978), pp. 301-327.

[13] J. L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, Berlin-Heidelberg-New York, 1971.

[14] ———, *Interpolation linéaire et non linéaire et régularité*, Symposia Mathematica vol. VII, pp. 443–458, Academic Press, London-New York, 1971.

[15] J. L. Lions and E. Magenes, *Non-Homogeneous Boundary Value Problems and Applications*, vol. I, II, Springer-Verlag, Berlin-Heidelberg-New York, 1972.

[16] J. J. Moreau, *Fonctionnelles convexes*, Séminaire sur les équations aux dérivées partielles, Collège de France, 1966–1967.

[17] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton NJ, 1969.

[18] ———, *Convex integral functionals and duality*, Contributions to Nonlinear Functional Analysis, E. Zarantonello, ed., Academic Press, New York, 1971, pp. 215–236.

[19] K. Yosida, *Functional Analysis*, Springer-Verlag, Berlin-Heidelberg-New York, 1971.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 3, May 1980

0363-0129/80/1803-0001$01.00/0

FINITE ENERGY APPROXIMATIONS IN STOCHASTIC AND DETERMINISTIC DIFFERENTIAL GAMES*

EMMANUEL NICHOLAS BARRON†

Abstract. By penalizing the players in a differential game of fixed duration for large energy use the game will always have (i) a value (ii) a saddle point in pure strategies and (iii) a well-formulated Hamilton–Jacobi equation. Furthermore, under certain regularity conditions, the saddle points can be characterized as the solution to a system of ordinary differential equations. The case where the control functions can be any measurable function have none of the properties (i)–(iii) above and are difficult to solve. However, by considering our finite energy games we can approximate the measureable case. It also seems more realistic to have an "inertial effect" in the controls. We consider both deterministic and stochastic versions.

Introduction. The concept of a differential game was introduced by R. Isaacs in the early 1950s and was applied to the solution of various practical problems. In the 1960s, W. Fleming, A. Friedman, R. Elliott and N. Kalton, and others rigorized the formalism introduced by Isaacs to present and develop a mathematical theory of differential games, both deterministic and stochastic. These authors, furthermore, applied the theory of differential games to obtain theorems in other mathematical disciplines, particularly in the global theory of first order, semilinear partial differential equations. Later, R. Jensen obtained a very nice characterization of the asymptotic behavior of the solution of a second-order semilinear parabolic equation based on the theory of stochastic Lipschitz differential games. The theory continues to be developed by these and other authors and has led to new results in optimal stopping problems and nonlinear variational inequalities.

For applications the systematic method of constructing the value function and optimal strategies consists of solving a minimaximization problem to find the strategies and then a highly nonlinear first order Cauchy problem by the method of Cauchy characteristics. Assuming that this local solution is in fact a global solution with strong regularity then one can show that this is the value function.

To be precise, consider the differential game associated with dynamics

$$d\xi/d\tau = f(\tau, \xi, \eta, \zeta), \qquad t < \tau \leqq T, \tag{I.1}$$

$$\xi(t) = x \tag{I.2}$$

and payoff

$$J(\eta, \zeta) = \int_t^T h(s, \xi, \eta, \zeta)\, ds. \tag{I.3}$$

Here η is the maximizer, ζ is the minimizer. The control sets are Y and Z, respectively. If the Isaacs' condition

$$\begin{aligned} H^+(t, x, r) &= \min_{\zeta \in Z} \max_{\eta \in Y} \{r \cdot f(t, x, \eta, \zeta) + h(t, x, \eta, \zeta)\} \\ &= \max_{\eta \in Y} \min_{\zeta \in Z} \{r \cdot f(t, x, \eta, \zeta) + h(t, x, \eta, \zeta)\} \\ &= H^-(t, x, r) \end{aligned} \tag{I.4}$$

holds, then the game has value $V = V(t, x)$ (i.e., $V \equiv V^+ = V^-$, $V^+ =$ upper value,

* Received by the editors May 3, 1979.

† Bell Laboratories, Inc., Naperville, Illinois 60540.

V^- = lower value). If the functions f and h are sufficiently smooth, $V(t, x)$ satisfies the Hamilton–Jacobi equation

$$V_t + H^+(t, x, \nabla_x V) = 0 \tag{I.5}$$

almost everywhere, with

$$V(T, x) = 0. \tag{I.6}$$

If there are function $\eta^*(t, x, r)$ and $\zeta^*(t, x, r)$ attaining the max and min in (I.4) which are smooth and $V(t, x)$ is C^1 then the pair (η^*, ζ^*) is a saddle point in pure strategies. These are obviously very stringent conditions. However, the hardest part of the above for applications is (i) verifying (I.4) holds; (ii) finding η^*, ζ^* and (iii) solving (I.5), (I.6). When f and h are linear in η, ζ the problem becomes much simpler but then the model is often less credible.

In an attempt to simplify the problem above we introduced in [1], [2] the restriction of forcing the players to choose Lipschitz control functions starting at fixed control positions y and z. We showed that in this case Lipschitz value will *always* exist even in the absence of (I.4). The Hamilton–Jacobi equation for this function is much simplified but still very difficult to solve: if $V^{M,L}(t, x, y, z)$ is the value then it satisfies a.e.

$$V_t^{M,L} + V_x^{M,L} \cdot f(t, x, y, z) + M|V_y^{M,L}| - L|V_z^{M,L}| + h(t, x, y, z) = 0. \tag{I.7}$$

Furthermore an approximate saddle point in pure strategies will exist. This is an obvious improvement over the measurable case but still not entirely satisfactory.

In this paper we will restrict the controls in a different manner. Here they are required to be absolutely continuous with first derivative square integrable. Thus only the energy is required to be finite. Without a fixed bound on the energy this results in identical results as in the measurable case. However if we also force the controls to start at fixed positions y and z and further penalize the players for large energy uses we arrive at new, and simpler results. We penalize by taking the new payoff.

$$\begin{aligned} P_{M,L}(\eta, \zeta) = &\int_t^T h(s, \xi, \eta, \zeta)\, ds \\ &+ \frac{1}{L}\int_t^T \dot{\zeta}^2(s)\, ds - \frac{1}{M}\int_t^T \dot{\eta}^2(s)\, ds \end{aligned} \tag{I.8}$$

for M, $L > 0$. Call this new game a Sobolev differential game. Then we prove the following:

(i) The value of a Sobolev game *always* exists even in the absence of the Isaacs condition. Denote it by $W^{M,L}(t, x, y, z)$, indicating the dependence on the initial time t, state x and control positions y and z.

(ii) $W^{M,L}$ satisfies a.e.

$$W_t^{M,L} + W_x^{M,L} \cdot f + M/4|W_y^{M,L}|^2 - L/4|W_z^{M,L}|^2 + h = 0. \tag{I.9}$$

(iii) If $W^{M,L}$ is C^2 the functions η^* and ζ^* solving

$$\begin{aligned} &d\xi/d\tau = f(\tau, \xi, \eta, \zeta), \qquad t < \tau \leqq T, \\ &d\eta/d\tau = M/2\, W_y^{M,L}(\tau, \xi, \eta, \zeta), \\ &d\zeta/d\tau = -L/2\, W_z^{M,L}(\tau, \xi, \eta, \zeta), \\ &\xi(t) = x, \qquad \eta(t) = y, \qquad \zeta(t) = z \end{aligned} \tag{I.10}$$

form a saddle point in pure strategies.

Furthermore, we study the corresponding Sobolev game when the dynamics are stochastic differential equations and the control functions are perturbed by a Brownian motion. We show that this game has a value and that a saddle point in pure strategies always exists.

Finally, to relate Sobolev games with measureable games we consider the behavior of $W^{M,L}$ as $M, L \to \infty$; that is the penalty terms in (I.8) goes to zero. Then using the results of Jensen [9] we show that

$$\lim_{L\to\infty} \lim_{M\to\infty} W^{M,L} = V^+, \qquad \lim_{M\to\infty} \lim_{L\to\infty} W^{M,L} = V^-,$$

a result analogous to that obtained in Barron [1] for Lipschitz games.

1. Sobolev games—definitions and elementary properties. For any positive integer m, R^m will denote Euclidean m-space and $|\cdot|$ the norm in R^m. Let $I \equiv [0, 1]$ and I^m denote the cartesian product of I with itself m times.

We are given functions $f: [0, T] \times R^m \times I^p \times I^q \to R^m$ and $h: [0, T] \times R^m \times I^p \times I^q \to R^1$. Throughout this paper the following assumptions will hold regarding the functions f and h:

(A) f and h are bounded and uniformly Lipschitz continuous on $[0, T] \times R^m \times I^p \times I^q$.

Consider the system of m ordinary differential equations for the function $\xi(\cdot)$

$$d\xi/d\tau = f(\tau, \xi(\tau), \eta, \zeta) \qquad (0 \leqq t < \tau \leqq T) \tag{1.1}$$

with initial conditions

$$\xi(t) = x, \qquad x \in R^m. \tag{1.2}$$

Substituting any pair of measurable functions $\eta = \eta(\tau)$ and $\zeta = \zeta(\tau)$ on $0 \leqq t \leqq \tau \leqq T$ with values a.e. in I^p and I^q, respectively, yields a unique, absolutely continuous $\xi(\tau)$ on $[t, T]$—called the *trajectory* corresponding to $\eta(\tau)$, $\zeta(\tau)$—which satisfies (1.1) almost everywhere.

Any pair of functions $\eta(\cdot)$, $\zeta(\cdot)$ which yields a unique trajectory will be called *control functions.*

Let J be any subinterval of $[0, T]$, $J_t \equiv [t, T]$ and $A \subseteq R^p$, $B \subseteq R^q$ be any sets. Define the following classes of functions

$$Y(J; A) = \{\eta(\cdot): J \to A | \eta(\tau) = (\eta_1(\tau), \cdots, \eta_p(\tau)),$$
$$\eta_i(\tau) \text{ is absolutely continuous on } J, \int_J \eta_i^2(\tau)\, d\tau < \infty, 1 \leqq i \leqq p\}$$

and

$$Z(J; B) = \{\xi(\cdot): J \to B | \zeta(\tau) = (\zeta_1(\tau), \cdots, \zeta_q(\tau)),$$
$$\zeta_j(\tau) \text{ is absolutely continuous on } J, \int_J \dot{\zeta}_j^2(\tau)\, d\tau < \infty, 1 \leqq j \leqq q\},$$

where "$\cdot$" denotes $d/d\tau$.

For fixed $t \in [0, T]$, $y \in A$, $z \in B$, let

$$Y_y(J_t; A) = \{\eta(\cdot) \in Y(J_t; A) | \eta(t) = y\}$$

and

$$Z_z(J_t; B) = \{\zeta(\cdot) \in Z(J_t; B) | \zeta(t) = z\}.$$

Given any functions $\eta(\cdot) \in Y(J_t; I^p)$ and $\zeta(\cdot) \in Z(J_t; I^q)$ let $\xi(\cdot)$ on $[t, T]$ denote the corresponding trajectory. For given positive constants M and L let $P(\eta, \zeta)$ denote the corresponding *payoff*:

$$(1.3) \qquad \begin{aligned} P(\eta, \zeta) = &\int_t^T h(s, \xi(s), \eta(s), \zeta(s))\, ds \\ &+ \frac{1}{L}\int_t^T |\dot{\zeta}(s)|^2\, ds - \frac{1}{M}\int_t^T |\dot{\eta}(s)|^2\, ds. \end{aligned}$$

Using the function classes $Y_y(J_t; A)$ and $Z_z(J_t; B)$ we can define a *differential game* associated with (1.1)–(1.3) in a manner similar to that in Friedman [4], [5] for measurable games and Barron [1], [2] for Lipschitz games. We will omit the precise details and simply refer the reader to the above references for clarifications.

To set the notation, for the partition of $J_t = [t, T]$ into n subintervals of equal length $\delta = n^{-1}(T - t)$ we let Γ^δ, Γ_δ and Δ^δ, Δ_δ denote the upper and lower δ-strategies for the functions η and ζ, respectively. Thus, for example, if $(\zeta_\delta, \eta^\delta)$ denotes the *outcome* of a pair $(\Delta_\delta, \Gamma^\delta)$ then $\zeta_\delta(\cdot) \in Z_z(J_t; B)$ and $\eta^\delta(\cdot) \in Y_y(J_t; A)$.

DEFINITION. The δ-strategies Γ^δ, Γ_δ for η and Δ^δ, Δ_δ for ζ mapping into $Y_y(J_t; A)$ and $Z_z(J_t; B)$, respectively, will be called Sobolev δ-strategies. Write Γ^δ or $\Gamma_\delta \in Y_y(J_t; A)$ and Δ^δ or $\Delta_\delta \in Z_z(J_t; B)$. A δ-game in which a player chooses a Sobolev δ-strategy will be called a *Sobolev δ-game* or a *finite energy* δ-game.

With $A = I^p$ and $B = I^q$ define the Sobolev upper and lower δ-values by

$$W^\delta_{M,L}(t, x, y, z) \equiv \inf_{\Delta_\delta} \sup_{\Gamma^\delta} P[\Delta_\delta, \Gamma^\delta]$$

and

$$W_{\delta,M,L}(t, x, y, z) \equiv \sup_{\Gamma_\delta} \inf_{\Delta^\delta} P[\Gamma_\delta, \Delta^\delta],$$

respectively. We have indicated the dependence of the δ-values on the initial time $t \in [0, T]$, the initial state $x \in R^m$ and the initial control positions $y \in I^p$ and $z \in I^q$.

Remarks. a) The defining term "Sobolev" is used to point up the fact that we are taking our control functions out of the well-known Sobolev space $W^{1,2}$ of functions with first derivative square integrable. In a "measurable" differential game the controls are allowed to be any measurable functions. Thus in a Sobolev game we consider a much smaller class of functions.

b) It follows from Gronwall's inequality that the set of all trajectories $\xi(\tau)$ determined by control functions from Y_y and Z_z is a uniformly bounded set.

In what follows we shall need to extend the Sobolev δ-values defined above on $[0, T] \times R^m \times I^p \times I^q$ to the entire strip $[0, T] \times R^m \times R^p \times R^q$. We do this as follows.

Define $F(t, x, y, z)$ and $H(t, x, y, z)$ on $[0, T] \times R^m \times R^p \times R^q$ as the even, periodic (period 2) extensions of the functions $f(t, x, \cdot, \cdot)$ and $h(t, x, \cdot, \cdot)$, respectively, in the y and z coordinates. Thus, F and H are of period 2 in y and z separately and, for any $y \in R^p$, $z \in R^q$

$$g(t, x, y, z) = g(t, x, \pm y, \pm z),$$

$$g(t, x, y, z) = g(t, x, y, 2 - z)$$

and

$$g(t, x, y, z) = g(t, x, 2-y, z)$$

holds for both $g = F$ and $g = H$. Obviously, F and H are uniformly Lipschitz continuous in (t, x, y, z) and are bounded with the same constants as those for f and h. Furthermore $F(t, x, y, z) = f(t, x, y, z)$ and $H(t, x, y, z) = h(t, x, y, z)$ on $[0, T]\times R^m \times I^p \times I^q$.

Given a pair of Sobolev δ-strategies $(\Delta_\delta, \Gamma^\delta)$ with $\Delta_\delta \in Z_z(J_t; R^q)$ and $\Gamma^\delta \in Y_y(J_t; R^p)$ we define the extended Sobolev upper δ-value associated with the dynamics (1.1)–(1.2) with f replaced by F and payoff (1.3) with h replaced by H and denote this extended δ-value by $\bar{W}^\delta_{M,L}(t, x, y, z)$. Note that the outcome functions $(\zeta_\delta, \eta^\delta)$ associated with $(\Delta_\delta, \Gamma^\delta)$ satisfy $\zeta_\delta(\cdot) \in Z_z(J_t; R^p)$ and $\eta^\delta(\cdot) \in Y_y(J_t; R^q)$.

Denote by $\bar{P}(\eta, \zeta)$, the payoff (1.3) with h replaced by H. $\bar{P}(\eta, \zeta) = P(\eta, \zeta)$ when $\eta(\tau) \in I^p$ and $\zeta(\tau) \in I^q$ for all $\tau \in J_t$.

Similarly we define the extended Sobolev lower δ-value associated with (1.1)–(1.3) (with F and H) and denote it by $\bar{W}_{\delta,M,L}(t, x, y, z)$. Here a pair $(\Gamma_\delta, \Delta^\delta)$ satisfies $\Gamma_\delta \in Y_y(J_t; R^p)$ and $\Delta^\delta \in Z_z(J_t; R^q)$.

Notation. In the remainder of this paper we write $Y_y \equiv Y_y(J_t; R^p)$ $(y \in R^p)$ and $Y^p_y \equiv Y_y(J_t; I^p)$ $(y \in I^p)$. Similarly $Z_z \equiv Z_z(J_t; R^q)$ $(z \in R^q)$ and $Z^q_z \equiv Z_z(J_t; I^q)$ $(z \in I^q)$.

To show that $\bar{W}^\delta_{M,L}$ and $\bar{W}_{\delta,M,L}$ actually extend $W^\delta_{M,L}$ and $W_{\delta,M,L}$ we present the following

LEMMA 1.1. *For any* $t \in [0, T]$, $x \in R^m$, $y \in I^p$, $z \in I^q$ *and all* $\delta > 0$

(i) $\bar{W}^\delta_{M,L}(t, x, y, z) = W^\delta_{M,L}(t, x, y, z)$;

(ii) $\bar{W}_{\delta,M,L}(t, x, y, z) = W_{\delta,M,L}(t, x, y, z)$.

Proof. We only prove (i); (ii) is similar.

For $y \in I^p$ and $z \in I^q$ we will show that

$$\begin{aligned} \bar{W}^\delta_{M,L}(t, x, y, z) &\equiv \inf_{\Delta_\delta \in Z_z} \sup_{\Gamma^\delta \in Y_y} \bar{P}[\Delta_\delta, \Gamma^\delta] \\ &= \inf_{\Delta_\delta \in Z^q_z} \sup_{\Gamma^\delta \in Y^p_y} P[\Delta_\delta, \Gamma^\delta] \qquad (1.4) \\ &\equiv W^\delta_{M,L}(t, x, y, z) \end{aligned}$$

for all $t \in [0, T]$, $x \in R^m$. The difficulty occurs in the fact that the control functions determined by a pair $(\Delta_\delta, \Gamma^\delta)$ may leave the cubes I^p and I^q. However, even should the controls leave these cubes we can achieve a similar effect through using controls which remain in the cubes. To see this we will take for simplicity $p = q = 1$.

By considering classes of control functions we clearly have

$$(1.5) \qquad \inf_{\Delta_\delta \in Z_z} \sup_{\Gamma^\delta \in Y^1_y} \bar{P}[\Delta_\delta, \Gamma^\delta] \leqq \bar{W}^\delta_{M,L} \leqq \inf_{\Delta_\delta \in Z^1_z} \sup_{\Gamma^\delta \in Y_y} \bar{P}[\Delta_\delta, \Gamma^\delta].$$

Given $\varepsilon > 0$, by (1.5) there is an extended Sobolev upper δ-strategy $\tilde{\Gamma}^\delta \in Y_y$ for the player η so that

$$(1.6) \qquad \bar{W}^\delta_{M,L}(t, x, y, z) \leqq \bar{P}[\zeta, \tilde{\Gamma}^\delta(\zeta)] + \varepsilon$$

for every control $\zeta(\cdot) \in Z^1_z$.

For a given $\zeta(\cdot) \in Z^1_z$ define the Sobolev upper δ-strategy ${}^*\Gamma^\delta \in Y^1_y$ as follows:

Let $\tilde{\eta}^\delta(\tau) \equiv \tilde{\Gamma}^\delta(\zeta)(\tau)$, $t \leqq \tau \leqq T$. Then, since $\tilde{\Gamma}^\delta \in Y_y$ is an extended Sobolev δ-strategy, $\tilde{\eta}^\delta(t) = y$ and $\tilde{\eta}^\delta(\cdot) \in Y_y \equiv Y_y(J_t; R^1)$. If $\tilde{\eta}^\delta(\tau) \in I^1 = [0, 1]$ for all $\tau \in J_t$ define ${}^*\Gamma^\delta(\zeta)(\tau) \equiv \tilde{\eta}^\delta(\tau)$ and we are done. If $\tilde{\eta}^\delta(\tau)$ leaves $[0, 1]$ at some time $\hat{\tau}$, $t < \hat{\tau} < T$ then we reflect back into $[0, 1]$ with respect to the line of crossing (either $y = 0$ or $y = 1$). If this reflected function then again crosses $y = 0$ or $y = 1$ we reflect back into $[0, 1]$ again. Continue this process up to time T. See Fig. 1. Denote the function just defined by $*\eta^\delta(\tau)$. Then by the definition, $*\eta^\delta(t) = y$ (since $y \in [0, 1]$) and $*\eta^\delta(\tau) \in I^1 = [0, 1]$ for all $\tau \in J_t$.

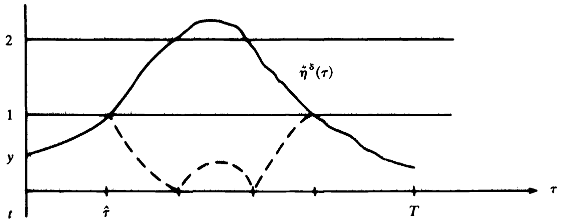


FIG. 1

Furthermore, for every τ at which $\tilde{\eta}^\delta(\cdot)$ is differentiable other than those τ's at which $\tilde{\eta}^\delta(\tau)$ leaves an interval $[n, n+1]$, we have, by definition that $*\eta^\delta(\cdot)$ will also be differentiable at these τ's and $*\dot{\eta}^\delta(\tau) = \pm\dot{\tilde{\eta}}^\delta(\tau)$. The + sign will hold on the even numbered reflections or when $\tilde{\eta}^\delta(\tau) \in [0, 1]$ and the − sign will hold on the odd numbered reflections.

Since $\tilde{\eta}^\delta(\tau)$ is an absolutely continuous function, $*\eta^\delta(\tau)$ will cross the boundary of $[0, 1]$ at most a countable number of times, namely the reflection times. At these times $*\eta^\delta$ may not have a derivative even though $\tilde{\eta}^\delta$ may. Hence, altogether $*\eta^\delta$ will not have a derivative on a subset E of $[0, \mathrm{T}]$ of Lebesgue measure zero. The set E consists of the reflection times and the times at which $\tilde{\eta}^\delta$ is not differentiable.

Taking these properties of $*\eta^\delta$ into account, it follows from the fact that $\tilde{\eta}^\delta \in Y_y$ that

$$(1.7)\qquad \int_{J_t} *\dot{\eta}^\delta(s)^2\,ds = \int_{J_t\setminus E} *\dot{\eta}^\delta(s)^2\,ds = \int_{J_t} \dot{\tilde{\eta}}^\delta(s)^2\,ds < +\infty$$

and so $*\eta^\delta \in Y_y^1$.

Define $*\Gamma^\delta(\zeta)(\tau) \equiv *\eta^\delta(\tau)$, $\tau \in J_t$.

Since F and H are even 2-periodic functions in y, we have from the definition of $*\eta^\delta$ that

$$(1.8)\qquad F(\tau, \cdot, \tilde{\eta}^\delta(\tau), \zeta(\tau)) = f(\tau, \cdot, *\eta^\delta(\tau), \zeta(\tau))$$

and

$$(1.9)\qquad H(\tau, \cdot, \tilde{\eta}^\delta(\tau), \zeta(\tau)) = h(\tau, \cdot, *\eta^\delta(\tau), \zeta(\tau))$$

for $\tau \in J_t$ and any $\zeta(\tau) \in Z_z^1$.

Let $\tilde{\xi}^\delta$ be the trajectory corresponding to $(\tilde{\eta}^\delta, \zeta)$, that is, the solution of (1.1), (1.2) with $\eta = \tilde{\eta}^\delta$ and f replaced by F. Let $*\xi^\delta$ be the solution of (1.1), (1.2) corresponding to $(*\eta^\delta, \zeta)$. Then, it follows from (1.8) and the uniqueness of solutions that $*\xi^\delta(\tau) = \tilde{\xi}^\delta(\tau)$ for all $\tau \in J_t$. Furthermore, from (1.7) and (1.9) we have that

$$(1.10)\qquad P(*\eta^\delta, \zeta) = P[\zeta, *\Gamma^\delta(\zeta)] = \bar{P}[\zeta, \tilde{\Gamma}^\delta(\zeta)] = \bar{P}(\tilde{\eta}^\delta, \zeta).$$

Using (1.10) in (1.6) and taking $\inf_{\Delta_\delta \in Z_z^1} \sup_{\Gamma^\delta \in Y_y^1}$ gives that

$$\bar{W}^\delta_{M,L}(t, x, y, z) \leqq \inf_{\Delta_\delta \in Z_z^1} \sup_{\Gamma^\delta \in Y_y^1} P[\Delta_\delta, \Gamma^\delta] \equiv W^\delta_{M,L}(t, x, y, z)$$

by the fact that ε was arbitrary. The opposite inequality is proved similarly using the first inequality in (1.5). This completes the proof.

LEMMA 1.2. *As a function of y and z, $\bar{W}^{\delta}_{M,L}(t, x, y, z)$ and $\bar{W}_{\delta,M,L}(t, x, y, z)$ are even periodic functions (in y and z separately) of period* 2.

Proof. Take $p = q = 1$. Given any $y \in R^1$ and a control function $\eta(\tau) \in Y_y = Y_y(J_t; R^1)$, define $\hat{\eta}(\tau) \in Y_{(-y)}(J_t; R^1)$ as the reflection of $\eta(\tau)$ about $y = 0$. Then, clearly, $\dot{\hat{\eta}}(\tau) = -\dot{\eta}(\tau)$ for almost every $\tau \in J_t$. Hence, using the evenness of F and H and the fact that $\int_{J_t} \dot{\hat{\eta}}^2(\tau)\,d\tau = \int_{J_t} \dot{\eta}^2(\tau)\,d\tau$ it is easy to see that $\bar{W}^{\delta}_{M,L}(t, x, y, z) = \bar{W}^{\delta}_{M,L}(t, x, -y, z)$; that is, $\bar{W}^{\delta}_{M,L}$ is an even function in y. The remaining definitions of even 2-periodic in y are also easily seen to hold and we leave these to the reader. Similarly, $\bar{W}^{\delta}_{M,L}$ is an even 2-periodic function of z.

Remark. It follows from Lemma 1.2 that the study of $\bar{W}^{\delta}_{M,L}(t, x, y, z)$ and $\bar{W}_{\delta,M,L}(t, x, y, z)$ may be restricted to the study of these functions for $y \in I^p$ and $z \in I^q$.

2. The equivalent L^2-differential game. In this section we formulate a differential game in which the control functions are allowed to be any L^2 functions. This game will be shown to be equivalent to the Sobolev game presented in § 1; however, it is much easier to work with and makes the theorems more readily apparent.

Denote by $L^2(A)$ the Lebesgue space of square integrable functions on J_t with values in the subset A of some Euclidean space.

Consider the system of $m + p + q$ ordinary differential equations for the functions $\xi(\tau)$, $\eta(\tau)$ and $\zeta(\tau)$ on J_t:

$$d\xi/d\tau = F(\tau, \xi(\tau), \eta(\tau), \zeta(\tau)), \tag{2.1}$$

$$d\eta/d\tau = u(\tau), \tag{2.2}$$

$$d\zeta/d\tau = v(\tau) \tag{2.3}$$

with initial conditions

$$\xi(t) = x \in R^m, \qquad \eta(t) = y \in R^p, \qquad \zeta(t) = z \in R^q. \tag{2.4}$$

Here $u(\tau) = (u_1(\tau), \cdots, u_p(\tau))$ and $v(\tau) = (v_1(\tau), \cdots, v_q(\tau))$ are measurable functions with $u \in L^2(R^p)$, $v \in L^2(R^q)$.

Consider the L^2-differential game as defined in Friedman [4, Chap. 7] associated with the dynamics (2.1)–(2.4) and the payoff

$$\begin{aligned} K(u, v) = &\int_t^T H(s, \xi(s), \eta(s), \zeta(s))\,ds \\ &+ \frac{1}{L}\int_t^T |v(s)|^2\,ds - \frac{1}{M}\int_t^T |u(s)|^2\,ds. \end{aligned} \tag{2.5}$$

The players in this game are u and v, with u the maximizer of K and v the minimizer of K; u is any function in $L^2(R^p)$ and v is any function in $L^2(R^q)$. Note that for any pair (u, v), there is a unique solution (ξ, η, ζ) of (2.1)–(2.4) and hence K is well-defined.

Denote by $K^{\delta}(t, x, y, z)$ and $K_{\delta}(t, x, y, z)$ the upper and lower δ-values, respectively, of the L^2 game defined above and denote the upper and lower δ-strategies for the players u and v by Φ^{δ}, Φ_{δ} and Ψ^{δ}, Ψ_{δ}, respectively.

The following lemma is readily seen.

LEMMA 2.1. *$K^{\delta}(t, x, y, z) \equiv \bar{W}^{\delta}_{M,L}(t, x, y, z)$ and $K_{\delta}(t, x, y, z) \equiv \bar{W}_{\delta,M,L}(t, x, y, z)$ for any $t \in [0, T]$, $x \in R^m$, $y \in R^p$ and $z \in R^q$, for all $\delta = n^{-1}(T - t) > 0$.*

We omit the proof of Lemma 2.1 since it is similar to the proof of Lemma 2.1 in Barron [2]. The lemma establishes the equivalence of the L^2 and Sobolev differential games.

Our next lemma shows that in computing K^δ and K_δ we can restrict ourselves to looking at δ-strategies which will keep the trajectoris η and ζ in the cubes I^p and I^q if the initial trajectory positions are in these cubes. To be precise we make the following

DEFINITION. We define the following classes of L^2 δ-strategies.

$$\Lambda_\delta = \{\Psi_\delta \text{ for } v;\ \forall\Phi^\delta \text{ for } u, \text{ if } (v_\delta, u^\delta) \text{ is the outcome of } (\Psi_\delta, \Phi^\delta) \text{ and } (\xi, \eta, \zeta) \text{ the corresponding solution of (2.1)–(2.4), then } \zeta(\tau)\in I^q,\ \tau\in J_t\},$$

$$\Sigma^\delta = \{\Phi^\delta \text{ for } u;\ \forall\Psi_\delta \text{ for } v, \text{ if } (v_\delta, u^\delta) \text{ is the outcome of } (\Psi_\delta, \Phi^\delta) \text{ and } (\xi, \eta, \zeta) \text{ the corresponding solution of (2.1)–(2.4), then } \eta(\tau)\in I^p,\ \tau\in J_t\}.$$

Similarly we define Λ^δ (for Ψ^δ) and Σ_δ (for Φ_δ). Note that, for example $\Lambda_\delta = \varnothing$ if $z\notin I^q$.

LEMMA 2.2. *For* $t\in[0, T]$, $x\in R^m$, $y\in I^p$, $z\in I^q$ *and all* $\delta>0$

(i) $K^\delta(t, x, y, z) = \inf_{\Psi_\delta\in\Lambda_\delta} \sup_{\Phi^\delta\in\Sigma^\delta} K[\Psi_\delta, \Phi^\delta]$;

(ii) $K_\delta(t, x, y, z) = \sup_{\Phi_\delta\in\Sigma_\delta} \inf_{\Psi^\delta\in\Lambda^\delta} K[\Phi_\delta, \Psi^\delta]$,

Proof. We will only prove (i).

Since Λ_δ is a proper subset of the class of all Ψ_δ strategies for v we have

$$K^\delta(t, x, y, z) \leqq \inf_{\Psi_\delta\in\Lambda_\delta}\ \sup_{,\Phi^\delta} K[\Psi_\delta, \Phi^\delta]. \tag{2.6}$$

Given $\varepsilon>0$ there is an upper δ-strategy $\hat\Phi^\delta$ for u such that

$$K^\delta \leqq K[\Psi_\delta, \hat\Phi^\delta] + \varepsilon \tag{2.7}$$

for every $\Psi_\delta\in\Lambda_\delta$.

Given $\Psi_\delta\in\Lambda_\delta$, let $(v_\delta, \hat u^\delta)$ denote the outcome of $(\Psi_\delta, \hat\Phi^\delta)$ and let $(\xi^\delta, \hat\eta^\delta, \zeta_\delta)$ denote the corresponding solution of (2.1)–(2.4). As shown in the proof of Lemma 1.2, given $\hat\eta^\delta(\cdot)$, there is a function which can be constructed by reflection, call it $\tilde\eta^\delta(\cdot)$ on J_t so that $\tilde\eta^\delta(\tau)\in I^p$ for all $\tau\in J_t$ and $\dot{\tilde\eta}^\delta(\tau) = \pm\dot{\hat\eta}^\delta(\tau)$ for almost every $\tau\in J_t$. By the properties of F and H in the y and z variables we have

$$\begin{aligned} &F(\tau, \cdot, \tilde\eta^\delta(\tau), \zeta_\delta(\tau)) = F(\tau, \cdot, \hat\eta^\delta(\tau), \zeta_\delta(\tau)) \quad \text{and} \\ &H(\tau, \cdot, \tilde\eta^\delta(\tau), \zeta_\delta(\tau)) = H(\tau, \cdot, \hat\eta^\delta(\tau), \zeta_\delta(\tau)). \end{aligned} \tag{2.8}$$

Define $\tilde u^\delta(t) = 0$ and $\tilde u^\delta(\tau) \equiv \dot{\tilde\eta}^\delta(\tau)$, $t<\tau\leqq T$. Set $\tilde\Phi^\delta(v_\delta)(\tau) \equiv \tilde u^\delta(\tau)$. Since $\hat\Phi^\delta$ is an L^2 δ-strategy, $\tilde u^\delta(\cdot)\in L^2(R^p)$. Also, by the definition of $\tilde\Phi^\delta$ we have

$$\tilde\Phi^\delta \in \{\Phi^\delta;\ \text{if } (v_\delta, u^\delta) \text{ is the outcome of } (\Psi_\delta, \Phi^\delta) \text{ and } \Psi_\delta\in\Lambda_\delta \text{ then the } \eta \text{ solution of (2.1)–(2.4) has image in } I^p\} \equiv A^\delta.$$

Since $A^\delta \subseteq \Sigma^\delta$, we have $\tilde\Phi^\delta\in\Sigma^\delta$.

It follows from (2.8) and the fact $\dot{\tilde\eta}^\delta(\tau) = \pm\dot{\hat\eta}^\delta(\tau)$ a.e. that if $(\tilde\xi^\delta, \tilde\eta^\delta, \zeta_\delta)$ denotes the solution of (2.1)–(2.4) corresponding to $(\Psi_\delta, \tilde\Phi^\delta)$ then $\tilde\xi^\delta \equiv \xi^\delta$ and as a consequence $K[\Psi_\delta, \tilde\Phi^\delta] = K[\Psi_\delta, \hat\Phi^\delta]$.

From (2.6) and (2.7) we then have

$$\begin{aligned} K^\delta(t, x, y, z) &\leqq \inf_{\Psi_\delta\in\Lambda_\delta}\ \sup_{\Phi^\delta\in A^\delta} K[\Psi_\delta, \Phi^\delta] \\ &\leqq \inf_{\Psi_\delta\in\Lambda_\delta}\ \sup_{\Phi^\delta\in\Sigma^\delta} K[\Psi_\delta, \Phi^\delta]. \end{aligned}$$

Similarly we prove the reverse inequality. This completes the proof.

To apply the theorems of Friedman [4], [5] we need bounded control sets. Thus we introduce another class of games. Let r and s be given positive integers. Let

$$\mathcal{U}^r=\{u\in R^p;\,|u|\leqq r\}\quad\text{and}\quad \mathcal{V}^s=\{v\in R^q;\,|v|\leqq s\}\tag{2.9}$$

The sets $\mathcal{U}^r$ and $\mathcal{V}^s$ will be called *control sets.*

Consider the differential game associated with dynamics (2.1)–(2.4) and payoff (2.5) when the control functions u and v are allowed to be any measurable functions with values almost everywhere in $\mathcal{U}^r$ and $\mathcal{V}^s$, respectively. Denote the upper and lower δ-values for this game by $K^{\delta}_{r,s}(t,x,y,z)$ and $K^{r,s}_{\delta}(t,x,y,z)$ respectively.

Since the control functions u and v appear linearly in the dynamics and separated in the payoff, it follows from Friedman [4, Thm. 2.3.1] that this game has a value which we will denote by $K^{r,s}(t,x,y,z)$. That is,

$$\begin{aligned}K^{r,s}(t,x,y,z)&\equiv K^{+}_{r,s}(t,x,y,z)\equiv\lim_{\delta\to0}K^{\delta}_{r,s}(t,x,y,z)\\&=\lim_{\delta\to0}K^{r,s}_{\delta}(t,x,y,z)\equiv K^{-}_{r,s}(t,x,y,z).\end{aligned}\tag{2.10}$$

Furthermore, by Friedman [4, Thms. 2.6.3, 4.2.1] we have that $K^{r,s}(t,x,y,z)$ is uniformly Lipschitz continuous in (t,x,y,z), has a total derivative at almost all (t,x,y,z) and satisfies the Hamilton–Jacobi–Isaacs equation at points of differentiability:

$$\begin{aligned}\partial K^{r,s}/\partial t+\min_{v\in\mathcal{V}^s}\max_{u\in\mathcal{U}^r}\Big\{&\nabla_xK^{r,s}\cdot F(t,x,y,z)+\nabla_yK^{r,s}\cdot u+\nabla_zK^{r,s}\cdot v\\&+H(t,x,y,z)+\frac{1}{L}|v|^2-\frac{1}{M}|u|^2\Big\}=0\end{aligned}\tag{2.11}$$

and

$$K^{r,s}(T,x,y,z)=0.\tag{2.12}$$

3. Sobolev differential games associated with Ito differential equations. In this section we develop the stochastic counterpart of the differential games introduced in §§ 1 and 2. The utility in doing so, apart from interest in white noise perturbed control systems, allows us to apply powerful partial differential equations techniques to our games.

Consider the system of $m+p+q$ stochastic Ito differential equations for the processes $(X(\tau),Y(\tau),Z(\tau))$ on $0\leqq t<\tau\leqq T$

$$dX=F(\tau,X(\tau),Y(\tau),Z(\tau))\,d\tau+\varepsilon\,dw_1(\tau),\tag{3.1}$$

$$dY=u(\tau)\,d\tau+\varepsilon\,dw_2(\tau),\tag{3.2}$$

$$dZ=v(\tau)\,d\tau+\varepsilon\,dw_3(\tau)\tag{3.3}$$

with deterministic initial conditions

$$X(t)=x\in R^m,\qquad Y(t)=y\in R^p,\qquad Z(t)=z\in R^q.\tag{3.4}$$

Here $\varepsilon>0$ is given and $w(\tau)=(w_1(\tau),w_2(\tau),w_3(\tau))$ is a standard $m+p+q$-dimensional Brownian motion on $[t,T]$. For each $u(\tau)\in L^2(R^p)$ and $v(\tau)\in L^2(R^q)$ the form of (3.1)–(3.4) with our assumptions on F implies by a minor modification of the arguments in Gikhman and Skorohod [8] (see also Fleming and Rishel [3, Thms. 4.1 and 10.3]) that there is a unique solution (X,Y,Z) of (3.1)–(3.4) with probability one.

Corresponding to a pair $u(\tau)\in L^2$ and $v(\tau)\in L^2$ we compute the number $\mathscr{K}(u, v)$ from the conditioned expectation

$$\mathscr{K}(u,v)=E_{t,x,y,z}\left\{\int_t^T H(s,X(s),Y(s),Z(s))\,ds + \frac{1}{L}\int_t^T |v(s)|^2\,ds-\frac{1}{M}\int_t^T |u(s)|^2\,ds\right\}. \tag{3.5}$$

In an obvious way we can define the stochastic Sobolev (or L^2) δ-games associated with the dynamics (3.1)–(3.4) and payoff (3.5). Let

$$\begin{aligned}\mathscr{K}^\delta(t,x,y,z)&\equiv\inf_{\Psi_\delta}\sup_{\Phi^\delta}\mathscr{K}[\Psi_\delta,\Phi^\delta],\\ \mathscr{K}_\delta(t,x,y,z)&\equiv\sup_{\Phi_\delta}\inf_{\Psi^\delta}\mathscr{K}[\Phi_\delta,\Psi^\delta]\end{aligned} \tag{3.6}$$

denote the upper and lower δ-values, respectively. Here, as in § 2, Φ^δ, Φ_δ are L^2 δ-strategies for u and Ψ^δ, Ψ_δ are L^2 δ-strategies for v.

The next theorem states that there is a saddle point in pure strategies for the L^2 game associated with (3.1)–(3.5).

THEOREM 3.1. *There exists a pair of functions* (u^*, v^*) *with* $u^*\in L^2(R^p)$ *and* $v^*\in L^2(R^q)$ *satisfying*

$$\mathscr{K}(u,v^*)\leqq\mathscr{K}(u^*,v^*)\leqq\mathscr{K}(u^*,v) \tag{3.7}$$

for any $u\in L^2$, $v\in L^2$.

Remark. Any pair of functions (u^*, v^*) satisfying (3.7) is called a saddle point in pure strategies, or a Nash equilibrium pair.

Proof of Theorem 3.1. Let L denote the second-order backward parabolic differential operator

$$\mathscr{L}=\partial/\partial t+(\varepsilon^2/2)\Delta_{x,y,z}, \tag{3.8}$$

where

$$\begin{aligned}\Delta_{x,y,z}k&=\sum_{i=1}^m\frac{\partial^2k}{\partial x_i^2}+\sum_{i=1}^p\frac{\partial^2k}{\partial y_i^2}+\sum_{i=1}^q\frac{\partial^2k}{\partial z_i^2}\\ &=\text{Laplacian in } x, y, z \text{ of } k.\end{aligned} \tag{3.9}$$

It follows from Ladyzenskaja–Solonnikov–Ural'ceva [10, V Thm. 8.1] that there exists a unique solution $U(t, x, y, z)$ of the backward semilinear Cauchy problem

$$\begin{aligned}\mathscr{L}U+\nabla_xU\cdot F(t,x,y,z)&+M/4|\nabla_yU|^2\\ &-L/4|\nabla_zU|^2+H(t,x,y,z)=0\end{aligned} \tag{3.10}$$

and

$$U(T,x,y,z)=0. \tag{3.11}$$

Furthermore, $U(t, x, y, z)$ is at least twice continuously differentiable in (x, y, z) and $\partial U/\partial t$ is at least continuous.

For this function $U(t, x, y, z)$ let $(\xi^*, \eta^*, \zeta^*)(\tau)$ denote the unique solution of the system of deterministic ordinary differential equations

$$d\xi/d\tau=F(\tau,\xi,\eta,\zeta)\qquad(0\leqq t<\tau\leqq T), \tag{3.12}$$

$$d\eta/d\tau = M/2\nabla_y U(\tau, \xi, \eta, \zeta), \tag{3.13}$$

$$d\zeta/d\tau = -L/2\nabla_z U(\tau, \xi, \eta, \zeta) \tag{3.14}$$

with initial conditions

$$\xi(t) = x, \qquad \eta(t) = y, \qquad \zeta(t) = z. \tag{3.15}$$

That (ξ^*, η^*, ζ^*) exists follows from the properties of the solution U and standard existence theorems in the theory of o.d.e.s.

Define the functions $u^*(\tau)$, $v^*(\tau)$ by

$$\begin{aligned} u^*(\tau) &= M/2\nabla_y U(\tau, \xi^*(\tau), \eta^*(\tau), \zeta^*(\tau)), \\ v^*(\tau) &= -L/2\nabla_z U(\tau, \xi^*(\tau), \eta^*(\tau), \zeta^*(\tau)). \end{aligned} \tag{3.16}$$

Then by [10, V Thm. 8.1], $u^* \in L^2(R^p)$, $v^* \in L^2(R^q)$.

Now, for any $u(\tau) \in L^2(R^p)$, $v(\tau) \in L^2(R^q)$ if we denote $\mathcal{K}(u, v)$ given in (3.5) by $\psi(t, x, y, z)$, then ψ satisfies

$$\mathcal{L}\psi + \nabla_x\psi \cdot F + \nabla_y\psi \cdot u + \nabla_z\psi \cdot v + H - \frac{1}{M}|u|^2 + \frac{1}{L}|v|^2 = 0 \tag{3.17}$$

and

$$\psi(T, x, y, z) = 0. \tag{3.18}$$

The argument in ψ, F and H is (t, x, y, z). The proof is a minor modification of Fleming–Rishel [3, pp. 128 and 148].

Then $\psi^*(t, x, y, z) \equiv \mathcal{K}(u^*, v^*)$ satisfies $\psi^*(T, x, y, z) = 0$ and

$$\begin{aligned} &\mathcal{L}\psi^* + \nabla_x\psi^* \cdot F + \nabla_y\psi^* \cdot M/2\nabla_y U + \nabla_z\psi^* \cdot (-L/2\nabla_z U) \\ &\qquad + H - \frac{1}{M}|M/2\nabla_y U|^2 + \frac{1}{L}|(-L/2\nabla_z U)|^2 = 0 \end{aligned} \tag{3.19}$$

with the argument of ψ^*, F, H and U being (t, x, y, z). $\mathcal{K}(u^*, v^*)$ is the payoff of the game (3.1)–(3.4) with $u \equiv u^*$, $v \equiv v^*$.

Consider now (3.10) for U; it can be written as

$$\begin{aligned} &\mathcal{L}U + \nabla_x U \cdot F + \nabla_y U \cdot (M/2\nabla_y U) - M/4|\nabla_y U|^2 \\ &\qquad + \nabla_z U \cdot (-L/2\nabla_z U) + L/4|\nabla_z U|^2 + H = 0. \end{aligned} \tag{3.20}$$

Comparing (3.20) with (3.19) we see that ψ^* and U satisfy the same equation (3.10). Since also $U(T, x, y, z) = \psi^*(T, x, y, z) = 0$ we have by uniqueness

$$U(t, x, y, z) \equiv \psi^*(t, x, y, z) \equiv \mathcal{K}(u^*, v^*). \tag{3.21}$$

Given any $v(\tau) \in L^2(R^q)$, let $\phi(t, x, y, z)$ solve $\phi(T, x, y, z) = 0$ and

$$\begin{aligned} &\mathcal{L}\phi + \nabla_x\phi \cdot F + \nabla_y\phi \cdot u^* + \nabla_2\phi \cdot v \\ &\qquad - \frac{1}{M}|u^*|^2 + \frac{1}{L}|v|^2 + H = 0, \end{aligned} \tag{3.22}$$

where u^* is given by (3.16) and is $u^*(t)$ in (3.22). For the function $U(t, x, y, z)$ defined as the solution of (3.10)–(3.11) we have for any $v \in L^2$

$$\mathscr{L}U+\nabla_x U\cdot F+\nabla_y U\cdot u^*+\nabla_z U\cdot v-\frac{1}{M}|u^*|^2+\frac{1}{L}|v|^2+H$$

$$(3.23)\qquad =\mathscr{L}U+\nabla_x U\cdot F+M/4|\nabla_y U|^2-L/4|\nabla_z U|^2+\frac{1}{L}|v+L/2\nabla_z U|^2+H$$

$$=\frac{1}{L}|v+L/2\nabla_z U|\geqq 0.$$

Since also $U(T,x,y,z)=\phi(T,x,y,z)=0$, by the maximum principle for parabolic equations [10] we have from (3.22) and (3.23) that

$$(3.24)\qquad U(t,x,y,z)\leqq\phi(t,x,y,z)\qquad (t,x,y,z)\in[0,T]\times R^m\times R^p\times R^q.$$

Given $v\in L^2$, let $\mathscr{K}(u^*,v)$ denote the payoff (3.5) for the stochastic L^2 game with dynamics (3.1)–(3.4) with u replaced by u^* given in (3.16). Let $\phi^*(t,x,y,z)\equiv\mathscr{K}(u^*,v)$. Then by (3.17) we see that ϕ^* satisfies (3.22) and $\phi^*(T,x,y,z)=0$. By uniqueness we have that $\phi^*\equiv\phi$. Then combining (3.21) with (3.24) and using the fact that $\phi^*=\phi$ we have shown that for any $v\in L^2$

$$\mathscr{K}(u^*,v^*)\leqq\mathscr{K}(u^*,v).$$

In a similar manner we see that for any $u\in L^2$

$$\mathscr{K}(u^*,v^*)\geqq\mathscr{K}(u,v^*);$$

that is

$$\mathscr{K}(u,v^*)\leqq\mathscr{K}(u^*,v^*)\leqq\mathscr{K}(u^*,v)$$

for any $u\in L^2(R^p)$, $v\in L^2(R^q)$. This completes the proof.

Since any game having a saddle point has a value we immediately deduce the following corollary (c.f. Friedman [4, Thm. 1.6.2]).

COROLLARY 3.2. *For any* $(t,x,y,z)\in[0,T]\times R^m\times R^p\times R^q$

$$\lim_{\delta\to 0}\mathscr{K}^\delta(t,x,y,z)=\lim_{\delta\to 0}\mathscr{K}_\delta(t,x,y,z)\equiv\mathscr{K}(t,x,y,z).$$

Furthermore $\mathscr{K}(t,x,y,z)=\mathscr{K}(u^*,v^*)$.

DEFINITION. The function $\mathscr{K}(t,x,y,z)$ is called the L^2*-value* of the stochastic game associated with (3.1)–(3.5). Corollary 3.2 implies that L^2-value exists whenever condition (A) holds even for nonlinear games and in the absence of the Isaacs condition [5].

Regarding the properties of the function $\mathscr{K}(t,x,y,z)$ we have the following

THEOREM 3.3. $\mathscr{K}(t,x,y,z)$ *is the unique solution of the backward, semilinear, parabolic Cauchy problem*

$$(3.25)\qquad \mathscr{K}_t+\varepsilon^2/2\Delta_{x,y,z}\mathscr{K}+\nabla_x\mathscr{K}\cdot F+M/4|\nabla_y\mathscr{K}|^2-L/4|\nabla_z\mathscr{K}|^2+H=0,$$

$$(3.26)\qquad \mathscr{K}(T,x,y,z)=0.$$

Furthermore $\mathscr{K}(t,x,y,z)$ *is an even periodic function in y and z of period* 2.

Proof. The first assertion is immediate from Corollary 3.2 and the proof of Theorem 3.1.

Let $y=(y_1,\cdots,y_p)$, $-y=(y_1,\cdots,-y_i,\cdots,y_p)$ for any $1\leqq i\leqq p$. The function $\mathscr{K}(t,x,-y,z)$ also satisfies the problems (3.25), (3.26) by the fact that F and H are even in y. By uniqueness $\mathcal{K}(t, x, -y, z) = \mathcal{K}(t, x, y, z)$ and hence $\mathcal{K}$ is even in y. The remaining assertions are shown similarly.

4. The Hamilton–Jacobi–Isaacs equation for deterministic Sobolev games. In this section we use our results developed in § 3 and further results developed below regarding stochastic games to determine some facts regarding the deterministic case.

For r and s given positive integers let $\mathcal{U}^r$ and $\mathcal{V}^s$ be the control sets given in (2.9); i.e., $\mathcal{U}^r$ is the ball of radius r in R^p and $\mathcal{V}^s$ is the ball of radius s in R^q.

Consider the stochastic (measurable) differential game associated with the dynamics (3.1)–(3.4) and payoff (3.5). Here u is allowed to be any measurable function with values a.e. in U^r and v is any measurable function with values a.e. in V^s. See Friedman [5], [6] for the definition of such games.

Since the functions u and v appear linearly in the dynamics (3.1)–(3.4) and separated in the payoff (3.5), this game has a value by Friedman [5], [6] which we denote by $\mathcal{W}^{r,s}(t, x, y, z)$. Furthermore $\mathcal{W}^{r,s}$ is the unique solution of the Cauchy problem

$$\begin{aligned}\mathcal{W}_t^{r,s} + \varepsilon^2/2\Delta_{x,yz}\mathcal{W}^{r,s} + \min_{|v|\leqq s}\max_{|u|\leqq r}\Big\{\nabla_x\mathcal{W}^{r,s}\cdot F + H + \nabla_y\mathcal{W}^{r,s}\cdot u \\ + \nabla_z\mathcal{W}^{r,s}\cdot v + \frac{1}{L}|v|^2 - \frac{1}{M}|u|^2\Big\} = 0,\end{aligned} \tag{4.1}$$

$$\mathcal{W}^{r,s}(T, x, y, z) = 0. \tag{4.2}$$

See Friedman [5, Lemma 3.5] for the proof.

The argument of $\mathcal{W}^{r,s}$, F and H in (4.1) is (t, x, y, z). Since these functions are independent of u and v (4.1) can be simplified and rewritten in a more useful form as

$$\begin{aligned}\mathcal{W}_t^{r,s} + \varepsilon^2/2\Delta_{x,yz}\mathcal{W}^{r,s} + \nabla_x\mathcal{W}^{r,s}\cdot F + H + M/4|\nabla_y\mathcal{W}^{r,s}|^2 - L/4|\nabla_z\mathcal{W}^{r,s}|^2 \\ + \frac{1}{L}\max_{|v|\leqq s}|v + L/2\nabla_z\mathcal{W}^{r,s}|^2 - \frac{1}{M}\min_{|u|\leqq r}|u - M/2\nabla_y\mathcal{W}^{r,s}|^2 = 0.\end{aligned} \tag{4.3}$$

THEOREM 4.1. *There is a constant C, independent of r, s, ε, M and L so that*

$$|\nabla_y\mathcal{W}^{r,s}(t, x, y, z)| \leqq C \quad \text{and} \quad |\nabla_z\mathcal{W}^{r,s}(t, x, y, z)| \leqq C.$$

Proof. The proof can be given using the maximum principle as in Jensen [9] or Friedman [7]. However, we present a proof based on differential games. We will prove only the first inequality.

For simplicity we take $p = q = 1$.

Let the initial position y for Y be in $[0, 1]$ and denote by (X^y, Y^y, Z^y) the solution of (3.1)–(3.4) given control functions $u(\cdot)$ and $v(\cdot)$ a.e. valued in $\mathcal{U}^r$ and $\mathcal{V}^s$, respectively. Also for $y' \in [0, 1]$, $y' \neq y$, denote by $(X^{y'}, Y^{y'}, Z^{y'})$ the solution of (3.1)–(3.4) with y replaced by y' but using the same control functions u and v. Denote the corresponding payoffs given by (3.5) as $\mathcal{K}^y(y, v)$ and $\mathcal{K}^{y'}(u, v)$.

Then, from (3.1)–(3.3) in integrated form, we immediately conclude from standard estimates in stochastic differential equations that

$$E|Y^y(\tau) - Y^{y'}(\tau)|^2 = |y - y'|^2$$

and from the Lipschitz continuity of F that

$$E|(X^y, Y^y, Z^y)(\tau)-(X^{y'}, Y^{y'}, Z^{y'})(\tau)|^2 \leqq A|y-y'|^2, \tag{4.4}$$

where A is a constant depending only on F and is independent of u, v, and ε, r, s, M and L.

From (4.4) and the uniform Lipschitz continuity of $H(t, \cdot, \cdot, \cdot)$ it follows from (3.5) that

$$|\mathcal{K}^y(y, v)-\mathcal{K}^{y'}(u, v)| \leqq A'|y-y'| \tag{4.5}$$

with A' independent of ε, r, s, M and L since we use the same control functions y and v in computing both payoffs.

By (4.5) we have from Friedman [5, Lemma 3.5, pp. 18 and 21] that

$$|\mathcal{W}^{r,s}(t, x, y, z)-\mathcal{W}^{r,s}(t, x, y, z)| \leqq C|y-y'|$$

and hence

$$|\partial/\partial y\, \mathcal{W}^{r,s}(t, x, y, z)| \leqq C \tag{4.6}$$

with C independent of ε, r, s, M and L, for all $y \in [0, 1]$. Finally from the form of the parabolic equation (4.3) and the even 2-periodicity properties of F and H we have that $W^{r,s}$ (t, x, y, z) is also an even 2-periodic function of y. The minimum will be attained at different functions u but the equation remains the same for both $\mathcal{W}^{r,s}(t, x, y, z)$ and $\mathcal{W}^{r,s}(t, x, -y, z)$. Hence the properties of $\mathcal{W}^{r,s}$ as a function of y are determined by $y \in [0, 1]$ and thus by (4.6) the theorem follows.

From Theorem 4.1 we immediately have from (4.3) that for any $r \geqq r_0$, $s \geqq s_0$ where

$$r_0=\frac{M}{2}C, \qquad s_0=\frac{L}{2}C \tag{4.7}$$

the function $\mathcal{W}^{r,s}(t, x, y, z)$ satisfies

$$\begin{aligned} \mathcal{W}_t^{r,s}+\varepsilon^2/2\Delta_{x,y,z}\mathcal{W}^{r,s}+\nabla_x\mathcal{W}^{r,s}\cdot F+H \\ +M/4|\nabla_y\mathcal{W}^{r,s}|^2-L/4|\nabla_z\mathcal{W}^{r,s}|^2=0 \end{aligned} \tag{4.8}$$

with $\mathcal{W}^{r,s}(T, x, y, z)=0$. By uniqueness we have $\mathcal{W}^{r,s}=\mathcal{W}^{r',s'}$ for r, $r' \geqq r_0$, s, $s' \geqq s_0$.

Since by Theorem 3.3, $\mathcal{K}(t, x, y, z)$, the value of the stochastic L^2 game also satisfies (4.8) with $\mathcal{K}(T, x, y, z)=0$ we have shown by uniqueness of solutions the following theorem.

THEOREM 4.2. *For any* $r \geqq r_0$, $s \geqq s_0$, $\mathcal{W}^{r,s}(t, x, y, z) \equiv \mathcal{K}(t, x, y, z)$ *and hence the stochastic (measurable) differential game has the value* $\mathcal{K}(t, x, y, z)$ *whenever* $r \geqq r_0$, $s \geqq s_0$.

COROLLARY 4.3. *There exists an equilibrium pair* (u^*, v^*) *for the stochastic measurable game when* $r \geqq r_0$, $s \geqq s_0$ *given by*

$$u^*(t, x, y, z)=M/2\nabla_y\mathcal{K}(t, x, y, z), \qquad v^*(t, x, y, z)=-L/2\nabla_z\mathcal{K}(t, x, y, z).$$

This equilibrium pair is a Nash point in feedback strategies.

Now denote $\mathcal{W}^{r,s}$, the solution of (4.1), (4.2) by $\mathcal{W}_\varepsilon^{r,s}$. Of course as seen above, when $r \geqq r_0$, $s \geqq s_0$, $\mathcal{W}_\varepsilon^{r,s}$ satisfies (4.8). We will investigate the behavior of $\mathcal{W}_\varepsilon^{r,s}$ as $\varepsilon \to 0$.

THEOREM 4.3. *Let* $\mathcal{W}_\varepsilon^{r,s}(t, x, y, z)$ *be as defined above and let* $K^{r,s}(t, x, y, z)$ *be the value function for the deterministic differential game associated with dynamics* (2.1)–(2.4) *and payoff* (2.5). *For this function* $K^{r,s}$ *which satisfies the first order backward*

Cauchy problem (2.11), (2.12) *for any* $r \geqq 0$, $s \geqq 0$, *we have*

$$\lim_{\varepsilon \to 0} \mathcal{W}_\varepsilon^{r,s}(t, x, y, z) = K^{r,s}(t, x, y, z) \tag{4.9}$$

uniformly on compact subsets. Furthermore there is a constant C *(*C *is the same constant as in Theorem* 4.1*) so that* $|\nabla_y K^{r,s}| \leqq C$ *and* $|\nabla_z K^{r,s}| \leqq C$ *for almost every* (t, x, y, z).

Proof. The proof is immediate from Friedman [7, Thms. 4.1 and 4.2, pp. 37–38].

COROLLARY 4.4. *When* $r \geqq r_0$, $s \geqq s_0$, $K^{r,s}(t, x, y, z)$ *satisfies, almost everywhere, the equation*

$$K_t^{r,s} + \nabla_x K^{r,s} \cdot F + M/4|\nabla_y K^{r,s}|^2 - L/4|\nabla_z K^{r,s}|^2 + H = 0 \tag{4.10}$$

and

$$K^{r,s}(T, x, y, z) = 0 \quad \textit{everywhere}. \tag{4.11}$$

Remark. Since $\mathcal{W}_\varepsilon^{r,s} = \mathcal{W}_\varepsilon^{r',s'}$ for $r, r' \geqq r_0$ and $s, s' \geqq s_0$ we also have $K^{r,s} = K^{r',s'}$ for $r, r' \geqq r_0$, $s, s' \geqq s_0$.

Let $\mathcal{K}_\varepsilon(t, x, y, z)$ denote the value of the stochastic L^2 game (we previously denoted it simply by $\mathcal{K}$). Let $\mathcal{K}_\varepsilon^\delta$ denote the upper δ-value and $\mathcal{K}_{\varepsilon,\delta}$ denote the lower δ-value. Let $K^\delta(t, x, y, z)$ and $K_\delta(t, x, y, z)$ denote the upper and lower δ-values for the deterministic L^2-game as presented in § 2.

THEOREM 4.5. (i) $|\mathcal{K}_\varepsilon^\delta(t, x, y, z) - K^\delta(t, x, y, z)| \leqq C\varepsilon$;

(ii) $|\mathcal{K}_{\varepsilon,\delta}(t, x, y, z) - K_\delta(t, x, y, z)| \leqq C\varepsilon$ *for some constant* C *independent of* δ, ε, *and* (t, x, y, z).

Proof. We only prove (i); (ii) is similar.

Since $\mathcal{K}_\varepsilon^\delta = \inf_{\Psi_\delta} \sup_{\Phi^\delta} \mathcal{K}_\varepsilon[\Psi_\delta, \Phi^\delta]$ and $K^\delta = \inf_{\Psi_\delta} \sup_{\Phi^\delta} K[\Psi_\delta, \Phi^\delta]$ where $\mathcal{K}_\varepsilon$ is the payoff (3.5) and K is the payoff (2.5) it suffices to show that

$$|\mathcal{K}_\varepsilon(u, v) - K(u, v)| \leqq C\varepsilon \tag{4.12}$$

with C independent of u, v.

Given $u, v \in L^2$, let $(X, Y, Z)(\tau)$ denote the solution of (3.1)–(3.4) and let $(\xi, \eta, \zeta)(\tau)$ denote the solution of (2.1)–(2. 4). Then by the properties of a (standard) Brownian motion $w(\tau)$ we easily see that

$$E|(X, Y, Z)(\tau) - (\xi, \eta, \zeta)(\tau)|^2 \leqq C\varepsilon^2(\tau - t). \tag{4.13}$$

Further, since we are using the same controls u and v we have from (4.13) the inequality (4.12).

COROLLARY 4.6. (i) $K(t, x, y, z) \equiv \lim_{\delta \to 0} K^\delta(t, x, y, z) = \lim_{\delta \to 0} K_\delta(t, x, y, z)$ *with the limits existing.*

(ii) $K(t, x, y, z) = \lim_{\varepsilon \to 0} \mathcal{K}_\varepsilon(t, x, y, z)$.

Proof. (i) $|K^\delta - K_\delta| \leqq |K^\delta - \mathcal{K}_\varepsilon^\delta| + |\mathcal{K}_\varepsilon^\delta - \mathcal{K}_{\varepsilon,\delta}| + |\mathcal{K}_{\varepsilon,\delta} - K_\delta|$. The middle term goes to zero with δ by Corollary 3.2. The other terms go to zero with δ by the theorem.

(ii) immediate from Theorem 4.5.

COROLLARY 4.7. $K(t, x, y, z)$ *is a uniformly Lipschitz continuous function which satisfies almost everywhere the Hamilton–Jacobi–Isaacs equation*

$$K_t + \nabla_x K \cdot F + M/4|\nabla_y K|^2 - L/4|\nabla_z K|^2 + H = 0 \tag{4.14}$$

with

$$K(T, x, y, z) = 0 \quad \textit{everywhere}. \tag{4.15}$$

Remark. The function K is called the L^2-value of the L^2 differential game. Corollary 4.6 states that the L^2-value exists even for nonlinear F and H.

Remark. It immediately follows from Lemma 2.1 that the extended Sobolev game has value $\bar{W}_{M,L}(t, x, y, z) = K(t, x, y, z)$ and from Lemma 1.1 that if $y \in [0, 1]$, $z \in [0, 1]$ then the Sobolev game defined in § 1 has value $W_{M,L}(t, x, y, z) = \bar{W}_{M,L}(t, x, y, z)$. Furthermore $\bar{W}_{M,L}$ also satisfies the problem (4.14), (4.15) at points of differentiability.

5. Saddle points in pure strategies for Sobolev games. Assume $\bar{W}_{M,L}(t, x, y, z) = K(t, x, y, z)$ is a twice continuously differentiable solution of the problem (4.14), (4.15). Denote by (ξ^*, η^*, ζ^*) the unique solution of the system of ordinary differential equations

$$d\xi/d\tau = F(\tau, \xi, \eta, \zeta) \qquad (0 \leqq t < \tau \leqq T), \tag{5.1}$$

$$d\eta/d\tau = M/2\nabla_y \bar{W}_{M,L}(\tau, \xi, \eta, \zeta), \tag{5.2}$$

$$d\zeta/d\tau = -L/2\nabla_z \bar{W}_{M,L}(\tau, \xi, \eta, \zeta), \tag{5.3}$$

$$\xi(t) = x \in R^m, \qquad \eta(t) = y \in R^p, \qquad \zeta(t) = z \in R^q. \tag{5.4}$$

THEOREM 5.1. *With $\bar{P}(\eta, \zeta)$ given by* (1.3) *with f, h replaced by F, H for* $(t, x, y, z) \in [0, T] \times R^m \times R^p \times R^q$ *we have* $\bar{W}_{M,L}(t, x, y, z) = \bar{P}(\eta^*, \zeta^*)$.

Proof. Note that (ξ^*, η^*, ζ^*) is well-defined by our assumption that $\bar{W}_{M,L}$ is C^2.

If $a(\tau) = \bar{W}_{M,L}(\tau, \xi^*(\tau), \eta^*(\tau), \zeta^*(\tau)) + \int_\tau^T H(s, \xi^*(s), \eta^*(s), \zeta^*(s))\, ds - (1/M) \cdot \int_\tau^T (\dot{\eta}^*(s))^2\, ds + (1/L) \int_\tau^T (\dot{\zeta}^*(s))^2\, ds$ then

$$\frac{da}{d\tau} = \partial \bar{W}_{M,L}/\partial\tau + \nabla_x \bar{W}_{M,L} \cdot F + M/4|\nabla_y \bar{W}_{M,L}|^2 - L/4|\nabla_z \bar{W}_{M,L}|^2 + H = 0.$$

So $a(\tau) =$ constant. Hence $a(t) = a(T)$ and the theorem is proved.

Given any function $\zeta(\tau) \in Z_z(J_t; R^q)$, let $(\tilde{\xi}, \tilde{\eta})$ denote the unique solution of

$$d\tilde{\xi}/d\tau = F(\tau, \tilde{\xi}, \tilde{\eta}, \zeta) \qquad (0 \leqq t < \tau \leqq T), \tag{5.5}$$

$$d\tilde{\eta}/d\tau = M/2\nabla_y \bar{W}_{M,L}(\tau, \tilde{\xi}, \tilde{\eta}, \zeta), \tag{5.6}$$

$$\tilde{\xi}(t) = x, \qquad \tilde{\eta}(t) = y. \tag{5.7}$$

Also for any $\eta(\tau) \in Y_y(J_t; R^p)$, let $(\hat{\xi}, \hat{\zeta})$ denote the unique solution of

$$d\hat{\xi}/d\tau = F(\tau, \hat{\xi}, \eta, \hat{\zeta}) \qquad (0 \leqq t < \tau \leqq T), \tag{5.8}$$

$$d\hat{\zeta}/d\tau = -L/2\nabla_z \bar{W}_{M,L}(\tau, \hat{\xi}, \eta, \hat{\zeta}), \tag{5.9}$$

$$\hat{\xi}(t) = x, \qquad \hat{\zeta}(t) = z. \tag{5.10}$$

Then our saddle point theorem becomes

THEOREM 5.2. *For any $\eta(\tau) \in Y_y(J_t; R^p)$, $\zeta(\tau) \in Z_z(J_t; R^q)$ we have*

$$\bar{P}(\eta, \hat{\zeta}) \leqq \bar{P}(\eta^*, \zeta^*) \leqq \bar{P}(\tilde{\eta}, \zeta),$$

That is, (η^*, ζ^*) is an equilibrium point for the Sobolev differential game.

Proof. For $\tau \in [t, T]$ and $\zeta \in Z_z(J_t; R^q)$ let

$$b(\tau) = \bar{W}_{M,L}(\tau, \tilde{\xi}(\tau), \tilde{\eta}(\tau), \zeta(\tau)) + \int_\tau^T H(s, \tilde{\xi}(s), \tilde{\eta}(s), \zeta(s))\, ds$$
$$+ \frac{1}{L}\int_\tau^T |\dot{\zeta}(s)|^2\, ds + \frac{1}{M}\int_\tau^T |\dot{\tilde{\eta}}(s)|^2\, ds$$

with $(\tilde{\xi}, \tilde{\eta})$ as in (5.5)–(5.7). Then

$$
\begin{aligned}
\frac{db}{d\tau} &= \partial \bar{W}_{M,L}/\partial\tau + \nabla_x \bar{W}_{M,L} \cdot F(\tau, \tilde{\xi}, \tilde{\eta}, \zeta) + \nabla_y \bar{W}_{M,L} \cdot (M/2\nabla_y \bar{W}_{M,L}) \\
&\quad + \nabla_z \bar{W}_{M,L} \cdot \frac{d\zeta}{d\tau} + H(\tau, \tilde{\xi}, \tilde{\eta}, \zeta) + \frac{1}{L}\left|\frac{d\zeta}{d\tau}\right|^2 - M/4|\nabla_y \bar{W}_{M,L}|^2.
\end{aligned} \tag{5.11}
$$

Rewriting (5.11) we have by (4.14)

$$
\begin{aligned}
\frac{db}{d\tau} = \partial \bar{W}_{M,L}/\partial\tau + \nabla_x \bar{W}_{M,L} \cdot F + M/4|\nabla_y \bar{W}_{M,L}|^2 + H + \frac{1}{L}\left|\frac{d\zeta}{d\tau} + L/2\nabla_z \bar{W}_{M,L}\right|^2 \\
- L/4|\nabla_z \bar{W}_{M,L}|^2 = \frac{1}{L}\left|\frac{d\zeta}{d\tau} + L/2\nabla_z \bar{W}_{M,L}\right|^2 \geqq 0.
\end{aligned}
$$

Hence $b(t) \leqq b(T)$ and so by the definition of b and (4.15)

$$\bar{W}_{M,L}(t, x, y, z) \leqq \bar{P}(\tilde{\eta}, \zeta).$$

Using Theorem 5.1 we obtain the second part of the inequality in the assertion of the theorem. We similarly obtain the first part and the theorem follows.

Remark. By Lemma 1.2 the value functions $\bar{W}_{M,L}(t, x, y, z) = K(t, x, y, z)$ are even 2-periodic functions in y and z. Hence, the properties of the function $\bar{W}_{M,L}$ holding for $(y, z) \in [0, 1] \times [0, 1]$ will hold for all y and z. Intuitively, therefore, one would expect that the equilibrium pair of functions (η^*, ζ^*) achieving the value $\bar{W}_{M,L}$ (and which are optimal) might as well remain in $[0, 1] \times [0, 1]$ for all time τ, assuming, of course, that they start in $[0, 1] \times [0, 1]$. This is the content of our next theorem. For simplicity, we take $p = q = 1$.

THEOREM 5.3. *Suppose* $y \in [0, 1]$. *Then given any* $\zeta(\tau) \in Z_z(J_t; R^1)$ *the function* $\tilde{\eta}(\tau)$ *given in* (5.5)–(5.7) *satisfies* $\tilde{\eta}(\tau) \in [0, 1]$ *for all* $t \leqq \tau \leqq T$. *Similarly if* $z \in [0, 1]$, *for any* $\eta(\tau) \in Y_y(J_t; R^1)$ *the function* $\hat{\zeta}(\tau)$ *given in* (5.8)–(5.10) *satisfies* $\hat{\zeta}(\tau) \in [0, 1]$ *for all* $t \leqq \tau \leqq T$.

Proof. We will only prove the first assertion.

Suppose, for example that $\tilde{\eta}(\sigma) = 0$ at some time $T > \sigma \geqq t$. Then $\tilde{\eta}(\tau) \geqq 0$ for all $\sigma \leqq \tau \leqq T$. Now suppose this is not the case. Then we can find a neighborhood of σ satisfying: (i) there is a σ_0 in this neighborhood, $\sigma_0 > \sigma$, so that $\tilde{\eta}(\sigma_0) < 0$; (ii) there exists $\tau_0 \in [\sigma, \sigma_0]$ with $|\tilde{\eta}(\tau_0)| \geqq |\tilde{\eta}(\tau)|$ for all $\tau \in [\sigma, \sigma_0]$; (iii) τ_0 satisfies $(\tau_0 - \sigma) < 2/MC$, where C is a Lipschitz constant for $\partial \bar{W}_{M,L}/\partial y$ as a function of y. Note that $\tau_0 > \sigma$ since $\tilde{\eta}(\sigma) = 0$. Then by (5.5)–(5.7) we have

$$
\begin{aligned}
\tilde{\eta}(\tau_0) &= \tilde{\eta}(\sigma) + \int_\sigma^{\tau_0} M/2\partial \bar{W}_{M,L}(s, \tilde{\xi}, \tilde{\eta}, \zeta)/\partial y \, ds \\
&= M/2 \int_\sigma^{\tau_0} [\partial \bar{W}_{M,L}(s, \tilde{\xi}, \tilde{\eta}, \zeta)/\partial y - \partial \bar{W}_{M,L}(s, \tilde{\xi}, 0, \zeta)/\partial y] \, ds.
\end{aligned}
$$

The second equality follows from the fact that $\bar{W}_{M,L}(t, x, y, z) = \bar{W}_{M,L}(t, x, -y, z)$ implies that $\partial \bar{W}_{M,L}(t, x, 0, z)/\partial y = 0$ for all (t, x, z). Hence by the definition of τ_0

$$|\tilde{\eta}(\tau_0)| \leqq CM/2 \int_\sigma^{\tau_0} |\tilde{\eta}(s)| \, ds \leqq CM/2(\tau_0 - \sigma)|\tilde{\eta}(\tau_0)| < |\tilde{\eta}(\tau_0)|$$

a contradiction.

Similarly we show that $\tilde{\eta}(\sigma)=1$, for some $T>\sigma\geqq t$, implies $\tilde{\eta}(\tau)\leqq 1$ for all $\sigma\leq\tau\leqq T$. Here we use the fact that $\bar{W}_{M,L}(t,x,y,z)=\bar{W}_{M,L}(t,x,2-y,z)$ implies $\partial\bar{W}_{M,L}(t,x,l,z)/\partial y=0$. This completes the proof.

Remark. The theorem implies that the equilibrium pair (η^*,ζ^*) given by (5.1)–(5.4) satisfy $\eta^*(\tau),\zeta^*(\tau)\in[0,1]$ for all $\tau\in J_t$ if $y,z\in[0,1]$. In view of Lemma 2.2, the theorem states that the "constant" lower δ-strategies Ψ_δ for ζ^* and Φ^δ for η^*; that is Ψ_δ has range $\{\zeta^*\}$ and Φ_δ has range $\{\eta^*\}$ for all $\delta>0$, satisfy $\Psi_\delta\in\Lambda_\delta$ and $\Phi_\delta\in\Sigma_\delta$.

6. Convergence as M and $L\to\infty$. In this section we denote the stochastic L^2-value function $\mathcal{K}(t,x,y,z)$ by $\mathcal{K}^\varepsilon_{M,L}(t,x,y,z)$ to indicate the dependence of $\mathcal{K}$ on the constants $\varepsilon, M, L>0$. We investigate the asymptotic behavior of $\mathcal{K}^\varepsilon_{M,L}$ as $M,L\to\infty$.

By Theorem 3.3, $\mathcal{K}^\varepsilon_{M,L}(t,x,y,z)$ is the unique solution of

$$\mathcal{K}_t+\varepsilon^2/2\Delta_{x,y,z}\mathcal{K}+\nabla_x\mathcal{K}\cdot F+M/4\Phi(\nabla_y\mathcal{K})-L/4\Psi(\nabla_z\mathcal{K})+H=0, \tag{6.1}$$

$$\mathcal{K}(T,x,y,z)=0, \tag{6.2}$$

where $\Phi\colon R^p\to R^1$ and $\Psi\colon R^q\to R^1$ are given by

$$\Phi(y)=|y|^2,\qquad \Psi(z)=|z|^2. \tag{6.3}$$

Let $V^\pm_\varepsilon(t,x)$ denote, respectively, the unique solutions of the problems

$$V^\pm_t+\varepsilon^2/2\Delta_x V^\pm+H^\pm(t,x,\nabla_x V^\pm)=0, \tag{6.4}$$

$$V^\pm(T,x)=0. \tag{6.5}$$

where the Hamiltonians $H^\pm(t,x,r)$ are given by

$$H^+(t,x,r)=\min_{z\in I^q}\max_{y\in I^p}\{r\cdot F(t,x,y,z)+(t,x,y,z)\} \tag{6.6}$$

and

$$H^-(t,x,r)=\max_{y\in I^p}\min_{z\in I^q}\{r\cdot F(t,x,y,z)+H(t,x,y,z)\} \tag{6.7}$$

with $r\in R^m$, $x\in R^m$, $t\in[0,T]$.

Note the regions over which we maximize and minimize. By the definitions of F and H as extensions of f and h as in § 1, we may substitute f, h for F, H.

THEOREM 6.1. (i) $\lim_{L\to\infty}\lim_{M\to\infty}\mathcal{K}^\varepsilon_{M,L}(t,x,y,z)=V^+_\varepsilon(t,x)$;

(ii) $\lim_{M\to\infty}\lim_{L\to\infty}\mathcal{K}^\varepsilon_{M,L}(t,x,y,z)=V^-_\varepsilon(t,x)$ *uniformly in* ε.

Proof. By the definitions of Φ and Ψ in (6.3) we have $\Phi(0)=D\Phi(0)=0$, $\Psi(0)=D\Psi(0)=0$ and $D^2\Phi(0)$, $D^2\Psi(0)$ are positive definite. These are exactly the conditions on the nonlinearity in the parabolic equation required to apply Theorem 1.2 of Jensen [9]. The conclusions of Jensen's theorem are exactly the conclusions of Theorem 6.1. This completes the proof.

Next, let $K_{M,L}$ denote the value of the (deterministic) L^2 differential game associated with dynamics (2.1)–(2.4) and payoff (2.5). Let $V^\pm(t,x)$ denote the upper and lower values of the "measurable" differential game of fixed duration associated with the dynamics

$$d\xi/d\tau=f(\tau,\xi,\eta,\zeta), \tag{6.8}$$

$$\xi(t)=x \tag{6.9}$$

and payoff

$$J(\eta, \zeta) = \int_t^T h(s, \xi, \eta, \zeta)\, ds. \tag{6.10}$$

The measurable functions η and ζ are the control functions; η is the maximizer and ζ is the minimizer of J. The control set for η is I^p and I^q is the control set for ζ. Then we have

THEOREM 6.2. (i) $\lim_{L\to\infty} \lim_{M\to\infty} K_{M,L} = V^+$,

(ii) $\lim_{M\to\infty} \lim_{L\to\infty} K_{M,L} = V^-$.

Proof of (i). By Lemma 3.2 of Friedman [5], $V_\varepsilon^\pm \to V^\pm$ as $\varepsilon \to 0$. By Theorem 4.5, $\mathcal{K}^\varepsilon_{M,L} \to K_{M,L}$ as $\varepsilon \to 0$. Using these facts and Theorem 6.1, (i) is immediate.

COROLLARY 6.3. *The value of the differential game* (6.8)–(6.10) *exists if and only if* $\lim_{L\to\infty} \lim_{M\to\infty} K_{M,L} = \lim_{M\to\infty} \lim_{L\to\infty} K_{M,L}$.

The proof is by definition of the existence of value [4], [5] and Theorem 6.2.

Remarks. (i) When there is only one player, say η, the maximizer, in our Sobolev and L^2 games, then the differential game is an optimal control problem. The results of this paper apply. In particular, if $A^M(t, x, y)$ denotes the optimal cost, then A^M satisfies the equation

$$A_t^M + A_x^M \cdot F(t, x, y) + M/4|A_y^M|^2 + H = 0. \tag{6.11}$$

According to Tamburro [11] and the references given there, there is a generalized solution of (6.11) in the sense of Kruzkov. This solution is *unique* in the class Kruzkov considers (c.f. [11, p. 250]). The reason for this is the strict convexity of $\Gamma(t, x, y, r, x) \equiv r \cdot F(t, x, y) + H(t, x, y) + M/4|s|^2$ in s and linearity of Γ in r. We cannot apply this result to the Sobolev *game* because the Hamiltonian is convex-concave; that is, uniqueness will not hold in general in the game problem.

(ii) The results of this paper hold for more general payoffs than that considered here. In particular we may take the payoff $P(\eta, \zeta) = g(\xi(T)) + \int_t^T h(s, \xi(s), \eta(s), \zeta(s))\, ds$ which includes a terminal part $g(\xi(T))$.

Acknowledgment. I would like to thank Robert Jensen for many helpful discussions, and in particular for the proof of Theorem 5.3.

REFERENCES

[1] E. N. BARRON, *Differential games with Lipschitz control functions and applications to games with partial differential equations*, Trans. Amer. Math. Soc., 219 (1976), pp. 39–76.

[2] ———, *Differential games with Lipschitz control functions and fixed initial control positions*, J. Differential Equations, 26 (1977), pp. 161–180.

[3] W. H. FLEMING AND R. W. RISHEL, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, 1975.

[4] A. FRIEDMAN, *Differential Games*, Wiley-Interscience, New York, 1971.

[5] ———, *Differential Games*, CBMS No. 18, American Mathematical Society, Providence, RI, 1974.

[6] ———, *Stochastic differential games*, J. Differential Equations, 11 (1972), pp. 79–108.

[7] ———, *The Cauchy problem for first order partial differential equations*, Indiana Univ. Math. J., 23 (1973), pp. 27–40.

[8] I. I. GIKHMAN AND A. V. SKOROHOD, *Introduction to the Theory of Random Processes*, Saunders, Philadelphia, 1969.

[9] R. JENSEN, *Asymptotic perturbation of a class of semilinear parabolic partial differential equations*, Trans. Amer. Math. Soc., to appear.

[10] O. A. LADYZENSKAJA, V. A. SOLONNIKOV AND N. N. URAL'CEVA, *Linear and Quasilinear Equations of Parabolic Type*, American Mathematical Society, Providence, RI, 1968.

[11] M. B. TAMBURRO, *The evolution operator solution of the Cauchy problem for the Hamilton–Jacobi equation*, Israel J. Math., 26 (1977), pp. 232–264.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 3, May 1980

A CUTTING-PLANE GAME FOR FACIAL DISJUNCTIVE PROGRAMS*

ROBERT G. JEROSLOW†

Abstract. Balas' characterization, of the convex span of feasible solutions to a system of facial constraints, is generalized through the device of first viewing the characterization as a two person "game" on a polytope, and then enlarging the class of "moves" open to one of the "players." Both primal and dual cutting-plane algorithms are presented for facial constraint systems, and are then proven finitely-convergent by use of our generalization of Balas' result.

Introduction. We obtain an extension of Balas' characterization (Balas (1974)) of the convex hull of feasible solutions for a bounded system of "facial constraints" (in the terminology of Balas), by viewing the characterization as a two-person "game" on polytopes, and then enlarging the class of moves open to one of the players. We then illustrate how this "game" can be used to develop cutting-plane algorithms, of both a primal and dual nature, which optimize a linear form subject to facial constraints.

The scope of practical problems which can be modeled by bounded facial constraints is essentially that of the bounded integer program, though the scope is different when the facial constraints may be unbounded. In the bounded case, an important distinction can be made between the modeling via facial constraints and that via integer variables; details are given in Appendix A.

The game we consider here is played on a polytope P by two players, the "indicating player" (Player 1) and the "cutting player" (Player 2). Certain of the extreme points of P are "essential" and certain are "inessential;" the "essential" extreme points are those which satisfy the facial constraints, and all other extreme points are "inessential."

Each round of the game is as follows. Player 1 indicates a nonessential extreme point, if he wishes to and can find one; if he indicates no point the game terminates and Player 2 wins.

Assuming that Player 1 indicates a nonessential point, Player 2 must produce a "cutting plane" (i.e., valid linear inequality for the facial constraints) which is not satisfied by this point, but which is satisfied by every essential extreme point. The game then continues on the polytope P' obtained by adjoining the cut to P. There is a restriction on the cutting-planes available to Player 2, which we will next discuss. Player 1 wins if the game continues indefinitely.

Any nonessential extreme point violates at least one, and usually several, of the logical conditions of the facial constraints system. Player 2 must choose only one of these violated conditions, and the cutting-plane he produces must depend only on this one condition; i.e. must be valid even if the other violated logical constraints were omitted.

From the nature of the game when Player 1 uses all available nonessential points, Player 2 wins exactly if, after adding finitely many cuts, the convex hull of the essential extreme points is obtained. Hence, if there were no restrictions on Player 2, he will always win, by adding a defining inequality for this convex hull at each round. However, typically a defining inequality depends on several logical conditions being violated, hence Player 2 cannot follow this strategem.

* Received by the editors August 11, 1977, and in revised form August 17, 1979.

† College of Industrial Management, Georgia Institute of Technology, Atlanta, Georgia 30332.

Nevertheless, our main result (Theorem G of § 2) is that Player 2 can always win. This is true, despite the fact that the addition of a cutting-plane often increases the number of nonessential points. The method of proof is to have Player 2 guided by a sequence of "partial convex hulls" and their defining inequalities. Since no linear form need increase or decrease, and no lexicographic vector need increase or decrease, as Player 1 freely chooses nonessential points to be cut away, this is a fundamentally different method of obtaining finite convergence than the earlier methods. Nevertheless, there are lexicographic aspects in the changes of state (see § 1 below) and hence the reasoning by which we prove finiteness.

This game yields finitely-convergent dual algorithms, by viewing the next extreme point found by dual simplex re-optimization, after a cutting-plane has been added, as the point indicated by Player 1, when it does not satisfy the facial constraints. Obtaining primal algorithms can also be done. Both primal and dual algorithms will be discussed in § 2. While our result can be extended to the case of an unbounded feasible region, by means of adjoining an infinite quantity as we did in Jeroslow (1976), this extension does not appear to be computationally promising, because of the high order of arithmetic operations needed to perform pivot steps in the field extension.

The restriction on Player 2, that his cuts must depend only on one logical condition, corresponds computationally to obtaining the cuts as extreme points of certain linear programs which are defined in § 1. Pivoting proceeds until an extreme point is found, at which a linear form, that is determined by the point to be cut off, becomes positive. In all cases, the linear form can be made positive. A "Phase 1" for the pivoting procedure can be avoided by use of a starting basis described in Appendix B, which often produces a cut that makes some problem constraint redundant.

The algorithms here do not have the rigorous upper bounds on the number of cuts to be retained, which one finds for the lexicographic methods. There are nevertheless many instances in which cuts can be dropped; we discuss this matter in § 2.

1. Terminology and basic lemmas. Throughout the paper, we shall consider constraint sets of the form:

$$Ax \geqq b \tag{1a}$$

and

(1b) for each $h = 1, \cdots, t$ there is at least one $i \in S_h$ for which $d^i x \geqq d_0^i$.

Constraint sets of this very general nature were studied by Balas (1974). As (1) incorporates both logical and linear restrictions on $x = (x_1, \cdots, x_r)$, the set of feasible points need not be convex, in fact, it often is discrete. In (1a) and (1b), of course, A is an $m \times r$ real matrix, b an $m \times 1$ real vector, each d^i is a 1 by r real vector, and d_0^i is a real scalar. The S_h $(h = 1, \cdots, t)$ are sets of indices, which can be assumed disjoint.

An important special case of Balas (1974) is the *generalized linear complementarity* problem:

$$\text{(GLC)} \qquad Gy + Hz \geqq h; \qquad y, z \geqq 0; \quad y \cdot z = 0.$$

In GLC, $y = (y_1, \cdots, y_s)$, $z = (z_1, \cdots, z_s)$, G and H are $m \times s$ and h is $m \times 1$. To see that GLC can be converted into the format (1), first put $A = [G \vdots H]$, $x = (y, z)$. Next, put $t = s$, $S_h = \{i, s+i\}$, and let $d^i x \geqq d_0^i$ resp. $d^{s+i} x \geqq d_0^{s+i}$ be $-y_i \geqq 0$ resp. $-z_i \geqq 0$ $(i = 1, \cdots, s)$. Since $y, z \geqq 0$ are among the constraints of GLC, if $d^i x \geqq d_0^i$ holds for some $i \in S_h$, we have $y_i z_i = 0$; hence if such holds for all $h = 1, \cdots, t$ then $y \cdot z = 0$, and (1) is equivalent to (GLC).

Another special case of (1) is the constraint set of the bivalent integer program:

(BIP) $Dx \geqq d$ and $x_j = 0$ or 1 for $j = 1, \cdots, r$.

In (BIP), it is understood that $Dx \geqq d$ includes the constraints $0 \leqq x \leqq e$, $e = (1, 1, \cdots, 1)$. Here $t = r$, $S_h = \{j, r+j\}$, $d^j x \geqq d_0^j$ is $-x_j \geqq 0$ and $d^{r+j}x \geqq d_0^{r+j}$ is $x_j \geqq 1$ $(j = 1, \cdots, r)$.

The constraint set (1) is called *facial* (Balas (1974)) if for all $h = 1, \cdots, t$ and all $i \in S_h$,

(FAC) $\{x \mid d^i x = d_0^i, Ax \geqq b\}$ is a face of $\{x \mid Ax \geqq b\}$.

The face described in (FAC) may, of course, be empty.

We assume that (1) *is facial throughout the paper. We also assume that* $\{x \mid Ax \geqq b\}$ *is nonempty and bounded.*

Both (GLC) and (BIP) are facial. For example, $\{(y, z) \mid -y_i \geqq 0, Gy + Hz \geqq h; y, z \geqq 0\} = \{(y, z) \mid y_i = 0, Gy + Hz \geqq h; y, z \geqq 0\}$ is a face of $\{(y, z) \mid Gy + Hz \geqq h; y, z \geqq 0\}$ since the inequality $y_i \geqq 0$ is included among those of (GLC). These two examples serve to motivate our interest in facial constraint sets of the form (1).

For facial constraints, all extreme points of the convex span of the feasible solutions to (1) (which clearly must themselves be feasible) are extreme points of $\{x \mid Ax \geqq b\}$ (see Balas (1974) for this and related results on facial constraints). However, $\{x \mid Ax \geqq b\}$ may (and often does) have other extreme points, and this paper is concerned with methods for "cutting" the "unwanted" extreme points away, in order to reveal portions of the convex span of the feasible points. We now begin the technical developments that are needed to accomplish this goal.

A 2θ-tuple $(h(1), T_1, h(2), T_2, \cdots, h(\theta), T_\theta)$ of integers $h(j)$, $1 \leqq h(j) \leqq t$, and nonempty sets T_j of real linear inequalities, will be called a *state* if certain requirements (to be described next) are met, and the *length* of such a state is θ. To each *initial segment* $(h(1), T_1, h(2), T_2, \cdots, h(i-1), T_{i-1}, h(i))$ of a state $(h(1), T_1, h(2), T_2, \cdots, h(\theta), T_\theta)(i \geqq 1)$ will be assigned a finite set F_i of real linear inequalities, and we shall require that

(2) $T_i \subseteq F_i$ for $i = 1, \cdots, \theta$.

The set $F_i = F_i(h(1), T_1, \cdots, h(i-1), T_{i-1}, h(i))$ is a function of the initial segment, although this fact is suppressed in our notation.

LEMMA 1. *For any integer* $\theta \geqq 0$, *there are finitely many states of length not exceeding* θ.

Proof. Trivial, as each set F_i is finite. Q.E.D.

The sets F_i we shall use will be of a very particular nature, and to describe them we shall need some preliminary discussion.

The next result is an easy adaptation of a consequence of results in Blair and Jeroslow (1978).

In what follows, conv(S) resp. clconv(S) denotes the convex hull resp. closed convex hull of S; notation is from Rockafellar (1970). We put conv $(\varnothing)$ = clconv $(\varnothing) = \varnothing$.

PROPOSITION 2. *Suppose that* $\{x \mid D^k x \geqq 0\} = \{0\}$ *for* $k = 1, \cdots, \tau$.

Then the linear inequality

(3) $\pi x \geqq \pi_0$

is valid for the set

$$(4)\qquad CH = \operatorname{clconv}\left(\bigcup_{k=1}^{\tau}\{x|D^kx \geqq d^k\}\right)$$

if and only if there are vectors of multipliers $\lambda^k \geqq 0$ *with*

$$(5a)\qquad \lambda^kD^k = \pi, \qquad k = 1, \cdots, \tau,$$

$$(5b)\qquad \lambda^kd^k \geqq \pi_0, \qquad k = 1, \cdots, \tau.$$

Proof. Note that (3) is valid for the set (4) if and only if (3) is valid for each of the sets

$$(6)\qquad P_k = \{x|D^kx \geqq d^k\}.$$

Clearly, if (5) holds, (3) is valid for each P_k, hence valid for (4). We need only prove that, if (3) is valid for each P_k, then (5) holds.

If $P_k \neq \varnothing$, since π_0 is a lower bound on the value of the bounded and consistent linear program $\min\{\pi x|D^kx \geqq d^k\}$, it is also a lower bound on the dual linear program $\max\{\lambda d^k|\lambda D^k = \pi, \lambda \geqq 0\}$, and hence there is λ^k with $\lambda^kd^k \geqq \pi_0, \lambda^kD^k = \pi, \lambda^k \geqq 0$.

If $P_k = \varnothing$, there is some vector f^k such that $P'_k = \{x|D^kx \geqq f^k\} \neq \varnothing$. By the hypothesis, πx is bounded on P'_k. Hence, as before, there is a vector $\theta^k \geqq 0$ with $\theta^kD^k = \pi$. But as $P_k = \varnothing$, there is a vector $\gamma^k \geqq 0$ with $\gamma^kD^k = 0$, $\gamma^kd^k > 0$. Then for the real scalar $\rho \geqq 0$ sufficiently large, and $\lambda^k = \theta^k + \rho\gamma^k$, we have $\lambda^kD^k = \pi$ and $\lambda^kd^k \geqq \pi_0$. Since k was arbitrary, (5) holds. Q.E.D.

We remark that the hypothesis $\{x|D^kx \geqq 0\} = \{0\}$ $(h = 1, \cdots, \tau)$ can be dropped if $\{x|D^kx \geqq d^k\} \neq \varnothing$ for all $h = 1, \cdots, \tau$. However, Proposition 2 is more useful, stated as it is, because in many applications below one does not wish to test if $\{x|D^kx \geqq d^k\} \neq \varnothing$ in order to discard inconsistent systems. In contrast, the condition $\{x|D^kx \geqq 0\} = \{0\}$ turns out to be easier to insure.

According to Proposition 2, all valid inequalities (3) for the set (4) arise as projections on to the (π, π_0)-coordinates of solutions $(\lambda^1, \cdots, \lambda^\tau, \pi, \pi_0)$ to the homogeneous linear system

$$(7a)\qquad \lambda^kD^k - \pi = 0, \quad \lambda^kd^k - \pi_0 \geqq 0, \quad \lambda^k \geqq 0 \quad \text{for } k = 1, \cdots, \tau.$$

If (4) is not R^r, (7a) will have nontrivial solution i.e., either $\pi \neq 0$ or $\pi = 0$ and $\pi_0 > 0$ (in fact $\pi \neq 0$, except if $CH = \varnothing$ in (4)). In this case, we choose to "normalize" the system (7a) by adding the inhomogeneous constraint

$$(7b)\qquad \sum_{k=1}^{\tau}\sum_{i=1}^{m_k}\lambda_i^k = 1$$

where D^k has m_k rows, and λ_i^k is the ith component of λ^k. Let Q denote the polyhedron of all points $(\lambda^1, \cdots, \lambda^\tau, \pi, \pi_0)$ that are described by (7a), (7b), the last $(r+1)$ coordinates of these points being (π, π_0).

LEMMA 3. *Suppose that* (4) *is not* R^r, *and that* $\{x|D^kx \geqq 0\} = \{0\}$ *for* $k = 1, \cdots, \tau$. *Then* $Q \neq \varnothing$ *and* Q *has* (*up to positive multiples*) *exactly one nonzero direction of recession, given by* $(\lambda^1, \cdots, \lambda^\tau, \pi, \pi_0) = (0, \cdots, 0, 0, -1)$.

For any valid inequality (3) *for* CH *of* (4) *such that either* $\pi \neq 0$ *or* $\pi = 0$ *and* $\pi_0 > 0$, *there is a valid inequality*

$$(8)\qquad \gamma x \geqq \gamma_0$$

and a scalar $\beta > 0$, *with both* $\gamma = \beta\pi$ *and* $\gamma_0 \geqq \beta\pi_0$, *having the property that* (γ, γ_0) *is a* *convex combination of the projections of extreme points of Q on their last $(r+1)$ coordinates. Moreover, such projections of extreme points of Q yield valid inequalities for* (4).

Proof. We have already seen that $Q \neq \varnothing$, since there are valid inequalities (3) for CH with $\pi \neq 0$, which requires that $\rho = \sum_{k=1}^{\tau} \sum_{i=1}^{m_k} \lambda_i^k > 0$, hence $\rho = 1$ for a suitable positive multiple of (3).

Clearly, $(0, \cdots, 0, 0, -1)$ is a direction of recession for Q. If $(\lambda^1, \cdots, \lambda^\tau, \pi, \pi_0)$ is a direction of recession, it must satisfy $\rho = \sum_{h=1}^{\tau} \sum_{i=1}^{m_k} \lambda_i^k = 0$. Since all $\lambda^k \geqq 0$, we obtain all $\lambda^k = 0$, hence $\pi = 0$ also. Then since $0 \leqq \lambda^k d^k - \pi_0 = -\pi_0$, we have $\pi_0 \leqq 0$. If $\pi_0 \neq 0$, we may take $\pi_0 = -1$ up to a positive multiple.

Next, suppose (3) is a nontrivial valid inequality for CH. Since $\pi \neq 0$ or $\pi = 0$ and $\pi_0 > 0$, we must have at least one $\lambda^k \neq 0$ in (7a). Therefore $\rho > 0$, and, up to a positive multiple β, we may take $\rho = 1$. By the finite Basis Theorem for a pointed polyhedron (see e.g. Rockafellar (1970)), there is a convex combination $(\lambda_c^1, \cdots, \lambda_c^\tau, \gamma, \gamma_0)$ of the extreme points of Q and a multiple $\alpha \geqq 0$ of the direction of recession $(0, \cdots, 0, 0, -1)$ of Q with

$$(9) \qquad (\lambda^1, \cdots, \lambda^\tau, \beta\pi, \beta\pi_0) = (\lambda_c^1, \cdots, \lambda_c^\tau, \gamma, \gamma_0) + \alpha(0, \cdots, 0, 0, -1).$$

Then (γ, γ_0) is a convex combination of the projection of extreme points of Q on their last $(r+1)$ coordinates, $\beta\pi = \gamma$ and $\beta\pi_0 \leqq \gamma_0$.

As we remarked before, the projection of any point of Q on its last $(r+1)$ coordinates is a valid inequality for CH. Q.E.D.

Suppose that a point x^* is given which is not in CH. Since $\pi x^* < \pi_0$ for some valid inequality (3) for CH, and either $\pi \neq 0$ or $\pi = 0$ and $\pi_0 > 0$, by Lemma 3 we may find such a (π, π_0) among the projection of extreme points of Q on its last $(r+1)$-coordinates. To find (π, π_0), one may use the simplex algorithm on (7) to maximize the linear form $(0, \cdots, 0, -x^*, 1) \cdot (\lambda^1, \cdots, \lambda^\tau, \pi, \pi_0) = \pi_0 - \pi x^*$, and stop when any extreme point is reached having $\pi_0 - \pi x^* > 0$.

Although several extreme points $(\lambda^1, \cdots, \lambda^\tau, \pi, \pi_0)$ of Q may have the same projection (π, π_0), all give the same criterion value $\pi_0 - \pi x^*$, so that no two points with the same projection will be encountered during Phase II pivoting (of course, the same degenerate extreme point may be repeated via different basis representations). When many points x^* are to be given, and cut off as in $\pi x^* < \pi_0$, it may be of value to tabulate the (π, π_0)-projection of several of the extreme points encountered during pivoting. Obviously, the natural place to begin pivoting for the successive x^* is from the extreme point where pivoting terminated for the last x^* given. A "Phase I" for this procedure is described in Appendix B.

We can now describe the sets F_i assigned to the initial segment $(h(1), T_1 \cdots, h(i-1), T_{i-1}, h(i))$ of the state $(h(1), T_1, \cdots, h(\theta), T_\theta)$. Let $A^{(j)}x \geqq b^{(j)}$ be the conjunction of the linear inequalities in T_j. For each $k = 1, \cdots, \tau = |S_{h(i)}|$, take, as the set of inequalities $D^k x \geqq d^k$ of (4), the inequalities

$$(10) \qquad Ax \geqq b, \quad A^{(1)}x \geqq b^{(1)}, \quad \cdots, \quad A^{(i-1)}x \geqq b^{(i-1)}, \quad d^{\rho(k)}x \geqq d_0^{\rho(k)}.$$

In (10), ρ is a 1-1 function from $\{1, \cdots, \tau\}$ onto $S_{h(i)}$. Thus, all the systems $D^k x \geqq d_0^k$ are identical, except for the very last inequality $d^{\rho(k)}x \geqq d_0^{\rho(k)}$. Finally, F_i is the set of all (π, π_0)-projections of extreme point solutions to (7).

Note that our hypothesis, that $\{x \mid Ax \geqq b\}$ is nonempty and bounded, gives $\{x \mid Ax \geqq 0\} = \{0\}$ hence also $\{x \mid D^k x \geqq 0\} = \{0\}$ for all $k = 1, \cdots, \tau$.

LEMMA 4. *F_i is finite. If x^* is not in the set*

$$\text{(11)} \qquad \operatorname{clconv}\left(\bigcup_{k=1}^{\tau} \{x | Ax \geqq b, A^{(1)}x \geqq b^{(1)}, \cdots, A^{(i-1)}x \geqq b^{(i-1)}, d^{\rho(k)}x \geqq d_0^{\rho(k)}\}\right)$$

then there is $(\pi, \pi_o) \in F_i$ with $\pi x^* < \pi_0$.

Should the system (7) be employed in connection with (10) as $D^k x \geqq d^k$, and then employed again at a later point of computation with the inequality sets $A^{(j)}x \geqq b^{(j)}$, $j = 1, \cdots, i-1$ enlarged (by addition of cutting-planes between uses of (7)), note that one needs only add new nonbasic columns to (7) that correspond to the added inequalities. In particular, the last used feasible basis for (7) remains a feasible basis, from which pivoting can be renewed to obtain more cutting-planes. The system (7) can be constructed, when needed, from the state (which is always known), and reconstructing the starting solution requires a knowledge of the indices of the columns of the last used basis.

Our next result gives a sufficient condition for x^* not to be in CH of (4), hence for some basic feasible solution of (7) to "cut off" x^*, under the hypotheses of Lemma 3.

PROPOSITION 5. *Suppose that $\{x | D^k x \geqq 0\} = \{0\}$ for $k = 1, \cdots, \tau$. Let x^* be an extreme point of a convex set C which contains the set CH of* (4).

If x^* does not satisfy any of the conditions $D^k x \geqq d^k$ for any $k = 1, \cdots, \tau$, then $x^* \not\in CH$. In particular, the (π, π_0)-projection of some extreme point of (7) *gives $\pi x^* < \pi_0$.*

Proof. If $x^* \in CH$, then for certain points $x^{(1)}, \cdots, x(\tau)$ with $D^k x^{(k)} \geqq d^k$ $(k = 1, \cdots, \tau)$ and multipliers $\lambda_1, \cdots, \lambda_\tau \geqq 0$ we have

$$\text{(12)} \qquad x^* = \sum_{k=1}^{\tau} \lambda_k x^{(k)}, \qquad \sum_{k=1}^{\tau} \lambda_k = 1.$$

Each $x^{(k)} \neq x^*$ and, as $C \supseteq CH$, we have each $x^{(k)} \in C$. From (12), x^* is not an extreme point of C, a contradiction. Q.E.D.

Let us say that a point x^* *fails condition h*, with respect to (1), if x^* does not satisfy any of the inequalities $d^i x \geqq d_0^i$, for $i \in S_h$.

To motivate our intended use of Proposition 5, suppose that the current state is $(h(1), T_1, \cdots, h(\theta), T_\theta)$, where $\bigcup_{j=1}^{\theta} T_j$ represents the "cutting-planes" (i.e., valid linear inequalities) which have been appended thus far to $Ax \geqq b$. Let x^* be an extreme point of the set C of all points satisfying

$$\text{(13)} \qquad Ax \geqq b, \quad A^{(1)}x \geqq b^{(1)}, \quad \cdots, \quad A^{(\theta)}x \geqq b^{(\theta)},$$

and suppose that x^* fails condition h.

We can now apply Proposition 5, taking $D^k x \geqq d^k$ to be

$$\text{(14)} \qquad Ax \geqq b, \quad A^{(1)}x \geqq b^{(1)}, \quad \cdots, \quad A^{(\theta)}x \geqq b^{(\theta)}, \quad d^{\rho(k)}x \geqq d_0^{\rho(k)}$$

where ρ is a 1-1 function from $\{1, \cdots, |S_h|\}$ onto S_h, for certainly C contains CH of (4). Thus we obtain a valid inequality $\pi x \geqq \pi_0$ for (1) with $\pi x^* < \pi_0$. However, adding this inequality to those of (13) may not result in a state if $h \in \{h(1), \cdots, h(\theta)\}$. In fact, if $h = h(i)$ for some $i = 1, \cdots, \theta$ we are permitted to obtain (π, π_0) only from (10) used as $D^k x \geqq d^k$, or else we violate the technical definition of a "state." The system (14) is typically much larger than (10).

The obvious tack to reduce the systems (14) and still maintain $x^* \not\in CH$, is to remove from (13) those inequalities which are slack at x^*, for x^* will still be an extreme point of the resulting inequality system. However, this may still leave some inequalities in $A^{(j)}x \geqq b^{(j)}$, $j \geqq i$, and thus appending $\pi x \geqq \pi_0$ may fail to yield a state. Yet, by a more careful analysis, some of the "tight" (i.e., nonslack) constraints can also be removed, so that $\pi x \geqq \pi_0$ can be added retaining a state, if h is such that k is largest among $\{k | x^*$ fails condition $h(k)\}$.

Of course, these technical issues present no problem in the "easy case" that $h \notin \{h(1), \cdots, h(\theta)\}$. One can then use the entire system (14) as $D^k x \geqq d^k$, by setting $h(\theta+1) = h$ and giving, as the next state, $(h(1), T_1, \cdots, h(\theta), T_\theta, h, T_{\theta+1})$ with $T_{\theta+1} = \{(\pi, \pi_0)\}$.

Similarly, if one is willing to increase the length of the state, the same device can be used for $h \in \{h(1), \cdots, h(\theta)\}$, since two $h(j)$ for different j are permitted to be equal.

PROPOSITION 6. *Suppose that x^* is an extreme point of the set C of all points satisfying* (13) *with state* $(h(1), T_1, \cdots, h(\theta), T_\theta)$, *and put:*

$$(15) \qquad F = \{j \in \{1, \cdots, \theta\} | x^* \textit{ fails condition } h(j)\}.$$

Then x^* is also an extreme point of the polyhedron C' of all points satisfying $Ax \geqq b$ together with all the inequalities

$$(16) \qquad A^{(j)} x \geqq b^{(j)} \quad \textit{for all } j \in F.$$

Remark. A slightly stronger version of Proposition 6 is false. Specifically, if x^* is an extreme point minimizing the linear form (say) cx subject to (13), then x^* need not also minimize cx subject to $Ax \geqq b$ and (16).

In simplex terminology, some of the constraints $A^{(j)}x \geqq b^{(j)}$ for some $j \notin F$ may provide a nonbasic slack variable relative to x^*, and the removal of this nonbasic variable (by a basic variable at a zero level) may allow some reduced cost to change sign.

Proof. Put $S = \{h(1), \cdots, h(\theta)\} \cap F'$, where $F' = \{h(j) | j \notin F\}$. Then for each $h \in S$ there is $k(h) \in S_h$ such that $d^{k(h)} x^* \geqq d_0^{k(h)}$.

Since (1) is facial, x^* lies in the face P of C which is described by the inequalities (13) together with

$$(17) \qquad d^{k(h)} x \geqq d_0^{k(h)} \quad \text{for all } h \in S.$$

Consequently, x^* is an extreme point of P.

To complete our proof, it suffices to show that P has an alternate description in terms of all points satisfying $Ax \geqq b$, (16) and (17). For if this were the case, P is also a face of the polyhedron C'. Since x^* is an extreme point of the extreme set P of C', x^* would be an extreme point of C'. (Here we use the facial nature of the constraints (1)).

To establish this alternate description $Ax \geqq b$, (16), (17), for P, it suffices to show that, if $\sigma x \geqq \sigma_0$ is any inequality in $A^{(j)} x \geqq b^{(j)}$ for any $h(j) \in S$, then the addition of $\sigma x \geqq \sigma_0$ to $Ax \geqq b$, (16), and (17) leaves the set of solutions unchanged. This is done by induction on $i = 1, \cdots, \theta$, but the "ground case" and "induction case" are essentially the same argument.

By induction, the set of all points satisfying (17) and

$$(18) \qquad Ax \geqq b, \quad A^{(1)} x \geqq b^{(1)}, \quad \cdots, \quad A^{(i-1)} x \geqq b^{(i-1)}$$

is the same as the set of points satisfying $Ax \geqq b$, (17), and

$$(19) \qquad A^{(j)} x \geqq b^{(j)} \quad \text{for all } j = 1, \cdots, i-1 \text{ with } j \in F.$$

We need only advance the index from $(i-1)$ to i, for the case $i = \theta$ gives the desired result. Without loss of generality, we may assume $h(i) \in S$.

However, $(\sigma, \sigma_0) \in F_i$, which in particular implies that $\sigma x \geqq \sigma_0$ is valid for all points satisfying (18) and $d^k x \geqq d_0^k$ for some $k \in S_{h(i)}$. Therefore, $\sigma x \geqq \sigma_0$ is valid for all points satisfying (17) and (18), since $d^{k(h(i))} x \geqq d_0^{k(h(i))}$, for all such points x (including $x = x^*$).

By the induction hypothesis, $\sigma x \geqq \sigma_0$ is valid for all points satisfying $Ax \geqq b$, (17) and (19). This completes the inductive step. Q.E.D.

Propositions 5 and 6 have the following consequence.

LEMMA 7. *Suppose that x^* is an extreme point of the set C of all points satisfying* (13) *with state $(h(1), T_1, \cdots, h(\theta), T_\theta)$, and that the set F of* (15) *is nonempty. Let i be the largest index $i \in F$.*

Then there exists $(\pi, \pi_0) \in F_i$ with $\pi x^* < \pi_0$.

Proof. From Proposition 6, x^* is also an extreme point of the set of points satisfying

$$(20) \qquad Ax \geqq b, \quad A^{(1)}x \geqq b^{(1)}, \quad \cdots, \quad A^{(i-1)}x \geqq b^{(i-1)}, \quad A^{(i)}x \geqq b^{(i)},$$

as $j \notin F$ for $j > i$. By Proposition 5, taking $D^k x \geqq d^k$ to be (20) together with

$$(21) \qquad d^{\rho(k)}x \geqq d_0^{\rho(k)}$$

where ρ is a 1-1 function from $\{1, \cdots, \tau = |S_{h(i)}|\}$, onto $S_{h(i)}$, there is a valid inequality (3) for CH with $\pi x^* < \pi_0$.

To insure that one may take $(\pi, \pi_0) \in F_i$, it suffices to show that

$$(22) \qquad \{x | D^k x \geqq d^k\} = \{x | D'^k x \geqq d'^k\}$$

for $k = 1, \cdots, \tau$, where $D'^k x \geqq d'^k$ is the set of all points satisfying

$$(20)' \qquad Ax \geqq b, \quad A^{(1)}x \geqq b^{(1)}, \quad \cdots, \quad A^{(i-1)}x \geqq b^{(i-1)}$$

together with (21). For (22), it in turn suffices to show that, if $\sigma x \geqq \sigma_0$ occurs in $A^{(i)}x \geqq b^{(i)}$, then $\sigma x \geqq \sigma_0$ is satisfied by any solution to (20)′ and (21).

As $(h(1), T_1, \cdots, h(\theta), T_\theta)$ is a state, $(\sigma, \sigma_0) \in F_i$, and hence $\sigma x \geqq \sigma_0$ is valid for all points x which satisfy (20)′ together with $d^k x \geqq d_0^k$ for at least one $k \in S_{h(i)}$. However, $\rho(k) \in S_{h(i)}$, hence indeed $\sigma x \geqq \sigma_0$ is valid for all points which satisfy (20)′ and (21). Q.E.D.

From Lemma 7, one may "cut away" a point x^* (as described there), by transforming from state $(h(1), T_1, \cdots, h(i), T_i, h(i+1), T_{i+1}, \cdots, h(\theta), T_\theta)$ to $(h(1), T_1, \cdots, h(i), T_i \cup \{(\pi, \pi_0)\}, \cdots, h(j), T_j)$ for any $j \geqq i+1$, or to $(h(1), T_1, \cdots, h(i), T_i \cup \{(\pi, \pi_0)\})$. The latter is a state for, as one easily checks (with F'_k denoting the set F_k for the latter) we have $F'_k = F_k$ for $k = 1, \cdots, i$ and $F'_k \supseteq F_k$ for $k = i+1, \cdots, j$. Then we retain $T_j \subseteq F'_j$ for $j = 1, \cdots, \theta$.

We call such changes of state a *change*, and i is called the *index* of the change. We also include, under the conception of a change, the "easy case" in which transition is made from state $(h(1), T_1, \cdots, h(\theta), T_\theta)$ to state $(h(1), T_1, \cdots, h(\theta), T_\theta, h(\theta+1), T_{\theta+1})$, $T_{\theta+1} = \{(\pi, \pi_0)\}$ by increasing the length of the state. The index of the latter kind of change is $(\theta + 1)$.

Our next result is essential to later proofs of finite convergence.

LEMMA 8. *There does not exist an infinite sequence of changes of bounded index.*

Proof. Suppose that there were an infinite sequence of changes, involving (in order) the states $\sigma^1, \sigma^2, \cdots$, such that all indices of changes do not exceed M, where M is some integer. Then there are integers $i \leqq M$ such that i is the index of infinitely many changes; let i^* be the least such i.

For sufficiently large k, all indices of changes in the sequence of states $\sigma^k, \sigma^{k+1}, \cdots$ are not less than i^*. Consequently, all σ^l for $l \geqq k$, have the same elements $h(1), \cdots, h(i^*)$ and $T_1, \cdots, T^*_{i-1}$. The set T^*_i in such σ^γ changes infinitely often, by increasing in size. This contradicts Lemma 1 for $\theta = i^*$. Q.E.D.

2. Main results. We continue use of the terminology and notation of § 1, as well as the standing assumptions, that $\{x|Ax \geqq b\}$ is nonempty and bounded and that (1) is facial. When we discuss algorithms below, it is also assumed that $x \geqq 0$ is among the constraints $Ax \geqq b$, so that the usual equivalence of extreme points and basic feasible solutions holds.

We next describe a game. Put $P_0 = \{x|Ax \geqq b\}$. We use the convention that the empty polytope has no extreme points (which is, in any case, implied by standard usage).

The game proceeds in rounds $j = 0, 1, \cdots$ and has two players. In round j, Player 1 either names Player 2 as the winner (and the game terminates), or Player 1 indicates an extreme point x^* of P_j which does not satisfy (1b). Player 2 then produces a vector (π, π_0) satisfying:

$$\pi x^* < \pi_0 \tag{23a}$$

$$\pi x \geqq \pi_0 \text{ is a valid inequality for}$$

$$P_j \cap \left\{ x \left| \begin{array}{l} \text{there is at least one } i \in S_{h^*} \\ \text{such that } d^i x \geqq d_0^i \end{array} \right. \right\} \tag{23b}$$

where $h^* \in V_j$ and

$$V_j = \{h | d^i x^* < d_0^i \text{ for all } i \in S_h\}. \tag{24}$$

Then we set $P_{j+1} = P_j \cap \{x|\pi x \geqq \pi_0\}$, $j = j+1$, and another round is entered.

THEOREM G. *Player* 2 *will win the game, regardless of the choices of Player* 1, *if he uses the following method to determine* (π, π_0):

—*Initially he sets the current state* $\sigma^0 = \emptyset$.

—*If the current state is* $\sigma^s = (h(1), T_1, \cdots, h(\theta), T_\theta)$ *and if F of* (15) *is nonempty with i the largest index in F then he puts* $h^* = h(i)$ *and he selects* $(\pi, \pi_0) \in F_i$ *satisfying* (23a) *and makes the state change*

$$\sigma^{j+1} = (h(1), T_1, \cdots, h(i), T_i \cup \{(\pi, \pi_0)\}). \tag{25a}$$

Put $j = j+1$.

—*If the current state is* $\sigma^j = (h(1), T_1, \cdots, h(\theta), T_\theta)$ *and if F of* (15) *is empty, then he selects any* $h^* \in V_j$ *and he sets*

$$\sigma^{j+1} = (h(1), T_1, \cdots, h(\theta), T_\theta, h^*, \{(\pi, \pi_0)\}) \tag{25b}$$

where (π, π_0) *satisfies* (23).

Put $j = j+1$.

Proof. It is possible for Player 2 to use the above method, by Proposition 5 and Lemma 7. Note that the length of all σ^j does not exceed t. The result follows by Lemma 8. Q.E.D.

Note that Theorem G still holds if (25a) is replaced by

$$\sigma^{j+1} = (h(1), T_1, \cdots, h(i), T_i \cup \{(\pi, \pi_0)\}, \cdots, h(\theta), T_\theta). \tag{25a$'$}$$

While (25a) allows for removal of accumulated cutting-planes, in dual algorithms this can cause the criterion function to deteriorate in value. The removal indicated in (25a) requires that, if a cut is removed, having currently a nonbasic slack, the slack must be first pivoted into the basis, retaining primal feasibility. In the primal algorithm discussed below, one easily checks that such a slack is pivoted in by a degenerate pivot, so the current solution does not change.

As a compromise between (25a) and (25a)′, when $i < \theta$ one may select an element $(\pi', \pi'_0) \in T_\theta$ and put

$$\text{(25a)}'' \qquad \sigma^{j+1} = (h(1), T_1, \cdots, h(i), T_i \cup \{(\pi, \pi_0)\}, \cdots, h(\theta), T_\theta \backslash \{(\pi', \pi'_0)\}).$$

The choice of (π', π'_0) can be done heuristically e.g. one may choose a $(\pi', \pi'_0) \in T_\theta$ with $\pi' x^* - \pi'_0$ large, or several of them. In this way, one cut has been added and one (or several) removed, hence the total number of cuts has not increased.

Thus, the only time the number of cuts must increase is if $i = \theta$ or one uses (25b). The latter cannot occur more than $(t - \theta)$ successive times. Thus the main cause of cut accumulation, if it occurs, is many repetitions of (25a) with $i = \theta$. When this occurs, and it is viewed that the size of T_θ would exceed desirable bounds, in dual algorithms a branch-and-bound approach can be used, with $d^i x \geqq d^i_0$ for some $i \in S_\theta$ enforced on each subproblem. The state of each subproblem is then $(h(1), T_1, \cdots, h(\theta-1) T_{\theta-1})$, and the method resumes on each.

We now discuss algorithms. These are obtained by using various strategies for Player 1.

The problem to be resolved by the algorithms, denoted *OP*, is the *optimization problem* of minimizing a linear form cx subject to (1), where $c = (c_1, \cdots, c_r)$ is a specified criterion vector.

Dual Algorithm. Choose as x^* any extreme point found after dual simplex re-optimization upon adding the cutting-plane $\pi x \geqq \pi_0$, if x^* fails (1b). If inconsistency results or x^* satisfies (1), declare Player 2 the winner.

It is immediate from Theorem G that Player 2 wins, i.e., the algorithm is finitely-convergent when one uses the procedure discussed in Theorem G.

We next discuss the primal algorithm, which consists of several applications of the *primal subroutine.*

The primal subroutine requires that a *current solution* $\bar{x}$ is given. $\bar{x}$ is to be a solution to (1) which is also an extreme point of the closed convex hull of solutions to (1). By Balas (1974), $\bar{x}$ will then be an extreme point of $\{x | Ax \geqq b\}$, hence there is a basic feasible representation for $\bar{x}$. It is assumed that one has such a representing basis in terms of the inequalities of $Ax \geqq b$ and the cuts so far appended.

Primal Subroutine. If reduced costs show that $\bar{x}$ is optimal, Player 2 wins. Otherwise, choose any non-basic variable to enter the basis with reduced cost for decreasing criterion value. Determine the point x^* obtained if this variable enters the basis under lexicographic pivoting.

—If $x^* = \bar{x}$, repeat this procedure.

—If $x^* \neq \bar{x}$ and x^* satisfies (1), Player 2 wins.

—If $x^* \neq \bar{x}$ and x^* does not satisfy (1), then indicate x^* to Player 2.

The primal subroutine cannot be repeated infinitely often due to $x^* = \bar{x}$, since no cuts are added, and the simplex method will either find an improving solution or indicate optimality of $\bar{x}$. Hence if optimality is not indicated, the case $x^* \neq \bar{x}$ arises. By Theorem G, Player 2 wins, hence if optimality is not indicated during the game, an $x^* \neq \bar{x}$ is found which satisfies (1). Since it is an extreme point of some P_j, and P_j contains the convex hull of solutions to (1), x^* is an extreme point of that convex hull. Hence x^* is also an extreme point of $\{x | Ax \geqq b\}$. Also, one has at hand a representation of x^* as a basic feasible solution to P_j.

For the primal algorithm, an initial current solution $\bar{x} = x^{(0)}$ is given and first $j = 0$.

Primal Algorithm. If the optimality of $x^{(j)}$ is indicated, stop. Otherwise, let x^* be the point found by the Primal Subroutine, which satisfies (1). Put $x^{(j+1)} = x^*$, $j = j + 1$, and repeat.

The primal algorithm clearly is finite, since each $x^{(j)}$ is an extreme point of $\{x|Ax \geqq b\}$ and $cx^{(0)} > cx^{(1)} > \cdots$.

To obtain $x^{(0)} = (y^{(0)}, z^{(0)})$ for (GLC) when $x^{(0)}$ is not otherwise available (say, from complementary pivoting), the following "Phase I" procedure can be used. Let I' be a diagonal matrix of size m by m, with i-th diagonal entry $+1(-1)$ if $h_i \geqq 0$ $(h_i < 0)$. (GLC) is consistent if and only if the following program has optimal value zero:

$$\text{(26a)} \qquad \min \sum_{h=1}^{m} w_k$$

$$\text{(26b)} \qquad \text{subject to } Gy + Hz + I'w \geqq h, \quad y, z, w \geqq 0$$

and

$$\text{(26c)} \qquad y \cdot z = 0.$$

A starting solution for (24) is $(y, z, w) = (0, 0, w^{(0)})$ where $w_i^{(0)}$ is $h_i(-h_i)$ if $h_i \geqq 0$ $(h_i < 0)$, and $w^{(0)} = (w_i^{(0)}, \cdots, w_m^{(0)})$. An extreme point $(y^{(0)}, z^{(0)}, 0)$ of (26b) clearly corresponds to an extreme point $x^{(0)} = (y^{(0)}, z^{(0)})$ of GLC, and when $(y^{(0)}, z^{(0)}, 0)$ is obtained, a basic representation for $x^{(0)} = (y^{(0)}, z^{(0)})$ is available once the basic w_h-variables (which are at value zero) are pivoted out of the basis and the variables of (GLC) (including possibly slack variables) are pivoted in to replace them.

We now relate Theorem G to Balas' result (Balas (1974), Corollary 5.3.1). (A similar analysis can be done for (Balas (1974), Theorem 5.3)).

We put inductively:

$$\text{(27a)} \qquad K_0 = \{x|Ax \geqq b\}$$

$$\text{(27b)} \qquad K_{h+1} = \text{clconv}\left\{ \bigcup_{i \in S_h} (K_h \cap \{x|d^i x \geqq d_0^i\}) \right\}, \qquad 0 \leqq h \leqq t-1.$$

COROLLARY (Balas (1974)). *K_t is the closed, convex span of the points feasible in* (1).

Proof. Let Player 1 use the following strategy. In the order $h = 1, \cdots, t$, Player 1 indicates an extreme point x^* of P_j, if any, such that $d^i x^* < d_0^i$ for all $i \in S_h$. If there are none, Player 2 sets $h = h + 1$ for $h < t$ (and does not return to $h - 1, h - 2, \cdots$ etc.) and stops if $h = t$.

By Theorem G, eventually $h = t$ and Player 2 wins. Let the state held by Player 2 at this time be $\sigma^* = (1, T_1, \cdots, t, T_t)$ and let the polytope be P_j^*.

Suppose there is an extreme point x^* of P_j^* and an index h^* such that $d^i x^* < d_0^i$ for $i \in S_h^*$, and without loss of generality, h^* is as large as possible. By Proposition 6, x^* is also an extreme point of polytope P_j, corresponding to the state $\sigma' = (h(1), T_1, \cdots, h(h^*), T_{h^*})$. Hence earlier in the game Player 1 indicated it to Player 2, and it was cut away. This is a contradiction. Q.E.D.

Theorem G generalizes the corollary, in that Player 1 need not be restricted to the order $h = 1, \cdots, t$ in indicating points, as done in the proof of the corollary. This generalization is necessary to obtain algorithms which always make progress on the optimization problem. For instance, after re-optimization in a dual algorithm, the violated constraint of index h may not be violated by the next solution. If one is required still to add cutting-planes that are based on the h-th constraint, all these would first have to be added before progress could result on the problem at hand.

We now relate Theorem G to our earlier characterization of the valid cutting-planes for (GLC) (Jeroslow (1978)). We showed that these were obtained by repeatedly applying two kinds of rules. One was taking nonnegative combinations of given inequalities (possibly lowering the r.h.s.). The second rules for $j = 1, \cdots, t$ had this form:

"If one has already obtained both

$$\alpha_1 y_1 + \cdots + u y_j + \cdots + \alpha_s y_s + \beta_1 z_1 + \cdots + v z_j + \cdots + \beta_s z_s \geqq \alpha_0 \tag{28a}$$

and

$$\alpha_1 y_1 + \cdots + u' y_j + \cdots + \alpha_s y_s + \beta_1 z_1 + \cdots + v' z_j + \cdots + \beta_s z_s \geqq \alpha_0 \tag{28b}$$

then one may obtain

$$\alpha_1 y_1 + \cdots + u y_j + \cdots + \alpha_s y_s + \beta_1 z_1 + \cdots + v' z_j + \cdots + \beta_s z_s \geqq \alpha_0 ." \tag{28c}$$

Let the current state of a problem (GLC) be $\sigma = (h(1), T_1, \cdots, h(\theta), T_\theta)$ and suppose one has an extreme point (y^*, z^*) with $y_j^* z_j^* > 0$. Further, suppose $h(i) = j$ and i is the largest index having this property. Player 2 cuts away (y^*, z^*) by the following means. He utilizes the system (7) to obtain a cutting-plane with (following (10)) $\tau = 2$. $D^1 x \geqq d^1$ is

$$Ax \geqq b, \quad A^{(1)} x \geqq b^{(1)}, \quad \cdots, \quad A^{(i-1)} x \geqq b^{(i-1)}, \quad -z_j \geqq 0, \tag{29a}$$

and $D^2 x \geqq d^2$ is

$$Ax \geqq b, \quad A^{(1)} x \geqq b^{(1)}, \quad \cdots, \quad A^{(i-1)} x \geqq b^{(i-1)} \quad -y_j \geqq 0. \tag{29b}$$

The inequality ($\pi x \geqq \pi_0$) obtained from (7), which cuts off x^*, is simultaneously a nonnegative combination of the inequalities of (29a), and a nonnegative combination of the inequalities of (29b), by (7a), except that the constant terms may differ. Let (28a) denote that part of the nonnegative combination from (29a) which is contributed by the inequalities in $Ax \geqq b$, $A^{(1)} x \geqq b^{(1)}, \cdots, A^{(i-1)} x \geqq b^{(i-1)}$. Hence $\pi x \geqq \pi_0$ is

$$\alpha_1 y_1 + \cdots + u y_j + \cdots + \alpha_s y_s + \beta_1 z_1 + \cdots + (v + \theta) z_j + \cdots + \beta_s z_s \geqq \alpha_0 \tag{30a}$$

for some nonpositive scalar θ. Similarly, let the inequality from (29b) be denoted

$$\alpha'_1 y_1 + \cdots + (u' + \theta') y_j + \cdots + \alpha'_s y_s + \beta'_1 z_1 + \cdots + v' z_j + \cdots + \beta'_s z_s \geqq \alpha'_0. \tag{30b}$$

Since these inequalities are the same, except possibly for constants, we have

$$\alpha_k = \alpha'_k \quad \text{for } k \neq j, \tag{31a}$$

$$\beta_k = \beta'_k \quad \text{for } k \neq j, \tag{31b}$$

$$u = u' + \theta' \quad \text{and} \quad v' = v + \theta. \tag{32}$$

Without loss of generality, $\alpha_0 = \min\{\alpha_0, \alpha'_0\}$. Hence both (28a) and (28b) will be valid for the current polytope P_j, and obtainable as nonnegative combinations of its defining inequalities. Also, $\pi x \geqq \pi_0$ is simply (28c). In brief, our algorithms do proceed by taking nonnegative combinations, and using the rule (28) of our earlier paper.

An interesting question, which our analysis above does not answer, is this one:

Question. Suppose that Player 1 can indicate, not only x^*, but also the index $h^* \in V_j$ to be used in (23). Does Player 2 still necessarily have a winning strategy?

Blair (1979) answers this question affirmatively and, in the process, provides new methods for establishing finite convergence. That paper also gives an example of nonconvergence when certain seemingly strong cuts are used, which happen to be the "wrong ones."

3. An example. We shall work the following example:

$$\text{(33a)} \qquad \begin{aligned} &\text{minimize} && x_1+4x_2+x_3+2x_4 \\ &\text{subject to} && 2x_1-3x_2-x_3+7x_4 \geqq 20 \\ &&& x_1+4x_2+2x_3+x_4 \leqq 10 \\ &&& x_1, x_2, x_3, x_4 \geqq 0 \end{aligned}$$

and

$$\text{(33b)} \qquad x_1 \leqq 0 \text{ or } x_2 \geqq 2.5 \quad \text{and} \quad x_4 \leqq 0 \text{ or } x_3 \geqq 5.$$

The logical conditions in (33b) are facial, due to the second constraint in (33a) and the nonnegativities; the second constraint forces $x_2 \leqq 2.5$ and $x_3 \leqq 5$.

This example is actually inconsistent, as we shall see. Upon solving (33a) alone, we obtain the solution $(x_1, x_2, x_3, x_4) = (0, 0, 0, 2.857)$.

To implement a dual cutting-plane algorithm, we generate a cut from the condition in (33b), that "$-x_4 \geqq 0$ or $x_3 \geqq 5$." (Recall all constraints are in "$\geqq$" format.) In the notation of equations (7) above, we have $\tau = 2$,

$$\text{(34a)} \qquad D^1 = \begin{bmatrix} 2 & -3 & -1 & 7 \\ -1 & -4 & -2 & -1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -1 \end{bmatrix}$$

and

$$\text{(34b)} \qquad d^1 = [20, -10, 0, 0, 0, 0, 0]^{\text{tr}},$$

where the superscript "tr" denotes transpose; also we have

$$\text{(35a)} \qquad D^2 = \begin{bmatrix} 2 & -3 & -1 & 7 \\ -1 & -4 & -2 & -1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

and

$$\text{(35b)} \qquad d^2 = [20, -10, 0, 0, 0, 0, 5]^{\text{tr}}.$$

(The large identity matrices in D^1 and D^2 of course are due to the nonnegativity constraints in (33a)). The cut obtained in this manner is

$$\text{(36)} \qquad .0879x_1 - .1319x_2 + .1758x_3 - .0330x_4 \geqq .8790.$$

When the cut (36) is added to the constraints of (33a), and the resulting linear program is solved, we obtain the solution $(x_1, x_2, x_3, x_4) = (10, 0, 0, 0)$. We then minimized x_1 subject to (33a) plus (36), and got the same solution. This solution is therefore the unique solution to (33a) plus (36), since the second row in (33a) is constraining.

The solution $(x_1, x_2, x_3, x_4) = (10, 0, 0, 0)$ violates the requirement "$x_1 \leqq 0$ or $x_2 \geqq 2.5$" of (33b). We again used (7), applied to the system of (33a) plus (36), and the cited disjunction, and we obtained the cut (up to a positive multiple)

$$x_1 \leqq 0. \tag{37}$$

Of course, when this cut (37) was appended to (33a) plus (36), the resulting linear program proved to be inconsistent.

Appendix A. Some modeling issues.

A.1. Faithful modeling. Often relations on a vector u of variables are modeled by adding in additional variables z and linear inequalities in u and z. The variables u and z are constrained in some special way, e.g. by integrality or by facial requirements.

The modeling is called "faithful" if, whenever the given relation does hold of u, and z is such that it satisfies the added linear inequalities in u and z, then z automatically also satisfies the special requirements.

For example, consider the relation

$$u_1 + u_2 \geqq 1,\ u_1 \geqq 0,\ u_2 \geqq 0 \quad \text{and either} \quad u_1 = 0 \text{ or } u_2 = 0. \tag{A.1}$$

This relation is easily modeled by adding the facial constraint $u_1 u_2 = 0$ in place of the logical condition. The modeling is clearly faithful, for if $u_1 = 0$ or $u_2 = 0$ does hold, so does $u_1 u_2 = 0$.

The same relation can be modeled with binary variables as follows. One must know an upper bound M on u_1 and u_2, and (A.1) is modeled as

$$\begin{aligned} &u_1 + u_2 \geqq 1, \quad u_1 \geqq 0, \quad u_2 \geqq 0 \\ &u_1 \leqq M z_1, \quad u_2 \leqq M z_2, \quad z_1 + z_2 = 1, \\ &z_1, z_2 \quad \text{are zero or one.} \end{aligned} \tag{A.2}$$

This modeling is not faithful, as one sees by the solution $u_1 = 0$, $u_2 = 1$, $z_1 = z_2 = \frac{1}{2}$, which is in fact an extreme point of the linear inequalities of (A.2) for $M = 2$.

If the cited solution of (A.2) causes computation to proceed merely to make z_1 and z_2 binary, no progress is made toward the problem that is modeled. In branch-and-bound codes, provisions can be entered to avoid arbitrating fractional variables when the relation they were introduced to model already holds. In cutting-plane algorithms however, the rules may require that e.g. z_1 is employed to generate a cut, as when a definite lexicographic ordering is used to insure finiteness of convergence.

A facial modeling need not be faithful; see Example 4 in A.2 below. However, it is faithful in many of the common situations, and an integer modeling is not faithful in most of these.

The issue of the faithfulness of an integer modeling does not arise in cases where the relation to be modeled itself stipulates integers, as e.g. a (whole) number of buses to run on a route. But as one easily shows, any binary-integer modeling can be constructed faithfully as a facial modeling. Thus (BIP) is equivalent to:

$$\begin{aligned} &Dx \geqq d, \quad x_j + x_j' = 1, \quad x_j \text{ and } x_j' \geqq 0 \\ &\text{and} \quad x_j \cdot x_j' = 0 \quad \text{for } j = 1, \cdots, r. \end{aligned} \tag{BIP)$'$}$$

A.2. Examples of modeling. We choose four more-or-less typical examples: many more can be given. An examination of these examples will show, that whatever is modeled by a bounded integer program can be modeled by facial constraints.

Example 1. Fixed charge problems. The relation to be modeled is

$$\text{(A.3)}\qquad \begin{aligned} &y=\begin{cases} a+bx, & \text{if } x>0\\ 0, & \text{if } x=0\end{cases}\\ &x\geqq 0.\end{aligned}$$

Here we assume a, $b\geqq 0$, that y is not otherwise constrained, and that $+y$ appears as a term in a linear criterion function to be minimized.

A modeling for (A.3) is

$$\text{(A.4)}\qquad y+z=a+bx,\quad x\cdot z=0,\quad x\geqq 0,\quad y\geqq 0,\quad z\geqq 0.$$

It is faithful.

Interestingly enough, in the unbounded case, (A.4) is still a model of (A.3), but Meyer has shown that there is no integer modeling for (A.3) when no bound on x is known, if one uses rational quantities (Meyer (1975)).

Example 2. Separable programming. The relation is

$$\text{(A.5)}\qquad \begin{aligned} &y=a+\sum_{i=1}^{p} s_ix_i,\ x_i+v_i=d_i \text{ for } i=1,\cdots,p\\ &x_i\geqq 0,\ v_i\geqq 0;\quad \text{and}\quad \text{also:}\quad \text{if } x_{j+1}>0 \text{ then } x_j=d_j \text{ for } i=1,\cdots,p.\end{aligned}$$

Here a, s_i and d_i are scalars with $d_i>0$. (A.5) represents a piecewise-linear approximation to a function of one variable.

The modeling for (1.5) is obtained by dropping the logical restriction and putting in its place $x_{i+1}v_i=0$ for $i=1,\cdots,p-1$. This modeling is faithful.

Example 3. Minima and maxima; absolute values. The relation to be modeled is

$$\text{(A.6)}\qquad y=\min\{w_1,w_2,\cdots,w_p\}$$

where each w_i may be a linear affine form.

The modeling is

$$\text{(A.7)}\qquad y+z_i=w_i \quad\text{and}\quad z_i\geqq 0 \quad\text{for}\quad i=1,\cdots,p,\quad\text{and}\quad z_1z_2\cdots z_p=0.$$

It is faithful.

Maxima of several affine forms can be handled similarly, and the process can be iterated when e.g. the w_i are maxima of affine forms, etc. Since $|x_j|=\max\{x_j,-x_j\}$, one may consider problems with constraints of the type $|x_1|-|x_2|+3|x_3|=5$, and similar objective functions.

Example 4. Disjunctive value; bounded integers. The relation is

$$\text{(A.8)}\qquad y=w_1\quad\text{or}\quad y=w_2\quad\text{or}\quad\cdots\quad\text{or}\quad y=w_p$$

where the w_i are affine forms.

The modeling is accomplished by

$$\text{(A.9)}\qquad \begin{aligned} &y_i+z_i=w_i,\qquad z_i=z_i^{(1)}-z_i^{(2)},\qquad v_i=z_i^{(1)}+z_i^{(2)},\\ &z_i^{(1)}\geqq 0,\qquad z_i^{(2)}\geqq 0,\quad \text{for } i=1,\end{aligned}$$

$$\text{and}\qquad z_i^{(1)}\cdot z_i^{(2)}=0,\qquad i=1,\cdots,p,\quad v_1v_2\cdots v_p=0.$$

Note that (A.8) is equivalent to $v_i=|z_i|=|y_i-w_i|$, so the modeling is correct. However, it is not faithful, since $y_i=w_i$ can hold without $v_i=|z_i|$ holding (the constraint $z_i^{(1)}\cdot z_i^{(2)}=0$ is not among the linear constraints).

By taking the linear affine form to be integers (e.g. $w_i \equiv (i-1)$), one thus obtains a bounded integer-constrained variable y.

Appendix B. Starting solutions for (GLC); producing redundancy. Suppose that the formula (25b) is to be used. We want a method which is likely to find the desired (π, π_0) in a relatively small number of pivots. This method is also available if one simply desires a valid cutting-plane that removes x^*, and one is not concerned with issues of finite convergence. For simplicity, we treat the case (GLC).

Let $Gx \geqq g$, with $x = (y, z)$, denote those inequalities of P_j whose slacks are nonbasic. G is r by r and nonsingular and x^* is the unique solution to $Gx = g$. Let the ith row of G be denoted $(g^{(i)})^T$ and the ith component of g be denoted g_i.

PROPOSITION 9. *The following matrix B is a feasible basis for* (7), *where the negative unit column $-e^*_{s+h}$ corresponds to the constraint $-z^*_h \leqq 0$ of D^1, and in* (7) (*with the D^k as in* (10)) *we view that $\tau = 2$ and the rows for $k = 1$ are listed above those for $k = 2$*:

$$(\text{B.1}) \qquad B = \left[\begin{array}{cc|cccc|c} g^{(1)} & -e^*_{s+h} & & & & & \\ & & & 0 & & & -I_{r+1} \\ g_1 & 0 & & & & & \\ \hline & & g^{(1)} & g^{(2)} & \cdots & g^{(r)} & \\ & 0 & & & & & -I_{r+1} \\ & & g_1 & g_2 & \cdots & g_r & \\ \hline 1 & 1 & 1 & 1 & \cdots & 1 & 0 \end{array}\right].$$

In the corresponding basic feasible solution, the variables for the first and third columns equal $\frac{1}{2}$, $\pi = \frac{1}{2}g^{(1)}$ and $\pi_0 = \frac{1}{2}g_1$, and all other variables are zero.

Proof. In (B.1) the I_r matrix contained in the upper left corner of each I_{r+1} corresponds to the columns for $\pi = (\pi_1, \cdots, \pi_r)$, and the last column of B is the column of π_0.

The solution described is clearly feasible, and all of its positive columns are in B. It suffices to show that B is invertible.

Let $(\alpha^{(1)}, \beta_1, \alpha^{(2)}, \beta_2, \lambda)$ be a vector of row multiples which gives row sum zero, with $\alpha^{(1)}$ and $\alpha^{(2)}$ r-vectors and β_1, β_2 and λ scalars. We will show that this vector is zero.

From the first column of B, we have

$$(\text{B.2}) \qquad \alpha^{(1)}g^{(1)} + \beta_1 g_1 + \lambda = 0$$

and from the third column we get

$$(\text{B.3}) \qquad \alpha^{(2)}g^{(1)} + \beta_2 g_1 + \lambda = 0.$$

Also from columns $r+3$ to $2r+3$ we obtain

$$(\text{B.4a}) \qquad \alpha^{(2)} = -\alpha^{(1)}$$

$$(\text{B.4b}) \qquad \beta_2 = -\beta_1.$$

Substituting (B.4) in (B.3) and then adding (B.2) we obtain $2\lambda = 0$, so $\lambda = 0$.

Now columns 3 through $r+2$ give

$$(\text{B.5}) \qquad \alpha^{(2)}g^{(i)} = -\beta_2 g_i, \qquad i = 1, \cdots, r.$$

If $\beta_2 = 0$, then from (B.5) and the nonsingularity of G, $\alpha^{(2)} = 0$. Then, (B.4) give $\alpha^{(1)} = 0$ and $\beta_1 = 0$, and we are done.

Suppose now $\beta_2 \neq 0$. Then since x^* is the unique solution to $Gx = g$, (B.5) gives $x^* = -\alpha^{(2)}/\beta_2$. The second column of B now gives $\alpha^{(1)}_{s+h}{}^* = 0$, and then with (B.4a) we find $x^*_{s+h^*} = z^*_h = 0$. However, by the choice of h^* for (25b), $z^*_h > 0$. This is a contradiction. Q.E.D.

The basic feasible solution of Proposition 9 has $\pi_0 - \pi x^* = 0$, and so, unless the degeneracy of the solution is a serious problem, one expects to obtain $\pi_0 - \pi x^* > 0$ in a few pivots.

In regard to the basic (B.1) for (7), it is particularly worthwhile to test if the vector

$$\text{(B.6)} \qquad \left[\begin{array}{c} 0 \\ 0 \\ \hdashline -e^*_h \\ 0 \\ \hdashline 1 \end{array}\right]$$

can be brought into the basis to make $\pi_0 - \pi x^* > 0$. If this can be done and (GLC) permits $z^*_{h+s} > 0$, then the first and second columns of (AII.1) remain basic, and so by (7) there is a multiplier $\theta \geqq 0$ with

$$\text{(B.7a)} \qquad \pi_i = \theta g^{(1)}_i, \quad i \neq h^*$$

$$\text{(B.7b)} \qquad \pi^*_h < \theta g^{(1)}_{h*}$$

$$\text{(B.7c)} \qquad \pi_0 = \theta g_1.$$

By (B.7), $\pi x \geqq \pi_0$ makes the problem constraint $g^{(1)}x \geqq g_1$ redundant.

If one is not concerned with finite convergence, it is often of value to produce redundant constraints, in the sense of (B.7), wherever possible, since this corresponds to "tightening" a problem constraint instead of "adding" a cutting-plane.

Geometrically, creating redundancy, by the method just described, involves "rotating" the constraint $g^{(1)}x = g_1$ about its intersection with $z^*_h = 0$, the rotation being toward the feasible set. This appears to account for most of the redundancies in cutting-planes observed in connection with (GLC). Of course, such a rotation is not always possible.

It is also worth noting that the "tightening" constraint $\pi x \geqq \pi_0$ of (B.7) has the same nonzero components as the problem constraint $g^{(1)}x \geqq g_1$ if $g^{(1)}_{h*} \neq 0$, and it has only one more nonzero component if $g^{(1)}_{h*} = 0$. Thus, sparseness is preserved.

REFERENCES

E. BALAS (1974), *Disjunctive programming: Properties of the convex hull of feasible points*, MSRR no 348, GSIA, Carnegie–Mellon University.

C. E. BLAIR (1976), *Two Rules for Deducing Valid Inequalities for* 0-1 *Problems*, SIAM J. Appl. Math., 31, pp. 614–617.

——— (1979), *Disjunctive programs and sequences of cutting-planes*, no. 576, Department of Business Administration, University of Illinois, Urbana.

C. E. BLAIR AND R. G. JEROSLOW (1978), *A Converse for Disjunctive Constraints*, J. Optimization Theory Appl. 25, pp. 195–206.

V. CHVÁTAL (1973), *Edmonds Polytopes and a Hierarchy of Combinatorial Problems*, Discrete Math. 4, pp. 305–337.

R. W. COTTLE AND G. B. DANTZIG (1968), *Complementary pivot theory of mathematical programming*, Linear Algebra and Its Appl. 1, 103–125.

B. C. EAVES (1971), *The linear complementarity problem*, Management Sci., 17, pp. 612–634.

F. GLOVER (1975), *Polyhedral annexation in mixed integer and combinatorial programming*, Math. Programming 9, pp. 161–188.

R. G. JEROSLOW (1978), *Cutting-planes for complementarity constraints*, this Journal, 16, pp. 56–62.

——— (1976), *Cutting-planes for complementarity constraints*, Notices Amer. Math. Soc., 23, A-364.

H. KONNO (1976), *A Cutting-plane algorithm for solving bilinear programs*, Math. Programming, 11, pp. 14–27.

C. E. LEMKE (1965), *Bimatrix equilibrium Points and Mathematical Programming*, Management Sci., 11, pp. 681–689.

C. E. LEMKE AND J. T. HOWSON (1964), *Equilibrium Points of Bimatrix Games*, J. Soc. Indust. Appl. Math., 12, pp. 413–423.

R. R. MEYER (1975), *Integer and Mixed Integer Programming Models*, J. Optimization Theory Appl., 16, pp. 191–206.

R. T. ROCKAFELLAR (1970), *Convex Analysis*, Princeton University Press, Princeton, NJ.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 3, May 1980

EFFICIENT ALGORITHMS FOR A SELECTION PROBLEM WITH NESTED CONSTRAINTS AND ITS APPLICATION TO A PRODUCTION-SALES PLANNING MODEL*

ARIE TAMIR†

Abstract. The following problem is considered. Given positive integers $(n_1, \cdots, n_m)$ and an $n_m \times m$ matrix D with the property that the n_m elements in each column form a monotone sequence, find a set A of n_m elements of D whose sum is maximum, and such that for any j, $j = 1, \cdots, m$, not more than n_j elements are chosen from columns $1, 2, \cdots, j$. An algorithm, solving the above problem in time $O(m^2 \log^2 n_m)$ is presented. The algorithm is applicable to a production-sales planning model with concave utilities. It is also demonstrated that the special case of equal columns is solvable in $O(m^2)$ time.

1. Introduction. A manufacturer producing a single, indivisible, product is faced with the following planning problem. Given the finite horizon $[0, T]$, items of the product can be sold only at m given times $0 < t_1 < t_2, \cdots, < t_m = T$. The utility of selling k items at time t_j, $j = 1, \cdots, m$ is a monotone nondecreasing concave function denoted by $U_j(k)$. It is assumed that (i) the production rate is constant and time independent with a denoting the time to produce one item; (ii) no inventory costs are incurred; and (iii) no initial stock is available. The problem is to find the production-sales scheme that maximizes the total utility over the given horizon. Using the above assumptions we focus only on production schedules that begin at $t = 0$ and are not idle until the last unit is produced.

Let x_j be the integer number of units sold at time t_j, $j = 1, \cdots, m$, then consider the following formulation,

Problem 1.

$$\text{Maximize} \quad \sum_{j=1}^{m} U_j(x_j) \quad \text{subject to}$$

$$(1) \qquad a \sum_{i=1}^{j} x_i \leqq t_j, \qquad x_j \geqq 0 \text{ integer}, j = 1, \cdots, m.$$

Using $[y]$ to denote the largest integer less than or equal to y, we define $n_j = [t_j/a]$. The feasible set becomes

$$(2) \qquad P = \left\{ x = (x_1, \cdots, x_m) | x_j \geqq 0 \text{ integer}, \sum_{i=1}^{j} x_i \leqq n_j, j = 1, \cdots, m \right\}$$

where $n_1 \leqq n_2 \leqq \cdots \leqq n_m$. We assume that $n_1 \geqq 1$. The well known problem of optimizing the distribution of effort [1], [2], [3], [5], [6], [7], [8] is obtained as a special case when $n_1 = n_2 = \cdots = n_m$, since then the inequalities $\sum_{i=1}^{j} x_i \leqq n_j$, $j = 1, \cdots, m-1$, are implied by the one corresponding to $j = m$. Following the dynamic programming procedure in [6], [8] for solving the above special case, one can easily extend it to solve Problem 1. However, this dynamic programming routine, which takes $O(mn_m^2)$ effort, i.e. $O(mn_m^2)$ comparisons and function evaluations, does not use the concavity properties of the utilities. The goal of this study, which is applicable only to the concave case, is to derive solution procedures which are more efficient than the above bound for the case when n_m, the total effort, is significantly larger than m.

Our study is based on the observation that the above problem is equivalent to the following restricted selection problem.

* Received by the editors June 12, 1978, and in final revised form July 17, 1979.

† Department of Statistics, Tel Aviv University, Tel Aviv, Israel.

Define the matrix $D=(d_{ij})$, where $d_{ij}=U_j(i)-U_j(i-1), j=1,\cdots,m, i=1,\cdots,n_m$.

Problem 2. Find a set, A, of n_m elements of the matrix D whose sum is maximum, and such that for any $j, j=1,\cdots,m$, not more than n_j elements are chosen from columns $1,2,\cdots,j$.

In the context of Problem 1, d_{ij} is the marginal contribution of the ith unit sold at time t_j. In particular, $U_j(x_j)-U_j(0)$ is the sum of the first x_j entries in column j of D.

The idea of using incremental analysis for solving Problem 1 is taken from [1], [3], where the same approach is used for solving the special case when $n_1=n_2=\cdots=n_m$. We have the following theorem, motivating Problem 2.

THEOREM 1. *$x=(x_1,\cdots,x_m)$ is an optimal solution to Problem 1 if and only if an optimal solution to Problem 2 is defined by choosing x_j entries in column j of D, $j=1,\cdots,m$.*

Two efficient algorithms of polynomial time complexity are presented. The first method is of time $O(mn_m)$ while the second procedure which is much more complex and utilizes the sophisticated technique (for the unrestricted selection problem) given in [2], yields the bound $O(m^2\log^2 n_m)$. The latter procedure, being sublinearly bounded in n_m, is recommended in those circumstances where n_m is significantly larger than m.

The organization of the paper is as follows. In the next section we prove the equivalence of Problems 1–2 and investigate properties of optimal solutions. In § 3 the above mentioned algorithms are presented, while § 4 treats the special case where the utility functions $U_j(x_j), j=1,\cdots,m$, are all identical. This special case is solved in $O(m^2)$ time.

2. Properties of optimal solutions. We start by proving Theorem 1.

First, note that the concavity of the functions $U_j(x_j)$ implies $d_{ij}\geqq d_{i+1,j}$ for all $j=1,\cdots,m$ and $i=1,\cdots,n_m-1$. Hence, one may assume that if x_j entries are chosen from column j in an optimal solution to Problem 2, these are the first x_j entries in this column. In particular, the sum of these x_j elements is $U_j(x_j)-U_j(0)$. Furthermore, the constraints stating that for any j no more than n_j elements are chosen from columns $1,2,\cdots,j$ are exactly the constraints of Problem 1 as expressed by (2). Problem 2 is now to find $x=(x_1,\cdots,x_m)$ in the set P defined by (2) such that $\sum_{j=1}^m(U_j(x_j)-U_j(0))$ is maximized. The latter is clearly equivalent to Problem 1, and the proof of Theorem 1 is complete.

As a corollary of Theorem 1 we see that the special case of $n_1=n_2=\cdots n_m$ is solved by finding the largest n_m elements in D.

We use the following definitions and notation.

Let $A\subseteq\{(i,j)|i=1,\cdots,n_m; j=1,\cdots,m\}$ be a set of cells and let x_j be the number of elements in A with right index j. We consider only those sets, A, such that

(i) $A=\{(i,j)|i=1,\cdots,x_j; j=1,\cdots,m\}$. A is said to be feasible if

(ii) $x=(x_1,\cdots,x_m)$ is in the set P defined by (2).

A feasible set A is optimal if $x=(x_1,\cdots,x_m)$ solves Problem 1. Similarly, given $x=(x_1,\cdots,x_m)$ in P, we say that it consists of the cells $A=\{(i,j)|i=1,\cdots,x_j; j=1,\cdots,m\}$. A set of cells B is said to be contained in the feasible solution x if $B\subseteq A$.

A set of cells, A, is p-largest if it consists of p cells and the (multi) set $d(A)=\{d_{ij}|(i,j)\in A\}$ is a set of p largest elements of D. (Note that due to possible ties a p-largest set is not necessarily unique).

Given a p-largest set A, let w be a smallest element in $d(A)$. Define $r(A)=(r_1,\cdots,r_m)$, where r_j, $(j=1,\cdots,m)$, is given by $r_j=|\{i|(i,j)\in A, d_{ij}=w\}|$. A is p-lexicolargest if $r(A)$ is lexicographically larger than $r(B)$ for all p-largest sets B, $B\neq A$.

($(u_1, \cdots, u_m)$ is lexicographically larger than $(v_1, \cdots, v_n)$ if for some j $v_j < u_j$ and $v_k = u_k$ for all $k > j$.) Given an integer p, a p-lexicolargest set is unique, and is denoted by A_p.

Solving Problems 1–2 now amounts to finding a feasible set A such that $\sum_{(i,j)\in A} d_{ij}$ is maximum.

Feasible p-largest, and p-lexicolargest sets play a key role in deriving our algorithm. Following are several of their properties.

PROPOSITION 1. *Given Problems* 1–2 *with a matrix* $D = (d_{ij})$ *and the set of bounds* $(n_1, n_2, \cdots, n_m)$, $n_1 \leqq n_2 \cdots \leqq n_m$, let d_{1j} *be a first row entry which is a largest element of the matrix D. If* $n_j \geqq 1$, *then there exists an optimal solution containing the cell* $(1, j)$.

Proof. Let x be an optimal solution and suppose that $(1, j)$ is not part of this solution. Hence $x_j = 0$. If $x_i \geqq 1$ for some $i < j$ define the solution y by $y_k = x_k$, $k \neq i, j$, $y_i = x_i - 1$ and $y_j = 1$. d_{1j} being a largest element yields the optimality of y for Problem 1. Thus assume $x_i = 0$, $i \leqq j$, and let $t > j$ be the smallest index such that $x_t \geqq 1$. Define the solution y by $y_k = x_k$, $k \neq j, t$, $y_t = x_t - 1$ and $y_j = 1$. Again, the feasibility and optimality of y is easily observed.

PROPOSITION 2. *Given Problems* 1–2 *with a matrix* $D = (d_{ij})$ *and the set of bounds* $(n_1, n_2, \cdots, n_m)$, $n_1 \leqq n_2 \cdots \leqq n_m$, *let A be a feasible p-largest set. Then there exists an optimal solution containing the set A.*

Proof. Suppose that the claim is true for feasible $(p-1)$-largest sets of all matrices D and bounds $(n_1, n_2, \cdots, n_m)$. Proposition 1 ensures its validity for $p = 1$. Let d_{1j} be a largest element in $d(A)$. By Proposition 1 there exists an optimal solution x that utilizes the cell $(1, j)$. Thus, due to the separability of Problem 1, the existence of the above x implies the sufficiency of solving Problems 1–2 with $D' = (d'_{i,v})$ and $(n'_1, n'_2, \cdots, n'_m)$, where $d'_{iv} = d_{iv}$ for $v \neq j$, $d'_{ij} = d_{i+1,j}$ for $i \geqq 1$ and $n'_k = n_k - 1$, $k \geqq j$, $n'_{j-1} = \min(n'_j, n_{j-1})$, $n'_{j-2} = \min(n'_{j-1}, n_{j-2})$, $\cdots$, $n'_1 = \min(n'_2, n_1)$. It is clearly observed that

$$A' = \{(i, v) | (i, v) \in A, v \neq j\} \cup \{(i, j) | (i+1, j) \in A, i \geqq 1\}$$

is a feasible $p-1$ largest set for the problem defined by D' and $(n'_1, n'_2, \cdots, n'_m)$. Thus, applying the induction hypothesis yields an optimal solution containing A' and the proposition follows.

Since a p-largest set is not necessarily unique one can easily construct examples where both feasible and infeasible p-largest sets are present. For our purposes this difficulty is resolved by the next proposition.

PROPOSITION 3. *Given* $D = (d_{ij})$ *and* $(n_1, \cdots, n_m)$, $n_1 \leqq n_2 \cdots \leqq n_m$, *if there exists a feasible p-largest set, then the p-lexicolargest set is feasible.*

Proof. Let $x = (x_1, \cdots, x_m)$ be a feasible solution corresponding to a feasible p-largest set A and let $z = (z_1, \cdots, z_m)$ be the vector corresponding to the p-lexicolargest set. Let $m(x)$ be the number of indices i such that $z_i \neq x_i$. Our proof is by induction on $m(x)$. Let j be such that $z_j > x_j$ and $x_k = z_k$, $\forall k > j$. Since $\sum_{i=1}^m x_i = \sum_{i=1}^m z_i$, there exists u such that $z_u < x_u$. Moreover, since both x and z correspond to sets containing p-largest elements it follows that $d_{rj} = d_{qu}$ for $x_j < r \leqq z_j$, $z_u < q \leqq x_u$. Define a solution y corresponding to a set of p-largest elements by the following $y_k = x_k$, $k \neq u, j$, $y_u = x_u - a$ and $y_j = x_j + a$, where $a = \min(z_j - x_j, x_u - z_u)$. It is easily verified that $m(y) \leqq m(x) - 1$ and that y is feasible for Problem 1. By the induction hypothesis our proof is now complete.

3. The Algorithms.

Proposition 1 validates the following algorithm.

ALGORITHM 1.

Step 1. Set $x_j = 0$, $j = 1, \cdots, m$. Let $AC = \{1, \cdots, m\}$ be the set of active columns; then set $E = \{d_{1j} | j \in AC\}$. Also set $r = 0$.

Step 2. Find the largest element in E, say d_{ij}, an element from column j. Increase x_j by 1.

Step 3. For every $k \geqq j$ decrease n_k by 1. Starting with n_{j-1}, replace n_t by $\min(n_t, n_{t+1})$, $t = j-1, j-2, \cdots, r+1$.

Step 4. If $n_{r+1} \geqq 1$ go to Step 5. Let r be the largest index such that $n_r = 0$. If $r = m$ terminate, otherwise delete the indices in AC which are smaller or equal to r.

Step 5. If $j \leqq r$ go to Step 2, otherwise replace d_{ij} in E by $d_{i+1,j}$ and go to Step 2.

It is easily verified that the time of the algorithm is determined in Step 3, which consumes $O(m)$ time per iteration. Since n_m iterations are performed the algorithm takes $O(n_m m)$ time. It should be remarked that several steps (Step 3 not included) of the algorithm can be improved upon to save time. But this does not affect the complexity bound. For example, in Step 2 we may start by first sorting all the elements in the first row of the matrix D (using only $O(m \log m)$ time). Then every time an element is replaced in E, (Step 5), the sorted set is updated in $O(\log m)$ time. Since the bound $O(n_m m)$ is not reduced we prefer to avoid further elaboration and present the second algorithm which is sublinearly bounded in n_m.

We use Propositions 2–3 to derive an algorithm whose time complexity is sublinear in n_m. We assume $n_m > m$, since otherwise Algorithm 1 will have a better asymptotic bound.

Start by finding the largest integer $r \leqq n_m$ such that the r-lexicolargest set A_r is feasible. By Proposition 2 there exists an optimal solution to Problems 1–2 that contains the cells of A_r. The maximality of r also implies that for some j the equality holds in the constraint $\sum_{i=1}^{j} x_i \leqq n_j$. (Consider the largest index with this property). Since, otherwise, if the strict inequality holds for all j, one more cell can be augmented to the solution to yield a feasible $(r+1)$ largest set. Proposition 3 then yields the contradiction. Thus the first j columns of D can be omitted from further consideration and x_i, $i \leqq j$, is determined by the number of cells in A_r, whose column index is i. Then, subtract n_j from n_k for all $k > j$, and find the largest index $r \leqq n_m - n_j$ such that the r-lexicolargest elements in $\{d_{uv} | u = 1, \cdots, n_m - n_j, v = j+1, \cdots, m\}$ satisfy the constraints

$$(3) \qquad \sum_{i=j+1}^{v} x_i \leqq n_v - n_j, \qquad v = j+1, \cdots, m.$$

The process is then continued with the above index j now being replaced by the (largest) index v such that equality holds in (3). The entire procedure is repeated until finally n_m elements are chosen. We label the above algorithm as Algorithm 2.

To compute the complexity of this algorithm we first see that the above procedure is not iterated more than m times. This is implied by the fact that at each iteration we omit at least one additional column of D from further consideration. To evaluate the computational effort spent at each iteration we focus on the first one first.

Given $q \leqq n_m$ and $j \leqq m$, we denote by $f(q, j)$ the time of finding the q-largest element in a matrix with q rows and j columns. (Monotonicity in the columns of the matrix is assumed). To compute the largest $r \leqq n_m$ required in the first iteration, we find the n_m-lexicolargest elements of D, and then apply a binary search on this set of elements to find r. Given a q-largest element of D, it takes $O(m \log q)$ to construct the set of q-lexicolargest elements of D. Since we compute at most $\log n_m$ lexicolargest sets and perform a feasibility test requiring only $O(m)$ time, for each one of them, the time complexity of the first iteration amounts to $O(f(n_m, m) \log n_m + m \log^2 n_m)$. (Note that finding the largest index j for which equality holds in the jth constraint is included in the feasibility test.) A similar analysis shows that in the second iteration the bound reduces to $O(f(n_m - n_j, m - j) \log(n_m - n_j) + (m - j) \log^2(n_m - n_j))$, where $j \geqq 1$ has been found in the first iteration. Now, recalling that the number of iterations is bounded by m, we conclude that performing the algorithm takes time $O(m(f(n_m, m)\log n_m + m\log^2 n_m))$. Finally, to evaluate a q-largest element of a matrix with column monotonicity (and hence the q-lexicolargest set) we suggest the selection algorithm of [2], which has $f(q, j) = O(j \log q)$. This yields the bound of $O(m^2 \log^2 n_m)$ for our algorithm. Note that this bound is expressed in terms of function evaluations when applied to Problem 1.

Remark. Considering the first iteration of Algorithm 2, it is pointed out that only n_j of the r-lexicolargest elements are used. Thus, if the remaining $r - n_j$ elements are recorded, some computational effort can be saved when we compute the lexicolargest set corresponding to the second iteration. This routine may be repeated at any iteration, but the complexity bound, being based on the worst possible case, is not improved.

4. The symmetric case. In this section we elaborate on the case when the utility functions $U_j(x)$ are equal, i.e. the matrix D has equal columns, and present a procedure to solve Problem 1 in time which is independent of n_m. Specifically the problem is solved in time $O(m^2)$. The procedure is motivated by Proposition 2, which suggests the construction of feasible p-largest sets as a possible direction for solution.

We start by introducing a routine to reduce the set of feasible solutions defined in (2). Given $(n_1, \cdots, n_m)$ we assume that $n_j \geqq 1, j = 1, \cdots, m$. (Otherwise omit the columns $k = 1, \cdots, j$ from further consideration and set $x_k = 0$, $k = 1, \cdots, j$.) Suppose that for some j, n_j is less than the number of variables appearing in the jth constraint, i.e. $n_j < j$. Hence at most n_j variables in the set $\{x_1, \cdots, x_j\}$ can take on positive integer values. Since the matrix D has equal columns, and due to the nested structure of the set in (2), we assume with no loss of generality that $x_1 = x_2 = \cdots = x_{j-n_j} = 0$, thus reducing the original m-variable problem into a $(m - j + n_j)$-variable one. Proceed in the same way with the reduced problems until no further reductions are possible. Since we can have at most m reductions the reduction routine terminates in time $O(m)$.

The following algorithm is applicable for the symmetric case. Note that it depends only on the equality of the columns but not on the specific values of the elements in D.

ALGORITHM 3.

Step 0. Let $(n_1, \cdots, n_m)$ be given and set $x_j = 0, j = 1, \cdots, m$.

Step 1. Apply the reduction routine to the current problem. If the jth variable is set to zero in this process, no further elements of D are selected in the jth column. Let $\{v + 1, \cdots, m\}$ be the indices of the remaining columns. (Stop if no columns remain).

Step 2. Let k be such that

$$\frac{n_k}{k - v} = \min\left\{\frac{n_j}{j - v};\, j = v + 1, \cdots, m\right\}, \text{ and define } t = \left[\frac{n_k}{k - v}\right] \geqq 1.$$

Step 3. For $j = v + 1, \cdots, m$ increase x_j by t.

Step 4. For $j = v + 1, \cdots, m$ subtract $(j - v)t$ from n_j and go to Step 1.

The validity of the above procedure is ensured by Proposition 2 and the equality of the columns of D. Also note that due to the reduction routine it is not necessary to maintain the monotonicity of the sequence $\{n_j\}$, as is done in Algorithm 1.

To find the time complexity of the procedure we first bound the number of iterations. We show that at each iteration at least one additional column is dropped from further consideration. Consider the index k defined in Step 2. If $n_k/(k - v)$ is integer then the updating performed in Step 4 yields $n_k = 0$, which in turn implies that the columns indexed $v + 1, \cdots, k$ are omitted. Thus suppose that $n_k/(k - v)$ is not integer. Then in Step 4 n_k is replaced by $n'_k = n_k - (k - v)[n_k/(k - v)] < n_k - (k - v)(n_k/(k - v) - 1) = k - v$. Therefore, the reduction routine is applied for $j = k$

since $n'_k/(k-v)<1$, and at least one more column is omitted from further discussion. Steps 1–4 are of time $O(m)$ and thus the entire procedure is of time complexity $O(m^2)$.

A comment is in order. Although it is not directly related to Problem 1, we point out that the integer vector $(x_1, \cdots, x_m)$, generated by Algorithm 3 is majorized (in the sense of Hardy, Littlewood, Polya [4]) by any other integer vector in the set defined by (2). In fact, this claim is implied by the independence of the algorithm on the specific values taken by the common utility.

Note added in proof. We have recently found another way to apply the algorithm in [2] to solve Problem 1. This application yields the bound $O(m^2 \log n_m)$.

REFERENCES

[1] B. Fox, *Discrete optimization via marginal analysis*, Mgt. Sci., 13 (1966), pp. 210–216.

[2] Z. Galil and N. Megiddo, *A fast selection algorithm and the problem of optimum distribution of effort*, J. Assoc. Comput. Mach., 26 (1979), pp. 58–64.

[3] O. Gross, *A class of discrete-type minimization problems*, RM-1644-PR, Rand Corp., Feb. 1956.

[4] G. H. Hardy, J. E. Littlewood and G. Polya, *Inequalities*, Cambridge University Press, London, England, 1934.

[5] E. P. C. Kao, *On incremental analysis in resource allocations*, Operational Res. Quart., 27, 3ii (1976), pp. 759–763.

[6] W. Karush, *A general algorithm for the optimal distribution of effort*, Mgt. Sci., 9 (1962), pp. 50–72.

[7] L. G. Proll, *Marginal analysis revisited*, Operational Res. Quart., 27, 3ii (1976), pp. 765–767.

[8] H. M. Wagner, *Principles of Operations Research*, Prentice-Hall, Englewood Cliffs, NJ, 1969.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 3, May 1980

0363-0129/80/1803-0004$01.00/0

NONSINGULAR FACTORS OF POLYNOMIAL MATRICES AND (A, B)-INVARIANT SUBSPACES*

E. EMRE†

Abstract. Given a polynomial matrix $B(s)$, we consider the class of nonsingular polynomial matrices $L(s)$ such that $B(s)=R(s)L(s)$ for some polynomial matrix $R(s)$. It is shown that finding such factorizations is equivalent to finding (A, B)-invariant subspaces in the kernel of C where A, B, C are linear maps determined by $B(s)$. In particular, the results yield, as a corollary, a method to determine simultaneously a row proper greatest right divisor of a left invertible polynomial matrix as well as the resulting polynomial matrix whose greatest right divisors are unimodular.

The results also relate, the same way, such subspaces of constant systems $(\bar{C}, \bar{A}, \bar{B})$ where $(\bar{C}, \bar{A})$ is observable, to the nonsingular right factors of the numerator polynomial matrices in factorizations of the form $D^{-1}(s)B(s)$ of their transfer matrices.

1. Introduction. Factorization of a polynomial matrix $B(s)$ has been a subject of several authors both in mathematics and system theory literature [1]–[3], [8]–[14]. In [12]–[14], $B(s)$ has been assumed to be square and monic (i.e., highest degree coefficient matrix is unit matrix), and only monic factors of $B(s)$ have been considered.

In [1]–[3], [8]–[11], $B(s)$ has been taken to be a left invertible polynomial matrix [1]–[3] and the main purpose has been the extraction of a greatest right divisor and obtaining the remaining factor as a polynomial matrix with all unity invariant factors [1]–[3].

In this paper, we consider a general polynomial matrix $B(s)$ with coefficients of its entries in a field, and its nonsingular right polynomial divisors (NRD) $L(s)$ (i.e., factorizations of the form $B(s)=R(s)L(s)$ for some polynomial matrix $R(s)$ such that $\det(L(s))\not\equiv 0$). Motivated by the results of [4] on exact matching, it is shown in § 2 that such factorizations are equivalent to finding (A, B) invariant subspaces in the kernel of C [5]–[7], where A, B, C are linear maps determined by $B(s)$. Every NRD yields such a subspace and, once such a subspace is found, it is shown that corresponding $L(s)$ can be found (in row proper form [1]–[3]). In particular, the results of the paper yield a method to determine a row proper greatest right divisor of a left invertible polynomial matrix as well as a resulting polynomial matrix which is a left factor whose invariant factors are all unity. Here we consider only the case of right factors because the case of nonsingular left factors can be approached by duality.

Finally, it is shown that if $(\bar{A}, \bar{B}, \bar{C})$ is any observable system, then the NRD's of $B(s)$ in a factorization [1]–[3] of $\bar{C}(sI-\bar{A})^{-1}\bar{B}$ as $D^{-1}(s)B(s)$ are related in the same way to $(\bar{A}, \bar{B})$-invariant subspaces in the kernel of $\bar{C}$.

The notation is such that the maps and their matrix representations are denoted by the same symbols and for a matrix R, $\{R\}$ denotes the span of the columns of R. If A is a linear map and ψ is an A-invariant subspace, $A \mid \psi: \psi \to \psi$ denotes the restriction of A to ψ. By a basis matrix for a subspace ψ we mean a matrix R whose columns are a basis for ψ. Ker C denotes the kernel of the mapping C.

2. Nonsingular right factors and (A, B)-invariant subspaces. Let $B(s)$ be an $f\times r$ polynomial matrix.

DEFINITION 1. An $r\times r$ polynomial matrix $L(s)$ is said to be a *nonsingular right divisor* of $B(s)$ (NRD) iff

* Received by the editors June 3, 1978, and in revised form August 7, 1979.

† Center for Mathematical System Theory, Department of Electrical Engineering, University of Florida, Gainesville, Florida 32611. This research was supported in part by the Department of Mathematics of Eindhoven University of Technology.

(1) $\det(L(s))$ is nonzero, and

(2) there exists an $f \times r$ polynomial matrix $R(s)$ such that

$$B(s) = R(s)L(s).$$

PROPOSITION 1 [1]–[3]. *If $L(s)$ is an $r \times r$ nonsingular polynomial matrix, then there exists a unimodular polynomial $M(s)$ and a row proper matrix $\bar{L}(s)$ such that*

$$M(s)L(s) = \bar{L}(s). \tag{1}$$

In general the polynomial matrices $M(s)$ and $\bar{L}(s)$ satisfying (1) *are not necessarily unique.*

However, if v_i is the degree of the i-th row of an $\bar{L}(s)$ as in (1), *then the set $\{v_1, \cdots, v_r\}$ is the same (modulo the ordering of v_i's) for every $\bar{L}(s)$ as in* (1).

It follows from Definition 1 and Proposition 1 that, if $L(s)$ is a NRD of $B(s)$, then the elements of the set

$$S_L = \{M(s)L(s) \,|\, M(s) \text{ is a unimodular polynomial matrix}\}$$

are all NRD's of $B(s)$. Further each S_L contains at least one element whose highest degree row coefficient matrix is nonsingular.

Another result that we use is the following:

LEMMA 1 [1]–[3]. *If $L(s)$ is an $r \times r$ row proper matrix with the i-th row degree v_i, then $L^{-1}(s)$ is a proper rational matrix. If $v_i \geqq 1$, $i = 1, \cdots, r$, then $L^{-1}(s)$ is strictly proper.*

Now motivated by the approach in [4] to the exact model matching, we have the following theorems characterizing NRD's of a polynomial matrix $B(s)$, where we assume, without loss of generality, that $B(s)$ has no zero rows. Let the ith row of $B(s)$ be

$$b_i(s) = \sum_{j=0}^{\lambda_i} b_j^i s^j, \qquad i = 1, \cdots, f,$$

where b_j^i's are constant row vectors, $\lambda_i \geqq 0$, and $b_{\lambda_i}^i \neq 0$, $i = 1, \cdots, f$. Let

$$\bar{B}_i = \begin{bmatrix} b_{\lambda_i}^i \\ \vdots \\ b_0^i \end{bmatrix}, \qquad B = \left[\begin{array}{c} \bar{B}_1 \\ \hdashline \vdots \\ \hdashline \bar{B}_f \end{array}\right],$$

$$P_i = \begin{bmatrix} 0 & 1 & & 0 & \cdots & 0 \\ \vdots & & \ddots & & \ddots & 1 \\ 0 & \cdots & & & \cdots & 0 \end{bmatrix} \quad \text{is } (\lambda_i+1)\times(\lambda_i+1) \text{ if } \lambda_i \geqq 1 \text{ and}$$

$P_i = 0$ if $\lambda_i = 0$;

$$A = \begin{bmatrix} P_1 & & 0 \\ & \ddots & \\ 0 & & P_f \end{bmatrix}, \qquad \bar{C}_i = [1 \quad 0 \quad \cdots \quad 0] \quad \text{is } 1 \times (\lambda_i + 1),$$

$$C = \begin{bmatrix} \bar{C}_1 & & 0 \\ & \ddots & \\ 0 & & \bar{C}_f \end{bmatrix}.$$

THEOREM 1. *Let $L(s)$ be a row proper NRD of $B(s)$ with the i-th row degree v_i and let (A_1, B_1, C_1, D_1) be a minimal realization of $L^{-1}(s)$, i.e.,*

$$L^{-1}(s) = C_1(sI - A_1)^{-1}B_1 + D_1. \tag{2}$$

Then the following hold:

(1) *There exists a subspace ψ, of dimension less than or equal to*

$$\bar{n} = \sum_{i=1}^{r} v_i$$

satisfying

$$A\psi \subset \psi + \{B\}, \qquad \psi \subset \operatorname{Ker} C. \tag{3}$$

(2) *There exists a matrix X such that $\psi = \{X\}$, and the matrices X, A_1, C_1 satisfy*

$$AX = XA_1 + BC_1. \tag{4}$$

(Thus in case $\dim \psi = \bar{n}$ and X is a basis matrix, there exists a feedback map F such that $(A+BF)\psi \subset \psi$, and $(A+BF)|\psi$ is represented by A_1 [5]–[6].)

Proof. If $L(s)$ is as in hypothesis, then (1) holds for some $f \times r$ polynomial matrix $R(s)$, or

$$B(s)C_1(sI - A_1)^{-1}B_1 = R(s) - B(s)D_1. \tag{5}$$

Then considering the formal power series expansion of $L_1^{-1}(s)$ and equating the coefficients in (5), row by row, we obtain

$$\underbrace{[b_{\lambda_i}^i \ : \ \cdots \ : \ b_0^i]}_{\bar{B}_i} \begin{bmatrix} C_1A_1^{\lambda_i}B_1 & : & C_1A_1^{\lambda_i+1}B_1 & : & \cdots \\ \vdots & & \vdots & & \\ C_1B_1 & : & C_1A_1B_1 & : & \cdots \end{bmatrix} = [0 \ : \ \cdots \ \ 0 \ : \ \cdots]$$

or

$$\bar{B}_i \begin{bmatrix} C_1A_1^{\lambda_i} \\ \vdots \\ C_1 \end{bmatrix} [B_1 \ : \ A_1B_1 \ : \ \cdots] = [0 \ : \ \cdots \ : \ 0 \ : \ \cdots]. \tag{6}$$

But, since (A_1, B_1) is reachable,

$$\bar{B}_i \begin{bmatrix} C_1A_i^{\lambda_i} \\ \vdots \\ C_1 \end{bmatrix} = 0, \qquad i = 1, \cdots, f. \tag{7}$$

But (7) shows that the polynomial matrix $b_i(s)C_1$ is right divisible by $(sI - A_1)$ [15], i.e., there exists a $1 \times \bar{n}$ polynomial matrix $\psi_i(s)$ such that

$$b_i(s)C_1 = \psi_i(s)(sI - A_1) \tag{8}$$

or, letting

$$\psi(s) = \begin{bmatrix} \psi_1(s) \\ \vdots \\ \psi_f(s) \end{bmatrix}, \qquad B(s)C_1 = \psi(s)(sI - A_1). \tag{9}$$

From (8) it follows that degree $(\psi_i(s))<\lambda_i$. Now let

$$\psi_i(s)=\sum_{j=0}^{\lambda_i-1}\psi_j^i s^j,$$

$$\bar{\psi}_i=\begin{bmatrix}0_{1,\bar{n}}\\ \psi^i_{\lambda_i-1}\\ \vdots\\ \psi^i_0\end{bmatrix},\qquad X=\begin{bmatrix}\bar{\psi}_1\\ \hline \vdots\\ \hline \bar{\psi}_f\end{bmatrix}. \tag{10}$$

Then (9) yields (4), i.e., if $\psi=\{X\}$, we have

$$P\psi\subset\psi+\{B\}.$$

$CX=0$ is clear and hence $\psi\subset\operatorname{Ker} C$. □

Remark 1. Note that we have

$$R(s)=\psi(s)B_1+B(s)D_1.$$

Also note that if $v_i\geqq 1$, $i=1,\cdots,r$, D_1 is zero.

Remark 2. In case $B(s)$ is left invertible from (9) it is seen that X in (10) has full column rank in which case dim $\psi=\bar{n}$ and A_1 always represents $A+BF\,|\,\psi$ where F is such that $(A+BF)\psi\subset\psi$.

THEOREM 2. *Let ψ be a subspace satisfying* (3). *Let X be a basis matrix for ψ. Let A_1, C_1 be matrices satisfying* (4). *Also, suppose that C_1 has full low rank. Then the following hold*:

(1) *(A_1, C_1) is observable,*

(2) *there exists a unique matrix B_1 such that (A_1, B_1) is reachable and such that*

$$L(s)=[C_1(sI-A_1)^{-1}B_1]^{-1}$$

is a NRD which is row proper with the ith row degree, v_i being $\geqq 1$, $i=1,2,\cdots,r$.

Proof. With the same notation as before, defining $\psi(s)$ as in (10) we see that (9) holds.

Now since

$$B(s)C_1(sI-A_1)^{-1}=\psi(s),$$

and X has full column rank, (C_1, A_1) is observable. Then since C_1 has full row rank, the observability indices v_i of (C_1, A_1) are $\geqq 1$. Then there exists a nonsingular constant matrix $\tilde{T}$ such that

$$C_1(sI-A_1)^{-1}=L^{-1}(s)W(s)\tilde{T}$$

where $L(s)$ is an $r\times r$ row proper polynomial matrix with row degrees being equal to v_i [1]–[3] and

$$W(s)=\begin{bmatrix}W_1(s) & & 0\\ & \ddots & \\ 0 & & W_r(s)\end{bmatrix},$$

where

$$W_i(s)=[s^{v_i-1}\quad\cdots\quad 1].$$

Thus if we let

$$\tilde{T}B_1 = \left[\begin{array}{c|ccc} v_1\left\{\begin{array}{c} 0\\ \vdots \\ 0 \\ 1 \end{array}\right. & & 0 & \\ \hline & \ddots & & \\ \hline & & & \\ 0 & & & \left.\begin{array}{c} 0\\ \vdots \\ 0 \\ 1 \end{array}\right\} v_r \end{array}\right],$$

we have

$$C_1(sI - A_1)^{-1}B_1 = L^{-1}(s).$$

Since $W(s)\tilde{T}B_1 = I_r$ is coprime with $L(s)$, (A_1, B_1) is reachable [1]–[3]. Then

$$B(s)L^{-1}(s) = \psi(s)B_1 = R(s)$$

and

$$B(s) = R(s)L(s). \quad \square$$

Remark 3. In Theorem 2, if C_1 does not have full row rank, let $\hat{T}$ be any nonsingular matrix such that

$$\hat{T}C_1 = \begin{bmatrix} \bar{C}_1 \\ 0 \end{bmatrix}$$

where $\bar{C}_1$ has full row rank. Then we again have

$$B(s)C_1 = \psi(s)(sI - A_1)$$

with (C_1, A_1) observable, and

$$B(s)\hat{T}^{-1}\begin{bmatrix} \bar{C}_1 \\ 0 \end{bmatrix} = \psi(s)(sI - A_1)$$

with $(\bar{C}_1, A_1)$ observable.

Let $B(s)\hat{T}^{-1} = [B_1(s) \ : \ B_2(s)]$. Then if we choose B_1 as in Theorem 2 for $(\bar{C}_1, A_1)$, the resulting $L(s)$ will satisfy

$$B_1(s)L^{-1}(s) = \psi(s)B_1 = \bar{R}(s).$$

Then

$$B(s)\hat{T}^{-1}\begin{bmatrix} L(s) & 0 \\ 0 & I \end{bmatrix}^{-1} = [\bar{R}(s) \ : \ B_2(s)] = R(s)$$

or

$$B(s) = R(s)\begin{bmatrix} L(s) & 0 \\ 0 & I \end{bmatrix}\hat{T}$$

yields

$$L_1(s) = \begin{bmatrix} L(s) & 0 \\ 0 & I \end{bmatrix} \hat{T}$$

as a row proper NRD of $B(s)$.

Now we have the following corollary which yields a method to find a greatest common right divisor [1]–[3] of two polynomial matrices $V(s)$, $T(s)$, where $T(s)$ is nonsingular, as well as the resulting coprime pair simultaneously. It is clear that this is equivalent to finding a greatest right divisor [1]–[3] of

$$B(s) = \begin{bmatrix} T(s) \\ V(s) \end{bmatrix}.$$

COROLLARY 1. *Let $B(s)$ be an $f \times r$ polynomial matrix with $f \geqq r$, which is left invertible (i.e., no zeros among the diagonal entries of its Smith form).*

Let $\psi_{\max}$ be the maximal dimensional subspace satisfying (3). *Let X, A_1, C_1 be as in Theorem* 2. *If C_1 has full row rank let B_1 be as in Theorem* 2 *and if C_1 does not have full row rank let $\bar{C}_1$ and B_1 be as in Remark* 3. *Then the resulting NRD, $L(s)$, is a row proper greatest right factor of $B(s)$.*

Proof. Suppose that $L(s)$ is not a greatest right divisor. Let $\bar{L}(s)$ be a greatest row proper right divisor. Then by Theorem 1 and Remark 2, there exists a subspace $\bar{\psi}$ satisfying (3) with dim $\bar{\psi}$ = degree (det $\bar{L}(s)$).

But then degree (det $\bar{L}(s)$) > degree (det $L(s)$). However, degree (det $L(s)$) is dimension of $\psi_{\max}$ by Theorem 2 and Remark 3. This is a contradiction. Thus $L(s)$ is a row proper greatest right divisor of $B(s)$. □

Remark 4. There are several methods to find a maximal (A, B)-invariant subspace $\psi_{\max}$ in Ker C [5]–[7].

Once we have found a basis matrix, $X_{\max}$, for $\psi_{\max}$, the corresponding $\psi(s)$ is already available.

Then applying Theorem 2 and Remark 3, we have both a row proper greatest right divisor as well as the resulting polynomial matrix whose only polynomial right divisors are unimodular polynomial matrices. Now the following corollary is immediate.

COROLLARY 2. *An $f \times r$ $(f \geqq r)$ left invertible polynomial matrix $B(s)$ has only unity invariant factors iff*

$$\psi_{\max} = \{0\}.$$

Based on Theorems 1 and 2 we also have the following result.

THEOREM 3. *Let $(\bar{A}, \bar{B}, \bar{C})$ be a system such that $(\bar{C}, \bar{A})$ is observable. Let*

$$\bar{C}(s\bar{I} - \bar{A})^{-1} = D^{-1}(s)\bar{S}(s)$$

be a coprime factorization such that $D(s)$ is row proper with the highest degree row coefficient matrix being the unit matrix (this can be always achieved by a nonsingular constant output transformation) and let $B(s) := \bar{S}(s)\bar{B}$ have no zero rows. Then in regard to the nonsingular right factors of $B(s)$, Theorems 1 *and* 2 *hold with $(\bar{A}, \bar{B}, \bar{C})$ replacing (A, B, C) as defined previously.*

Proof. Let the observability indices of (C, A) be $\bar{V}_i$, $i = 1, \cdots, f$. Since $G(s)$ is strictly proper and $B(s)$ has no zero rows, $\bar{V}_i \geqq 1$, $i = 1, \cdots, f$ [1]–[3]. Since $(\bar{C}, \bar{A})$ is observable, there exist matrices K, $\tilde{T}$ ($\tilde{T}$ being nonsingular) such that

$$\tilde{T}(\bar{A} + K\bar{C})\tilde{T}^{-1} = \tilde{A}, \qquad \bar{C}\tilde{T}^{-1} = \tilde{C} \tag{11}$$

[1]–[3], where

$$\tilde{A}=\begin{bmatrix}\tilde{A}_1 & & 0\\ & \ddots & \\ 0 & & \tilde{A}_f\end{bmatrix},$$

$\tilde{A}_i$ is a $\bar{V}_i\times\bar{V}_i$ matrix given as

$$\tilde{A}_i=\begin{bmatrix}0 & 1 & 0 & \cdots & 0\\ & & \ddots & & \vdots\\ & & & \ddots & 0\\ & 0 & & \ddots & 1\\ 0 & \cdots & \cdots & \cdots & 0\end{bmatrix}$$

if $\bar{V}_i>1$ and $\tilde{A}_i=0$ if $\bar{V}_i=1$.

$$\tilde{C}=\begin{bmatrix}\tilde{C}_1 & & 0\\ & \ddots & \\ 0 & & \tilde{C}_f\end{bmatrix};$$

$\tilde{C}_i$ is the $1\times\bar{V}_i$ matrix given as

$$\tilde{C}_i=[1\quad 0\quad\cdots\quad 0].$$

Such a pair $(\tilde{A},\tilde{C})$ is usually referred to as Brunovsky's canonical form. Then, since $D^{-1}(s)B(s)$ is strictly proper and $D(s)$ is row proper, $\lambda_i<\bar{V}_i$, $i=1,\cdots,f$. Now let

$$\tilde{B}=\tilde{T}\bar{B}=\begin{bmatrix}\hat{B}_1\\ \hline \vdots\\ \hline \hat{B}_f\end{bmatrix}.$$

Now we will show that $\hat{B}_i$ is a $\bar{V}_i\times r$ matrix given as

$$\hat{B}_i=\begin{bmatrix}0_{\bar{V}_i-\lambda_i-1,\,r}\\ \hline \bar{B}_i\end{bmatrix},\qquad i=1,\cdots,f$$

([1]–[3]) where $\bar{B}_i$ is related to $B(s)$ as at the beginning of § 2.

From the definition of $D(s)$ and $\bar{S}(s)$ we obtain

$$D(s)\tilde{C}+\bar{S}(s)\tilde{T}^{-1}\tilde{T}(sI-\tilde{A})\tilde{T}^{-1}=0. \tag{12}$$

Define

$$W(s):=\left[\begin{array}{ccc|ccc}s^{\bar{V}_1-1} & \cdots & 1 & 0 & \cdots & 0\\ \hline & & & \ddots & & \\ & & & & 0 & \\ \hline & 0 & & s^{\bar{V}_f-1} & \cdots & 1\end{array}\right].$$

Then there exists a constant matrix $\bar{K}$ such that

$$D(s)=\operatorname{diag}(s^{\bar{V}_i})+W(s)\bar{K}.$$

Thus we can write (12) as

$$\operatorname{diag}(s^{\bar{V}_i})\tilde{C}+W(s)\bar{K}\tilde{C}+\bar{S}(s)\tilde{T}\tilde{T}^{-1}(sI-\bar{A})\tilde{T}^{-1}=0. \tag{13}$$

On the other hand we have the identity

$$\operatorname{diag}(s^{\bar{V}_i})\tilde{C}+W(s)\tilde{T}(sI-\bar{A}-K\bar{C})\tilde{T}^{-1}=0. \tag{14}$$

Subtracting (14) from (13) we get

$$[\bar{S}(s)\tilde{T}^{-1}-W(s)]\tilde{T}(sI-\bar{A})\tilde{T}^{-1}+W(s)[K\tilde{C}+\tilde{T}K\tilde{C}]=0$$

which implies that

$$W(s)=\bar{S}(s)\tilde{T}^{-1} \quad \text{and} \quad \bar{K}\tilde{C}=-\tilde{T}K\tilde{C}.$$

Thus,

$$B(s)=\bar{S}(s)\bar{B}=W(s)\tilde{B}.$$

Hence, $\bar{B}_i$ are as defined at the beginning of § 2. Now, since the subspaces $\bar{\psi}$ satisfying $\bar{A}\bar{\psi}\subset\bar{\psi}+\{\bar{B}\}$, $\bar{\psi}\subset\operatorname{Ker}\bar{C}$ are independent of the type of transformations occurring in (11) which are invertible, they are the same as the subspaces $\tilde{\psi}$ satisfying

$$\tilde{A}\tilde{\psi}\subset\tilde{\psi}+\{\tilde{B}\}, \qquad \tilde{\psi}\subset\operatorname{Ker}\tilde{C}. \tag{15}$$

But the subspaces $\tilde{\psi}$ satisfying (15) are the same as the subspaces ψ satisfying

$$A\psi\subset\psi+\{B\}, \qquad \psi\subset\operatorname{Ker} C$$

embedded into a larger dimensional vector space. Also the matrices A_1, C_1 satisfying

$$\bar{A}\bar{X}=\bar{X}A_1+\bar{B}C_1$$

where $\bar{\psi}=\{\bar{X}\}$, satisfy

$$\tilde{A}\tilde{T}\bar{X}=\tilde{T}\bar{X}A_1+\tilde{B}C_1,$$

and thus they satisfy

$$AX=XA_1+BC_1$$

for some X such that $\{X\}=\psi$ which is the same as $\bar{\psi}$ (modulo embedding ψ into a larger vector space). Then, by Theorems 1 and 2 the proof follows. □

Remark 5. Theorem 3 shows that any given NRD, $L(s)$, of $B(s)$, in $G(s)=D^{-1}(s)B(s)$ which is a factorization of $\bar{C}(sI-\bar{A})^{-1}\bar{B}$ with $(\bar{C},\bar{A})$ being observable (equivalently the set S_L), there corresponds a unique $(\bar{A},\bar{B})$-invariant subspace in Ker $\bar{C}$. Also, given any such subspace, there corresponds at least one NRD of $B(s)$.

3. Conclusion. We have given a characterization of NRD's of a polynomial matrix in terms of (A, B)-invariant subspaces in Ker C. The results in particular yield a method to obtain simultaneously a row proper greatest right divisor of a left invertible polynomials matrix as well as the resulting polynomial matrix whose greatest right divisors are unimodular polynomial matrices. The results also yield a characterization of the NRD's of the numerator polynomial matrix in a factorization of a transfer matrix in terms of $(\bar{A},\bar{B})$ invariant subspaces in Ker $\bar{C}$ where $(\bar{A},\bar{B},\bar{C})$ is an observable realization of the transfer matrix.

Acknowledgments. The author would like to thank Prof. M. L. J. Hautus for several discussions, and the Dept. of Mathematics of Technische Hogeschool

Eindhoven for their financial support and hospitality while this research was being done. The author is also thankful to Prof. R. E. Kalman for several discussions related to the problems considered in this paper.

REFERENCES

[1] H. H. ROSENBROCK, *State-Space and Multivariable Theory*, Nelson, 1970.

[2] G. D. FORNEY, *Minimal bases of rational vector spaces with applications to multivariable linear systems*, this Journal, 13 (1975), pp. 493–520.

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

[4] E. EMRE, *On the exact matching of linear systems by dynamic compensation*, unpublished research note.

[5] 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.

[6] W. M. WONHAM, *Linear Multivariable Control: A Geometric Approach*, Springer-Verlag, New York, 1974.

[7] L. M. SILVERMAN, *Discrete Riccati equations: alternative algorithms, asymptotic properties and system theory interpretations*, Advances in Control and Dynamic Systems: Theory and Applications, vol. 12, Academic Press, New York, 1976.

[8] B. D. O. ANDERSON AND E. I. JURY, *Generalized Bezoutian and Sylvester matrices in linear multivariable control*, IEEE Trans. Automatic Control, AC-21 (1976), pp. 551–556.

[9] S. KUNG, T. KAILATH AND M. MORF, *A generalized resultant matrix for polynomial matrices*, Proc. IEEE Conf. Decision and Control (Florida), Dec. 1976.

[10] E. EMRE AND L. M. SILVERMAN, *Relatively prime polynomial matrices: algorithms*, Proc. IEEE Conf. Decision and Control (Texas), 1975.

[11] ———, *New criteria and system theoretical interpretations for relatively prime polynomial matrices*, IEEE Trans. Automatic Control, April 1977.

[12] P. A. FUHRMAN, *Simulation of linear systems and factorization of matrix polynomials*, M.I.T. Report ESL-R-762, 1977.

[13] I. GOHBERG, P. LANCASTER AND L. RODMAN, *Spectral analysis of matrix polynomials, I. Canonical forms and divisors*, Dept. of Mathematics and Statistics, Univ. of Calgary, Res. Paper 313, Calgary, Alberta, 1976.

[14] H. LANGER, *Factorization of operator pencils*, Acta Sci. Math., 38 (1976), pp. 83–96.

[15] F. R. GANTMACHER, *The Theory of Matrices*, vol. 1, Chelsea, New York, 1959.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 3, May 1980

0363-0129/80/1803-0005$01.00/0

FIXED AND VARIABLE CONSTRAINTS IN SENSITIVITY ANALYSIS*

J. E. SPINGARN†

Abstract. Sufficient conditions are obtained for the $\underset{\sim}{C}^1$ dependence on a parameter of a local minimizer and associated Lagrange multipliers for parametrized families of nonlinear programming problems in which some constraints do not vary with the parameter. A class of sets, called "cyrtohedra", with properties suitable for the representation of fixed constraint sets is discussed.

Introduction. Let $f, g_i (i \in I)$, and $h_j (j \in J)$ be real-valued twice continuously differentiable functions on $R^n \times P$, where I and J are finite index sets and $P \subset R^s$ is open. Let $C \subset R^n$, and consider the family of nonlinear programming problems

$$(\mathrm{Q}_p) \qquad \begin{aligned} &\min f(x, p) \text{ in } x \quad \text{subject to} \quad g_i(x, p) \leqq 0 \quad \forall i, \\ &h_j(x, p) = 0 \quad \forall j \quad \text{and} \quad x \in C. \end{aligned}$$

The "variable" constraints $g_i(x, p) \leqq 0$ and $h_j(x, p) = 0$ which depend on p, and the "structural", or "fixed" constraints represented by the set C, play fundamentally different roles in our analysis.

In the case where no fixed constraints are present (that is, where $C = R^n$), Fiacco [2] showed that if $\bar{x}$ is a local minimizer for $(\mathrm{Q}_{\bar{p}})$ and vectors $\bar{y} \in R^I_+$ and $\bar{z} \in R^J$ exist with $(\bar{x}, \bar{y}, \bar{z})$ satisfying the "strong second-order conditions" (see below), then $\underset{\sim}{C}^1$ functions $x(p), y(p)$, and $z(p)$ may be defined on a neighborhood of $\bar{p}$ such that $x(\bar{p}) = \bar{x}$, $y(\bar{p}) = \bar{y}$, $z(\bar{p}) = \bar{z}$, and for all p, $x(p)$ is a local minimizer for (Q_p) with unique multiplier vectors $y(p)$ and $z(p)$ and such that $(x(p), y(p), z(p))$ again satisfies the strong second-order conditions for (Q_p). The triple $(\bar{x}, \bar{y}, \bar{z}) \in R^n \times R^I_+ \times R^J$ satisfies the *strong second-order conditions* for the problem

$$(\mathrm{Q}) \qquad \min f(x) \quad \text{subject to} \quad g_i(x) \leqq 0 \quad \forall i, \quad h_j(x) = 0 \quad \forall j, \quad x \in R^n$$

provided that (letting $I_+ = \{i \in I: g_i(\bar{x}) = 0\}$ and $L(x, y, z) = f(x) + \sum_I y_i g_i(x) + \sum_J z_j h_j(x)$)

(A) (i) $\bar{x}$ is feasible for (Q);

(ii) $\nabla_x L(\bar{x}, \bar{y}, \bar{z}) = 0$;

(iii) $\bar{y}_i > 0$ if and only if $g_i(\bar{x}) = 0$;

(iv) $\{\nabla g_i(\bar{x}): i \in I_+\} \cup \{h_j(\bar{x}): j \in J\}$ is a linearly independent set.

(v) If $0 \neq \zeta \in R^n$ satisfies $\zeta \cdot \nabla g_i(\bar{x}) = 0$ $(\forall i \in I_+)$, and $\zeta \cdot \nabla h_j(\bar{x}) = 0$ $(\forall j \in J)$, then $\zeta \cdot H\zeta > 0$, where $H = \nabla^2_x L(\bar{x}, \bar{y}, \bar{z})$.

In case fixed constraints are present, it may well be that the set C itself is expressible as the set of points satisfying a finite number of "variable" type constraints, the dependence on p being trivial. If so, the set C could be eliminated, and the family

* Received by the editors November 7, 1978, and in revised form July 25, 1979.

† Department of Mathematics, Georgia Institute of Technology, Atlanta, Georgia 30332. This research was supported in part by the Air Force Office of Scientific Research, Air Force Systems Command, under Grant 77-3204.

(Q_p) could be reformulated in an equivalent way involving only variable constraints. However, the trouble with this approach (and hence our reason here for segregating the two types of constraints) is that the conditions (A) for the new form of the problem (the hypothesis for Fiacco's result) may be too strong to satisfy, even though the conclusion of Fiacco's theorem is actually valid. This obstacle arises even in the simple case where the fixed constraints are linear. We overcome this difficulty (for a certain class of fixed sets C) by showing that the weaker conditions introduced in § 3 are sufficient to establish the desired extension of Fiacco's theorem. These new conditions are then shown to be the weakest possible ones which give such a result.

In § 2, a class of sets, called "cyrtohedra", is described. These sets possess properties which make them suitable to represent the fixed set C. In [11], [12], this same class of sets will play a central role in our analysis of the "generic" necessity of the strong second-order conditions for optimality in (Q_p), generalizing results obtained in [14].

In § 3, the strong second-order conditions (A) are modified to obtain new conditions. The modified conditions depend on the set C, and reduce to the old conditions (A) when $C = R^n$.

In § 4, the tools developed in § 2 and § 3 are applied to obtain the generalization (4.2) of Fiacco's result, as well as a converse (4.10) which shows that the extended conditions are the weakest which imply the conclusion of (4.2).

1. Preliminaries. Let R^n denote n-dimensional Euclidean space with the inner product $x \cdot y = \Sigma\, x_i y_i$. For any finite set I, R^I denotes Euclidean space of dimension $|I|$. If $J \subset I$ then R^J is regarded, in the natural way, as a subspace of R^I. R^I_+ and R^I_{++} are the nonnegative and (strictly) positive orthants in R^I.

For any set $S \subset R^n$, "rank S" denotes the dimension of the linear subspace "span S" generated by S. "co S" is the convex hull of S, and "relint S" is the interior of S relative to the affine subspace generated by S. Also define $S^\perp = \{\zeta \in R^n : \zeta \cdot \xi = 0, \forall \xi \in S\}$. If h is a function whose domain contains S, "$h|S$" denotes the restriction of h to S.

Let $U \subset R^k$ and $V \subset R^l$ be open sets. A function $h: U \to V$ is *of class* $\underset{\sim}{C}^r$ $(r \geqq 1)$ if all of its partial derivatives of order $\geqq r$ exist and are continuous. If $X \subset R^k$ and $Y \subset R^l$ are arbitrary subsets, then $h: X \to Y$ is *of class* $\underset{\sim}{C}^r$ if for each $x \in X$ there is an open set $U \subset R^k$ containing x and a $\underset{\sim}{C}^r$ function $H: U \to R^l$ such that $h|U \cap X = H|U \cap X$. A $\underset{\sim}{C}^r$ *diffeomorphism* $(r \geqq 1)$ $h: X \to Y$ is a homeomorphism for which both h and h^{-1} are of class $\underset{\sim}{C}^r$ (X and Y are always given the inherited topologies).

$M \subset R^n$ is a k-dimensional $\underset{\sim}{C}^r$ *submanifold* $(r \geqq 1)$ if for each $x \in M$ there is an open set $U \subset R^k$ and a $\underset{\sim}{C}^r$ diffeomorphism ϕ mapping U onto a neighborhood of x in M. The map ϕ is a *local parametrization* for M. Let $J\phi(v)$ be the Jacobian of ϕ at v. The range of $J\phi(v)$ is a k-dimensional subspace of R^n called the *tangent space* to M at $x = \phi(v)$, denoted by "M_x". It does not depend on the choice of ϕ (cf. Milnor [8]).

Let $M \subset R^n$ be a k-dimensional $\underset{\sim}{C}^1$ submanifold, $h: M \to R$ a $\underset{\sim}{C}^1$ function. For each $x \in M$, there is a unique $\nabla h(x) \in M_x$, the *gradient* of h at x, such that for any $\zeta \in M_x$, if $\eta: (-1, 1) \to M$ is a differentiable curve with $\eta(0) = x$ and $\eta'(0) = \zeta$ then $d/dt|_{t=0}\, h(\eta(t)) = \nabla h(x) \cdot \zeta$. If $\nabla h(x) = 0$ then x is a *critical point* for h.

When M and h are of class $\underset{\sim}{C}^2$ and $x = \phi(v)$ is a critical point for h, we define a symmetric bilinear function $\nabla^2 h(x)$ on M_x, the *Hessian* of h at x, by setting

$$\nabla^2 h(x)(u, w) = [\nabla^2(h \circ \phi)(v)](\bar{u}, \bar{w}) \qquad (u, w \in M_x),$$

where $u = J\phi(v)\bar{u}$, $\dot{w} = J\phi(v)\bar{w}$, and where $\nabla^2(h \circ \phi)(v)$ denotes the usual Hessian of $h \circ \phi$ at v. This definition does not depend on the choice of ϕ [6]; x is termed a *nondegenerate critical point* for h if the matrix $\nabla^2(h \circ \phi)(v)$ is invertible.

We will frequently be dealing with the situation where $h = H|M$ for some $H: R^n \to R$. In this case, it is important not to confuse $\nabla h(x)$ with $\nabla H(x)$. (Here $\nabla H(x)$ is the ordinary gradient of H at x). If $\pi: R^n \to M_x$ is orthogonal projection, then

$$\nabla h(x) = \pi(\nabla H(x)). \tag{1.1}$$

If x is a critical point for h, we will say that x is a *critical point for H on M*. Similarly, it is important not to confuse $\nabla^2 H(x)$ with $\nabla^2 h(x)$. If M is affine, then the two bilinear functions agree on M_x.

Let X and Y be topological spaces. If $S(x)$ is a subset of Y for each $x \in X$, we will say that $S: X \to Y$ is a *multifunction*. S has *closed graph* if the set $Gr(S) = \{(x, y): y \in S(x)\}$ is closed in $X \times Y$.

2. Cyrtohedra. The assumptions we will need to make about the fixed set C have been incorporated into the definition of "cyrtohedron". These sets have a local structure similar to polyhedra, except that their "faces" are locally the intersection of zero sets of nonlinear functions.

Throughout this article, we let $U \subset R^n$ be an open set, and $G_\alpha, \alpha \in A$, and $H_\beta, \beta \in B$, finite collections of differentiable functions on U. For any $A_0 \subset A$ and $x \in U$, define

$$\Gamma(x, A_0) = \{\nabla G_\alpha(x): \alpha \in A_0\} \cup \{\nabla H_\beta(x): \beta \in B\},$$

$$Z(A_0) = \{y \in U: 0 = G_\alpha(y) = H_\beta(y)\ \forall \alpha \in A_0, \forall \beta \in B\}.$$

A nonempty connected set $C \subset R^n$ is a *cyrtohedron* of class $\underset{\sim}{C}^k (k \geqq 1)$ if for every $\bar{x} \in C$ there are $\underset{\sim}{C}^k$ functions $G_\alpha, \alpha \in A$, and $H_\beta, \beta \in B$ (for finite index sets A and B) defined on a neighborhood $U \subset R^n$ of $\bar{x}$ such that $\bar{x} \in Z(A)$ and

$$\begin{array}{ll} \text{(2.1a)} & \begin{array}{l}\text{for all } x \in U,\ x \in C \text{ if and only if} \\ G_\alpha(x) \leqq 0\ \forall \alpha \in A \text{ and } H_\beta(x) = 0\ \forall \beta \in B;\end{array} \\ \text{(2.1b)} & \begin{array}{l}\text{if } \sum\limits_A a_\alpha \nabla G_\alpha(\bar{x}) + \sum\limits_B b_\beta \nabla H_\beta(\bar{x}) = 0 \text{ for some} \\ a \in R_+^A \text{ and } b \in R^B, \text{ then } a = 0 \text{ and } b = 0;\end{array} \\ \text{(2.1c)} & \begin{array}{l}\text{for each } A_0 \subset A \text{ there is an integer } s(A_0) \\ \text{such that rank } \Gamma(x, A_0) = s(A_0) \text{ for all } x \in U.\end{array} \end{array}$$

(The sets A and B will both be empty precisely when x belongs to the interior of C relative to R^n). Condition (b) is the *Mangasarian–Fromovitz constraint qualification*. It is known [3, Corollary 2.10] to imply

$$\begin{array}{ll} \text{(2.1b')} & \begin{array}{l}\text{if } \sum\limits_A a_\alpha \nabla G_\alpha(x) + \sum\limits_B b_\beta \nabla H_\beta(x) = 0 \text{ for some} \\ x \in U,\ a \in R_+^A, \text{ and } b \in R^B, \text{ then } a = 0 \text{ and } b = 0\end{array} \end{array}$$

if U is sufficiently small. Condition (2.1c) implies, by an argument using the implicit function theorem (cf. Auslander and MacKenzie [1, p. 32]), that U can be chosen so that also

$$\begin{array}{ll} \text{(2.1d)} & \begin{array}{l}\text{for all } A_0 \subset A,\ Z(A_0) \text{ is a connected } (n - s(A_0))\text{-} \\ \text{dimensional submanifold.}\end{array} \end{array}$$

Similarly, it follows from (2.1c) that if U is taken small enough, one also has

$$\text{(2.1c')} \qquad \text{if } A_0 \subset A_1 \subset A \text{ and } s(A_0) = s(A_1) \text{ then } Z(A_0) = Z(A_1).$$

We will say that $(G_\alpha(\alpha \in A), H_\beta(\beta \in B), U)$, or more briefly (G_α, H_β, U), is a *local representation* (l.r.) for the cyrtohedron C if $Z(A) \neq \varnothing$ and (2.1a), (2.1b′), (2.1c′) and (2.1d) hold. It is clear from the above that we have

PROPOSITION 2.2. *If $C \subset R^n$ is a cyrtohedron and $x \in C$, then there is a l.r. (G_α, H_β, U) for C such that $x \in Z(A)$.*

Examples of cyrtohedra 2.3. (a) A *differentiable submanifold* in R^n is a cyrtohedron for which the set A in (2.1) may always be taken to be empty.

(b) Cyrtohedra for which the set A in (2.1) may always be taken either empty or of cardinality one are *submanifolds with boundary* (cf. Milnor [8, p. 12]).

(c) A *polyhedral convex set* is the intersection of a finite number of closed half-spaces in R^n (cf. Grünbaum [4]).

(d) Sets that can be expressed as $C = \{x \in R^n\colon g_i(x) \leqq 0,\ i = 1, \cdots, m, \text{ and } h_j(x) = 0,\ j = 1, \cdots, p\}$, where the functions g_i and h_j are of class $\underline{C}^k$ and have the property that for every $x \in C$, $\{\nabla g_i(x)\colon i \in I_+(x)\} \cup \{\nabla h_j(x)\colon j = 1, \cdots, p\}$ is linearly independent, where $I_+(x) = \{i\colon g_i(x) = 0\}$.

Let (G_α, H_β, U) be any l.r. for C. It can easily be demonstrated that for any $A_0 \subset A$,

(2.4) the multifunctions $x \to \operatorname{span} \Gamma(x, A_0)$ and
$$x \to \{\textstyle\sum_{A_0} a_\alpha \nabla G_\alpha(x) + \sum_B b_\beta \nabla H_\beta(x)\colon a \in R_+^{A_0} \text{ and } b \in R^B\},$$
defined for $x \in U$, have closed graph;

(2.5) the multifunctions $x \to \Gamma(x, A_0)^\perp$ and
$$x \to \{\zeta \in R^n\colon \zeta \cdot \nabla G_\alpha(x) \leqq 0 \text{ and } \zeta \cdot \nabla H_\beta(x) = 0\ \forall \alpha \in A_0, \beta \in B\},$$
defined for $x \in Z(A_0)$, have closed graph.

Let $C \subset R^n$ be a cyrtohedron, $x \in C$, (G_α, H_β, U) a l.r. such that $x \in U$. The *tangent cone* to C at x is

$$T_C(x) = \{\zeta \in R^n\colon \zeta \cdot \nabla G_\alpha(x) \leqq 0\ \forall \alpha \in A_+(x),\ \zeta \cdot \nabla H_\beta(x) = 0\ \forall \beta \in B\},$$

where $A_+(x) = \{\alpha \in A\colon G_\alpha(x) = 0\}$. The symbol "$L_C(x)$" will denote the largest linear subspace contained in $T_C(x)$. Thus, $L_C(x) = \Gamma(x, A_+(x))^\perp$. Equivalently, $L_C(x)$ is the tangent space to the $(n - s(A_+(x)))$-dimensional submanifold $Z(A_+(x))$ at x. The definitions of $T_C(x)$ and $L_C(x)$ do not depend on the local representation (G_α, H_β, U). To see this, let $\underline{A}(x, \zeta)$ denote the set of all $\underline{C}^1$ arcs $\phi\colon (-1, 1) \to U$ such that $\phi(0) = x$ and $\phi'(0) = \zeta$, and note that

$$\begin{aligned} T_C(x) &= \{\zeta\colon \exists \phi \in \underline{A}(x, \zeta) \text{ with } \phi(t) \in C\ \forall t \geqq 0\}, \\ L_C(x) &= \{\zeta\colon \exists \phi \in \underline{A}(x, \zeta) \text{ with } \phi(t) \in C\ \forall t\}. \end{aligned} \tag{2.6}$$

The *normal cone* to C at x is the set

$$N_C(x) = \left\{ \sum_{\alpha \in A_+(x)} a_\alpha \nabla G_\alpha(x) + \sum_{\beta \in B} b_\beta \nabla H_\beta(x)\colon a \in R_+^{A_+(x)} \text{ and } b \in R^B \right\}$$

which is the polar of the tangent cone. Note that

$$\operatorname{span} N_C(x) = L_C(x)^\perp = \operatorname{span} \Gamma(\mathrm{x}, \mathrm{A}_+(\mathrm{x})). \tag{2.7}$$

Every polyhedral set P has the property that each $x \in P$ belongs to the relative interior of exactly one face of P, where a "face" is defined to be the intersection of P with a supporting hyperplane to P [4]. We now define an analogous concept for cyrtohedra.

For any $x, y \in C$, define the equivalence relation $\sim$ by specifying $x \sim y$ if and only if there exists a sequence $x = x_0, x_1, \cdots, x_p = y$ in C such that for each pair (x_i, x_{i+1}) $(i = 0, \cdots, p-1)$, there exists a l.r. (G_α, H_β, U) such that $Z(A) \supset \{x_i, x_{i+1}\}$. The equivalence classes under this relation are the *faces* of C.

A few examples help to clarify the definition:

(a) The faces of a polyhedral convex set are the relative interiors of the "faces" in the sense of [4].

(b) A submanifold $C \subset R^n$ has only one face.

(c) If C is the hemisphere $C = \{x = (x_1, \cdots, x_n) \in R^n : |x| \leqq 1 \text{ and } x_n \geqq 0\}$, then C has four faces, corresponding to the choices of equality or strict inequality in the definition of C.

For any l.r. (G_α, H_β, U) for C and any $x \in Z(A)$, $L_C(x)$ is the tangent space at x to the $(n - s(A))$-dimensional submanifold $Z(A)$, and hence $\dim L_C(x) = n - s(A)$. For each pair (x_i, x_{i+1}) in the definition of "face" it follows that $\dim L_C(x_i) = \dim L_C(x_{i+1})$, so for any face F of C and any $x \in F$ we may define $\dim F = \dim L_C(x)$.

It will now be shown that the faces of C are submanifolds. Let F be a face of C, $\bar{x} \in F$, and let (G_α, H_β, U) be any l.r. for C such that $\bar{x} \in Z(A)$ (cf. Proposition 2.2). It is enough to show that $F \cap U = Z(A)$, since $Z(A)$ is a submanifold of dimension $n - s(A)$ (cf. (2.1d)). Clearly $Z(A) \subset F$ by the definition of "face". If $x \in U \backslash Z(A)$ then $s(A_+(x)) < s(A)$, since otherwise (2.1c′) would imply $x \in Z(A)$. Thus $\dim L_C(x) = n - s(A_+(x)) > n - s(A) = \dim L_C(\bar{x})$. But $\dim L_C(\cdot)$ is constant on F, so $x \notin F$. Then $F \cap U = Z(A)$, as desired. Note also that each face of C is a connected set. This follows from (2.1d) and the definition of "face". We summarize by the following theorem.

THEOREM 2.8. *Let $C \subset R^n$ be a cyrtohedron of class $\underline{C}^r$ $(r \geqq 1)$, $x \in C$. Then x lies on a unique face F of C, and F is a connected $\underline{C}^r$ submanifold of R^n. The tangent space F_x to F at x is $L_C(x)$. There exists a local representation (G_α, H_β, U) for C such that $x \in Z(A)$, and for any such representation, one has $Z(A) = F \cap U$, $\dim F = \dim L_C(x) = n - s(A)$, $\dim N_C(x) = s(A)$, and $\dim T_C(x) = n - |B|$.*

From (2.4) and (2.5), the next theorem follows immediately.

THEOREM 2.9. *Let $C \subset R^n$ be a cyrtohedron. Then $x \to L_C(x)^\perp$ and $x \to N_C(x)$ define multifunctions: $C \to R^n$ having closed graphs. If F is any face of C then $x \mapsto L_C(x)$ and $x \mapsto T_C(x)$ define multifunctions: $F \to R^n$ with closed graphs.*

The following will be needed in the proof of Theorem 4.2.

PROPOSITION 2.10. *Let $C \subset R^n$ be a cyrtohedron, $\bar{x} \in C$, and let F be the face of C containing $\bar{x}$. If $\bar{\zeta} \in \operatorname{relint} N_C(\bar{x})$, then there are neighborhoods $V \subset C$ of $\bar{x}$ and $W \subset R^n$ of $\bar{\zeta}$ such that: for any $x \in V$ and $\zeta \in W$, $\zeta \in N_C(x)$ implies $x \in F$.*

The proof of Proposition 2.10 will require

LEMMA 2.11. *Let $C \subset R^n$ be a cyrtohedron, (G_α, H_β, U) a l.r., $A' \subsetneqq A$, and $\bar{x} \in Z(A)$. Then either*

(a) *there exists $\xi \in R^n$ such that $\xi \cdot \nabla G_\alpha(\bar{x}) = \xi \cdot \nabla H_\beta(\bar{x}) = 0$, $\forall \alpha \in A'$, $\beta \in B$, and $\xi \cdot \nabla G_\alpha(\bar{x}) < 0$ $\forall \alpha \in A \backslash A'$, or*

(b) *there is a neighborhood $V \subset U$ of $\bar{x}$ such that no $x \in V$ satisfies*

$$\begin{aligned} &0 = G_\alpha(x) = H_\beta(x) && \forall \alpha \in A' \quad \forall \beta \in B, \\ &G_\alpha(x) < 0 && \forall \alpha \in A \backslash A'. \end{aligned} \tag{2.12}$$

Proof of Proposition 2.10. Assume Lemma 2.11 for the moment and suppose Proposition 2.10 to be false. Then for some $\bar{x} \in F$ and $\bar{\zeta} \in \operatorname{relint} N_C(\bar{x})$, there are sequences (x_k) and (ζ_k) such that

$$x_k \to \bar{x}, \qquad \zeta_k \to \bar{\zeta}, \qquad x_k \in C \backslash F, \quad \zeta_k \in N_C(x_k). \tag{2.13}$$

Let (G_α, H_β, U) be a l.r. for C such that $\bar{x} \in Z(A)$ (cf. Proposition 2.2). Assume (as we may) that $x_k \in U$ for all k, and define $A_+(x_k) = \{\alpha \in A: G_\alpha(x_k) = 0\}$. By Theorem 2.8, $Z(A) = F \cap U$, so $x_k \notin F$ implies $A_+(x_k) \subsetneqq A$. Passing to a subsequence, it may be assumed for some $A' \subsetneqq A$ that $A_+(x_k) = A'$ for all k. For this A', one of the alternatives of Lemma 2.11 must hold. However, $x = x_k$ satisfies (2.12) for every k, so alternative (b) is false.

Let ξ satisfy (a). Then $\xi \in T_C(\bar{x})$. Pick $\alpha_0 \in A \backslash A'$ and for $t \in R$, define $\zeta(t) = \bar{\zeta} + t\nabla G_{\alpha_0}(\bar{x})$. By choice of $\bar{\zeta}$, $\zeta(t) \in N_C(\bar{x})$ for all t in a neighborhood of $t = 0$. For each k, $\zeta_k \in \operatorname{span} \Gamma(x_k, A')$ by (2.7) and (2.13). By (2.4) and (2.13), this implies $\bar{\zeta} \in \operatorname{span} \Gamma(\bar{x}, A')$. So for $t < 0$, $\xi \cdot \zeta(t) = t(\xi \cdot \nabla G_{\alpha_0}(\bar{x})) > 0$, and hence $\zeta(t) \notin T_C(\bar{x})^0 = N_C(\bar{x})$, a contradiction. □

Proof of Lemma 2.11. If $A' = \varnothing$ then (a) holds by (2.1b'). Suppose, then, that $A' \neq \varnothing$ and neither (a) nor (b) holds. Let $M = Z(A')$, and let $\pi: R^n \to M_{\bar{x}}$ be projection onto the tangent space to M at $\bar{x}$. By (2.1c, d), $M_{\bar{x}}^{\perp} = \operatorname{span} \Gamma(\bar{x}, A')$. Since (a) is false, there exists no $\xi \in M_{\bar{x}}$ such that $\xi \cdot \pi(\nabla G_\alpha(\bar{x})) < 0 \ \forall \alpha \in A \backslash A'$, or equivalently, $0 \in \operatorname{co} \{\pi(\nabla G_\alpha(\bar{x})): \alpha \in A \backslash A'\}$. Hence there is a subset $A'' \subset A \backslash A'$ such that

$$(2.14) \qquad 0 \in \operatorname{relint} \operatorname{co} \{\pi(\nabla G_\alpha(\bar{x})): \alpha \in A''\}.$$

Let $L = Z(A' \cup A'')$. Then $\bar{x} \in L \subset M$ and by (2.1c, d) (letting $m = \dim M$ and $l = \dim L$),

$$(2.15) \qquad \begin{aligned} m - l &= \operatorname{rank} \Gamma(\bar{x}, A' \cup A'') - \operatorname{rank} \Gamma(\bar{x}, A') \\ &= \operatorname{rank} \{\pi(\nabla G_\alpha(\bar{x})): \alpha \in A''\}. \end{aligned}$$

Let V_l and V_{m-l} denote the subspaces of R^m consisting of all m-tuples whose last $m - l$ and whose first l coordinates, respectively, are zero. Let $\phi: W \to M$ be a local parametrization for M, $0 \in W \subset R^m$, $\phi(0) = \bar{x}$, and $\phi(V^l \cap W) = L \cap \phi(W)$. For each $\alpha \in A''$, define $g_\alpha = G_\alpha \circ \phi$. Then

$$(2.16) \qquad V^l \cap W = \{q \in W: g_\alpha(q) = 0 \ \forall \alpha \in A''\}.$$

By (2.14), $0 \in \operatorname{relint} \operatorname{co} (\{\nabla g_\alpha(0): \alpha \in A''\})$, and by (2.15), $V_{m-l} = \operatorname{span} (\{\nabla g_\alpha(0): \alpha \in A''\})$. Hence,

$$(2.17) \qquad \text{if } 0 \neq \eta \in V_{m-l}, \text{ then } \eta \cdot \nabla g_\alpha(0) > 0 \text{ for some } \alpha \in A''.$$

Since (b) is false, there is a sequence (x_k) in U converging to $\bar{x}$ such that each x_k satisfies (2.12). For all k, $x_k \in M$, and $x_k \in \phi(W)$ for k sufficiently large, so we may define $q_k = \phi^{-1}(x_k)$. Then $q_k \to 0$ and, since $A'' \subset A \backslash A'$, $g_\alpha(q_k) < 0$ for all $\alpha \in A''$ and all k. Each q_k may be written as $q_k = u_k + v_k$, $u_k \in V_l$, $v_k \in V_{m-l}$, and by (2.16), $v_k \neq 0$. Thus, passing to a subsequence, it can be assumed that

$$\frac{v_k}{|v_k|} \to \eta \in V_{m-l}.$$

By (2.17), there exists $\alpha_0 \in A''$ such that $\eta \cdot \nabla g_{\alpha_0}(0) > 0$. For each k, the mean-value theorem for functions of $m - l$ variables may be invoked to write

$$g_{\alpha_0}(u_k, v_k) = g_{\alpha_0}(u_k, v_k) - g_{\alpha_0}(u_k, 0) = \nabla_v g_{\alpha_0}(u_k, \theta_k v_k) \cdot v_k$$

for some $0 < \theta_k < 1$. Then

$$0 > \frac{g_{\alpha_0}(u_k, v_k)}{|v_k|} = \frac{v_k}{|v_k|} \cdot \nabla_v g_{\alpha_0}(u_k, \theta_k v_k).$$

Taking the limit as $k \to \infty$,

$$0 \geqq \eta \cdot \nabla_v g_{\alpha_0}(0) = \eta \cdot \nabla g_{\alpha_0}(0)$$

(since $\eta \in V_{m-1}$), a contradiction. □

PROPOSITION 2.18. *Let $C \subset R^n$ be a cyrtohedron, F a face,*

$$L = \{(x, \zeta) \in R^{2n} \colon x \in F, \zeta \in L_C(x)^\perp\}, \quad \textit{and}$$
$$M = \{(x, \zeta) \in L \colon \zeta \in \operatorname{relint} N_C(x)\}.$$

Then M is open relative to L.

Proof. Let (x_i, ζ_i) be a sequence in $L \backslash M$ converging to some $(\bar{x}, \bar{\zeta}) \in L$. $\zeta_i \notin \operatorname{relint} N_C(x_i)$ implies, by definition of "normal cone" and a separation argument [10, 11.3] that there exists $\eta_i \in R^n$ such that $\eta = \eta_i$, $x = x_i$, and $\zeta = \zeta_i$ satisfy

$$(2.19) \qquad |\eta| = 1, \qquad \eta \in T_C(x) \cap L_C(x)^\perp, \qquad \eta \cdot \zeta \geqq 0.$$

Passing to a subsequence, it may be assumed that $\eta_i \to \bar{\eta}$. By Theorem 2.9, $\eta = \bar{\eta}$, $x = \bar{x}$, and $\zeta = \bar{\zeta}$ satisfy (2.19), implying that $\bar{\zeta} \notin \operatorname{relint} N_C(\bar{x})$, and hence $(\bar{x}, \bar{\zeta}) \notin M$. □

3. The strong second-order conditions. *Suppose henceforth that f, g_i, and h_j are of class $\underline{C}^2$ on R^n, and that $C \subset R^n$ is a cyrtohedron of class $\underline{C}^2$. Let $r = n + |I| + |J|$* and let $\tilde{C} = C \times R^I_+ \times R^J \subset R^r$. For $w = (x, y, z) \in R^r$, define $L \colon R^r \to R$ and $\tau \colon R^r \to R^r$ by

$$L(w) = f(x) + \sum_I y_i g_i(x) + \sum_J z_j h_j(x),$$
$$\tau(w) = (\nabla_x L(w), -\nabla_y L(w), -\nabla_z L(w)).$$

In this section, first- and second-order conditions will be studied. Theorem 3.3 gives first-order conditions of Kuhn–Tucker type which are necessary under a constraint qualification called the "independence criterion". In Lemma 3.1, these conditions are shown to correspond to solutions w to $0 \in \tau(w) + N_{\tilde{C}}(w)$ as observed by Robinson [9] when $C = R^n$. New conditions (SSOC) are introduced which generalize (A), and in Theorem 3.5 the new conditions are further characterized. Theorem 3.5 will be useful in our study of sensitivity analysis, as well as in [11], [12], where we will show that for certain classes of problems, the conditions (SSOC) are "generically" necessary for optimality.

LEMMA. *$\tilde{C}$ is a cyrtohedron in R^r of class $\underline{C}^2$. For any $w = (x, y, z) \in \tilde{C}$,*

$$N_{\tilde{C}}(w) = N_C(x) \times \{t \in -R^I_+ \colon t_i y_i = 0, \forall i \in I\} \times \{0\}.$$

Proof. If $K \subset R^s$ and $K' \subset R^{s'}$ are cyrtohedra, it is easy to show that $K \times K'$ is a cyrtohedron in $R^s \times R^{s'}$ and that for any $(x, x') \in K \times K'$, $N_{K \times K'}(x, x') = N_K(x) \times N_{K'}(x')$. □

LEMMA 3.1. *For any $\bar{w} = (\bar{x}, \bar{y}, \bar{z}) \in \tilde{C}$,*

(a) $-\tau(\bar{w}) \in N_{\tilde{C}}(\bar{w})$ *if and only if $\bar{x}$ is feasible for* (Q) *and*

$$-\nabla_x L(\bar{w}) \in N_C(\bar{x})$$
$$\forall i \in I \quad \textit{if } \bar{y}_i > 0 \textit{ then } g_i(\bar{x}) = 0;$$

(b) $-\tau(\bar{w}) \in \operatorname{relint} N_{\tilde{C}}(\bar{w})$ *if and only if $\bar{x}$ is feasible for* (Q) *and*

$$-\nabla_x L(\bar{w}) \in \operatorname{relint},.\, N_C(\bar{x})$$
$$\forall i \in I, \qquad \bar{y}_i > 0 \quad \textit{if and only if} \quad g_i(\bar{x}) = 0.$$

(c) $\tau(\bar{w}) \in L_{\tilde{C}}(\bar{w})^\perp$ *if and only if* $\nabla_w L(\bar{w}) \in L_{\tilde{C}}(\bar{w})^\perp$.

Proof. By the previous lemma, $-\tau(\bar w)\in N_{\tilde C}(\bar w)$ *if and only if*

$$-\nabla_x L(\bar w)\in N_C(\bar x),$$
$$\nabla_y L(\bar w)\in\{t\in -R^I_+ : t_i\bar y_i=0,\ \forall i\in I\},$$
$$\nabla_z L(\bar w)=0,$$

so (a) follows from

$$\begin{aligned}\nabla_y L(\bar w)&=(\cdots, g_i(\bar x),\cdots)\\ \nabla_z L(\bar w)&=(\cdots, h_j(\bar x),\cdots).\end{aligned}\tag{3.2}$$

The proof of (b) is similar. We have

$$\nabla_w L(\bar w)=(\nabla_x L(\bar w), \nabla_y L(\bar w), \nabla_z L(\bar w)),$$
$$\tau(\bar w)=(\nabla_x L(\bar w), -\nabla_y L(\bar w), -\nabla_z L(\bar w)),$$
$$L_{\tilde C}(\bar w)^\perp = L_C(\bar x)^\perp\times L_{R^I_+}(\bar y)^\perp\times L_{R^J}(\bar z)^\perp$$

from which (c) follows. □

The *independence criterion* is satisfied for (Q) at the feasible point $\bar x$ if, for any $a\in R^{I_+}$ and $b\in R^J$,

$$\text{(IC)}\qquad \sum_{I_+} a_i\nabla g_i(\bar x)+\sum_J b_j\nabla h_j(\bar x)\in L_C(\bar x)^\perp \quad\text{implies}\quad a=0 \text{ and } b=0.$$

It is trivially satisfied if $I_+=J=\varnothing$. If F is the face of C containing $\bar x$, then (IC) holds at $\bar x$ if and only if $\{\nabla(g_i|F)(\bar x): i\in I_+\}\cup\{\nabla(h_j|F)(\bar x): j\in J\}$ is a linearly independent subset of $F_{\bar x}=L_C(\bar x)$.

From (2.1b′) and by the definition of $L_C(\bar x)$, it follows easily that if (IC) is satisfied for (Q) at $\bar x$, then for any l.r. (G_α, H_β, U) with $\bar x\in U$, the Mangasarian–Fromovitz constraint qualification [5, 4.10.4] is satisfied at $\bar x$ for the problem

$$\text{(Q}'\text{)}\qquad\begin{aligned}&\min f(x)\quad\text{subject to}\quad x\in U,\quad g_i(x)\leqq 0\quad\forall i,\\ &h_j(x)=0\quad\forall j,\quad G_\alpha(x)\leqq 0\quad\forall\alpha,\qquad H_\beta(x)=0\quad\forall\beta.\end{aligned}$$

If $\bar x$ happens to be a local minimizer for (Q), and hence also for (Q′), then by the Lagrange multiplier rule [5, 4.2.1] there exist multiplier vectors such that the standard first-order conditions hold at $\bar x$ for the problem (Q′). By the definition of $N_C(\bar x)$, this implies

THEOREM 3.3. *If $\bar x$ is a local minimizer for* (Q) *and if* (IC) *for* (Q) *is satisfied at $\bar x$, then there exists $(\bar y, \bar z)\in R^I_+\times R^J$ such that*

(i) $-\nabla_x L(\bar x, \bar y, \bar z)\in N_C(\bar x)$;

(ii) $\forall i\in I$, *if* $\bar y_i>0$ *then* $g_i(\bar x)=0$.

Let $\bar w=(\bar x, \bar y, \bar z)\in\tilde C$. Then $\bar w$ will be said to satisfy the *strong second-order conditions* for (Q) provided that

(SSOC)
(i) $\bar x$ is feasible for (Q);
(ii) $-\nabla_x L(\bar w)\in\operatorname{relint} N_C(\bar x)$;
(iii) $\forall i\in I$, $\bar y_i>0$ if and only if $g_i(\bar x)=0$;
(iv) condition (IC) holds for (Q) at $\bar x$;
(v) $(\nabla^2_x(L|F)(\bar w))(\zeta,\zeta)>0$ for all $\zeta\in R^n$ satisfying
$0\neq\zeta\in L_C(\bar x)$, $\zeta\cdot\nabla g_i(\bar x)=\zeta\cdot\nabla h_j(\bar x)=0\ \forall i\in I_+, j\in J.$

In (v), F is the face of C containing $\bar{x}$. The Hessian in (v) is well-defined since $\bar{w}$ is a critical point for L on F by (i), (ii), (iii), and (3.1c).

Let us remark on the relationship between (A) and (SSOC). Fix (G_α, H_β, U), a l.r. with $\bar{x} \in Z(A)$, and let (Q′) be as before. If $(\bar{x}, \bar{y}, \bar{a}, \bar{z}, \bar{b}) \in R^n \times R^I_+ \times R^A_+ \times R^J \times R^B$ satisfies (A) for (Q′), then it is easily checked that $(\bar{x}, \bar{y}, \bar{z})$ satisfies (SSOC) for (Q). If $(\bar{x}, \bar{y}, \bar{z}) \in R^n \times R^I_+ \times R^J$ satisfies (SSOC) for (Q), then it is possible to find $\bar{a} \in R^A_+$ and $\bar{b} \in R^B$ such that $(\bar{x}, \bar{y}, \bar{a}, \bar{z}, \bar{b})$ satisfies (Ai, ii, iii) and for any such $\bar{a}$ and $\bar{b}$, (Av) will automatically hold for (Q′). However, (Aiv) may fail, except in the special case where C is of the form (2.3d). If, for example, C is a four-sided pyramid in R^3 with apex $\bar{x}$, then (Aiv) can never be satisfied for (Q′) because no set of four vectors in R^3 can be linearly independent. In this case $L_C(\bar{x}) = \{0\}$, so $(\bar{x}, \bar{y}, \bar{z})$ satisfies (SSOC) for (Q) if and only if $\bar{x}$ is feasible for (Q), $I_+ = J = \emptyset$, and $-\nabla f(\bar{x}) \in \operatorname{relint} N_C(\bar{x})$. Of course, if $C = R^n$ (i.e., there are no fixed constraints), then (SSOC) is equivalent to (A).

Remark 3.4. Conditions (SSOC) implies $\bar{x}$ is a local minimizer for (Q). To see this, observe that (SSOC) implies the second-order sufficient conaitions [7, Thm. 6] for (Q′) at $\bar{x}$. Thus, $\bar{x}$ is an isolated local minimizer for (Q′), and hence also for (Q).

For any face F of C (Q_F) will denote the restriction of (Q) to F, namely

$$(\mathrm{Q}_F) \qquad \min f(x) \quad \text{subject to} \quad x \in F, \quad g_i(x) \leqq 0 \quad \forall i, \quad h_j(x) = 0 \quad \forall j.$$

If $\phi: U \to F$ is a local parametrization for F, (Q_ϕ) will denote the problem

$$(\mathrm{Q}_\phi) \qquad \begin{aligned} &\min f(\phi(q)) \quad \text{subject to} \quad q \in U, \quad g_i(\phi(q)) \leqq 0 \quad \forall i, \\ &h_j(\phi(q)) = 0 \quad \forall j. \end{aligned}$$

THEOREM 3.5. *Let $\bar{w} = (\bar{x}, \bar{y}, \bar{z}) \in \tilde{C}$, let F be the face of C containing $\bar{x}$, $\tilde{G}$ the face of $\tilde{C}$ containing $\bar{w}$. Then the following three conditions are equivalent:*

(3.6) $\bar{w}$ *satisfies* (SSOC) *for* (Q);

(3.7a) $-\tau(\bar{w}) \in \operatorname{relint} N_{\tilde{C}}(\bar{w})$,

(3.7b) *$\bar{w}$ is a nondegenerate critical point for L on $\tilde{G}$,*

(3.7c) *$\bar{x}$ is a local minimizer for* (Q_F);

(3.8a) $-\tau(\bar{w}) \in \operatorname{relint} N_{\tilde{C}}(\bar{w})$,

(3.8b) *if ϕ is a local parametrization for F, and $\bar{x} = \phi(\bar{q})$, then $\bar{v} = (\bar{q}, \bar{y}, \bar{z})$ satisfies* (A) *for* (Q_ϕ).

Proof. By Lemma 3.1, $-\tau(\bar{w}) \in \operatorname{relint} N_{\tilde{C}}(\bar{w})$ if and only if (SSOCi, ii, iii) hold. So assume these equivalent conditions hold. It must then be shown that $\bar{w}$ satisfies (SSOCiv, v)$\Leftrightarrow$(3.7b, c)$\Leftrightarrow$(3.8b).

Note that (SSOCiii) implies $\tilde{G} = F \times R^{I_+}_{++} \times R^J$. Let $\phi: U \to F$ be a local parametrization for F, where $U \subset R^e (e = \dim F)$ is open and $\bar{x} = \phi(\bar{q})$ for some $\bar{q} \in U$. The dimension of $\tilde{G}$ is $c = e + |I_+| + |J|$. Define $\tilde{U} = U \times R^{I_+}_{++} \times R^J \subset R^c$, and for any $v = (q, y, z) \in \tilde{U}$, let $\tilde{\phi}(v) = (\phi(q), y, z)$. Then $\tilde{\phi}: \tilde{U} \to \tilde{G}$ is a local parametrization for $\tilde{G}$. For $\bar{v} = (\bar{q}, \bar{y}, \bar{z})$, we have $\bar{w} = \tilde{\phi}(\bar{v})$.

If $\bar{w}$ satisfies (SSOCi, ii, iii), it easily follows that $\bar{v}$ satisfies (Ai, ii, iii) for (Q_ϕ).

The theorem is now a consequence of the following remarks, which are valid under the assumption that (SSOCi, ii, iii) hold.

Claim. Condition (3.7b) holds if and only if the Hessian at $\bar{v}$ of the map $L \circ \tilde{\phi}\colon \tilde{U} \to R$, namely the $(c \times c)$-matrix

$$\nabla_v^2 L(\tilde{\phi}(\bar{v})) = \begin{bmatrix} H & D \\ D^t & 0 \end{bmatrix} \tag{3.9}$$

is invertible, where $H = \nabla_q^2 L(\tilde{\phi}(\bar{v}))$ and D is the matrix whose columns are the vectors $\nabla_q g_i(\phi(\bar{q}))$, $i \in I_+$, and $\nabla_q h_j(\phi(\bar{q}))$, $j \in J$.

Claim. $\bar{x}$ is a local minimizer for (Q_F) if and only if $\bar{q}$ is a local minimizer for (Q_ϕ).

Claim. By known facts about the conditions (A) (cf. McCormick [7, Thm. 7]), $\bar{v}$ satisfies (A) for (Q_ϕ) if and only if $\bar{q}$ is a local minimizer for (Q_ϕ) and the matrix (3.9) is invertible.

Claim 3.10. $\bar{v}$ satisfies (A) for (Q_ϕ) if and only if

$$\text{the set } \{\nabla_q g_i(\phi(\bar{q}))\colon i \in I_+\} \cup \{\nabla_q h_j(\phi(\bar{q}))\colon j \in J\} \text{ is linearly independent, and} \tag{3.11}$$

$$\text{if } 0 \neq \xi \in R^e,\ \xi \cdot \nabla_q g_i(\phi(\bar{q})) = 0,\ \forall i \in I_+ \text{ and } \xi \cdot \nabla_q h_i(\phi(\bar{q})) = 0,\ \forall j \in J, \text{ then } \xi \cdot H\xi > 0. \tag{3.12}$$

Claim. Condition (SSOCiv) is equivalent to (3.11), and (SSOCv) is equivalent to (3.12).

4. Sensitivity analysis. Consider the family

$$(\mathrm{Q}_p) \qquad \min f(x, p) \text{ in } x \quad \text{subject to} \quad g_i(x, p) \leqq 0 \quad \forall i, \qquad h_j(x, p) = 0 \quad \forall j, \quad x \in C$$

of problems indexed by the parameter $p \in P$, *where* P *is open, the functions* f, g_i, *and* h_j *are of class* $\underset{\sim}{C}^2$ *on* $R^n \times P$, I *and* J *are finite index sets, and* $C \subset R^n$ *is a cyrtohedron of class* $\underset{\sim}{C}^2$. Let $r = n + |I| + |J|$ and let $\tilde{C} = C \times R_+^I \times R^J \subset R^r$. Some evident modifications in notation are forced by the parameter p; for example, we now have for $(w, p) \in R^r \times P$ $(w = (x, y, z))$,

$$\begin{aligned} L(w, p) &= f(x, p) + \sum_I y_i g_i(x, p) + \sum_J z_j h_j(x, p), \\ \tau(w, p) &= (\nabla_x L(w, p), -\nabla_y L(w, p), -\nabla_z L(w, p)). \end{aligned} \tag{4.1}$$

The principal result is

THEOREM 4.2. *Suppose, for some* $\bar{w} = (\bar{x}, \bar{y}, \bar{z}) \in \tilde{C}$ *and some* $\bar{p} \in P$, *that* (SSOC) *holds for* $(\mathrm{Q}_{\bar{p}})$. *Then there is a neighborhood* $\Pi \subset P$ *of* $\bar{p}$ *and a* $\underset{\sim}{C}^1$ *function* $w(p) = (x(p), y(p), z(p))$ *defined on* Π *such that* $w(\bar{p}) = \bar{w}$, *and for all* $p \in \Pi$, $w(p)$ *satisfies* (SSOC) *for* (Q_p). In particular, $x(p)$ is a local minimizer for (Q_p) *with unique multiplier vectors* $y(p)$ *and* $z(p)$. *Furthermore, there is a neighborhood* $\chi \subset R^n$ *of* $\bar{x}$ *such that for each* $p \in \Pi$, $x(p)$ *is the unique local minimizer for* (Q_p) *in* χ.

Proof. Fix $\bar{p}$ and $\bar{w}$ satisfying (SSOC) for $(\mathrm{Q}_{\bar{p}})$ and let $\bar{I} = \{i \in I\colon g_i(\bar{x}, \bar{p}) = 0\}$. By Theorem 3.5, $\bar{w}$ is a nondegenerate critical point for $L(\cdot, \bar{p})$ on $\tilde{G}$, where $\tilde{G} = F \times R_{++}^{\bar{I}} \times R^J$ is the face of $\tilde{C}$ containing $\bar{w}$ (and F is the face of C containing $\bar{x}$). This, and the implicit function theorem imply that there is a $\underset{\sim}{C}^1$ function $w(p)$ mapping a neighborhood $\Pi_0 \subset P$ of $\bar{p}$ into a neighborhood $\tilde{V}$ of $\bar{w}$ in $\tilde{G}$ such that

(4.3) (i) $w(\bar{p}) = \bar{w}$;

(ii) for all $p \in \Pi_0$, $w(p)$ is a nondegenerate critical point for $L(\cdot, p)$ on $\tilde{G}$;

(iii) if $p \in \Pi_0$, $w \in \tilde{V}$, and w is a critical point for $L(\cdot, p)$ on $\tilde{G}$, then $w = w(p)$.

By Theorem 3.5, we also have $-\tau(\bar{w}, \bar{p}) \in \operatorname{relint} N_{\tilde{C}}(\bar{w})$. By (4.3ii) and (1.1), $\nabla_w L(w(p), p) \in \tilde{G}^{\perp}_{w(p)}$ or equivalently (cf. Lemma 3.1(c)), $\tau(w(p), p) \in \tilde{G}^{\perp}_{w(p)}$ $(= L_{\tilde{C}}(w(p))^{\perp}$ by Theorem 2.8). So, applying Proposition 2.18 (with $\tilde{C}$ in place of C, $\tilde{G}$ in place of F),

(4.4) $-\tau(w(p), p) \in \operatorname{relint} N_{\tilde{C}}(w(p))$ for all p in a neighborhood of $\bar{p}$.

(In the terminology of Proposition 2.18, $(\bar{w}, -\tau(\bar{w}, \bar{p})) \in M$ so $(w(p), -\tau(w(p), p)) \in M$ for all p in a neighborhood of $\bar{p}$). In particular, $w(p)$ satisfies (SSOCi, ii, iii) for (Q_p).

Let ϕ be a local parametrization for F, $\Phi(\bar{q}) = \bar{x}$, $\bar{v} = (\bar{q}, \bar{y}, \bar{z})$. Also, let $q(p) = \phi^{-1}(x(p))$ and $v(p) = (q(p), y(p), z(p))$. By Theorem 3.5, (3.8b) holds, i.e., $v(\bar{p})$ satisfies (A) for $(\mathrm{Q}_{\phi, \bar{p}})$. This is equivalent, by (3.10), to (3.11) and (3.12). However, (3.11) and (3.12) clearly hold for v in a neighborhood of $\bar{v}$ if they hold at $\bar{v}$. So for all p in a neighborhood of $\bar{p}$, (3.8b) is satisfied (with $w(p)$ and $v(p)$ in place of $\bar{w}$ and $\bar{p}$). Since (3.8a) also holds by (4.4), Theorem 3.5 implies $w(p)$ satisfies (SSOC) for (Q_p).

Only the uniqueness claim of the theorem remains to be proved.

Claim 4.5. For all $(x, p) \in C \times P$ in a neighborhood of $(\bar{x}, \bar{p})$, if x is feasible for (Q_p), then (IC) is satisfied for (Q_p) at x.

Proof. For simplicity, we consider only the case $\bar{I} \neq \varnothing$, and $J = \varnothing$. If false, there are sequences $x^k \to \bar{x}$, $p^k \to \bar{p}$, and $a^k \in R^{\bar{I}}$ such that

$$\sum_{\bar{I}} a_i^k \nabla_x g_i(x^k, p^k) \in L_C(x^k)^{\perp}$$

and $|a^k| = 1$ for all k. Passing to a subsequence, it may be assumed that $a^k \to \bar{a} \neq 0$. By continuity of $\nabla_x g_i$ and because the multifunction $L_C(\cdot)^{\perp}$ has closed graph (Theorem 2.9),

$$\textstyle\sum_{\bar{I}} \bar{a}_i \nabla_x g_i(\bar{x}, \bar{p}) \in L_C(\bar{x})^{\perp},$$

contradicting (IC) for $(\mathrm{Q}_{\bar{p}})$ at $\bar{x}$. □

Claim 4.6. There exists a neighborhood N of $(\bar{y}, \bar{z})$ in $R^I \times R^J$ and a neighborhood M of $(\bar{x}, \bar{p})$ in $C \times P$, such that for every $(x, p) \in M$,

$$\nabla_x L(w, p) \in L_C(x)^{\perp}, \qquad y_i = 0 \quad \forall i \notin \bar{I} \tag{4.7}$$

admits only one solution $(y, z) \in R^I \times R^J$, and the function $(x, p) \to (y, z)$ is a $\underline{C}^1$ function mapping M into N.

Proof. Let (G_α, H_β, U) be a l.r. for C with $\bar{x} \in Z(A)$, and choose $A_0 \subset A$ minimal with respect to the property $s(A_0) = s(A)$. Then for any $x \in U$, $\Gamma(x, A_0)$ is a basis for span $\Gamma(x, A)$, so (4.7) implies

$$\nabla_x L(w, p) + \sum_{A_0} a_\alpha \nabla G_\alpha(x) + \sum_B b_\beta \nabla H_\beta(x) = 0$$

for some $a \in R^{A_0}$ and $b \in R^B$. For (x, p) near $(\bar{x}, \bar{p})$, the set

$$\Gamma(x, A_0) \cup \{\nabla_x g_i(x, p) \colon i \in \bar{I}\} \cup \{\nabla_x h_j(x, p) \colon j \in J\}$$

is linearly independent, since (IC) holds for $(\mathrm{Q}_{\bar{p}})$ at $\bar{x}$. The conclusion now follows by the implicit function theorem. □

Proof of Theorem 4.2 (continued). Suppose the uniqueness claim is false. Then there is a sequence $(x^k, p^k) \to (\bar{x}, \bar{p})$ in $C \times P$ such that for each k, $x^k \neq x(p^k)$ and x^k is a local minimizer for (Q_{p^k}). By Claim 4.5, for all large k (IC) holds for (Q_{p^k}) at x^k. Hence, by Theorem 3.3 there exists $(y^k, z^k) \in R^I_+ \times R^J$ such that

$$(4.8) \quad \begin{array}{lll} \text{(i)} & -\nabla_x L(w^k, p^k) \in N_C(x^k) & (w^k = (x^k, y^k, z^k)), \\ \text{(ii)} & \forall i \in I, \quad \text{if } y_1^k > 0 \quad \text{then} \quad g_i(x^k, p^k) = 0. & \end{array}$$

For large k, (4.8) implies that $y_i^k = 0, \forall i \notin \bar{I}$. So, by (2.7), Claim 4.6, and (4.8i), $(y^k, z^k) \to (\bar{y}, \bar{z})$. Hence, $\tau(w^k, p^k) \to \tau(\bar{w}, \bar{p})$. But $-\tau(\bar{w}, \bar{p}) \in \operatorname{relint} N_{\tilde{C}}(\bar{w})$ and by (3.1a), (4.8) implies

$$(4.9) \qquad -\tau(w^k, p^k) \in N_{\tilde{C}}(w^k).$$

So we can conclude from Proposition 2.10 that $w^k \in \tilde{G}$, and hence that $w^k \in \tilde{V}$, for k large. By (3.1c), (4.9) implies that w^k is a critical point for $L(\cdot, p^k)$ on $\tilde{G}$. Then (4.3iii) implies $w^k = w(p^k)$, and in particular $x^k = x(p^k)$, a contradiction that completes the proof of Theorem 4.2. □

Briefly, Theorem 4.2 states that (SSOC) is sufficient for $\underline{C}^1$ variation of the minimizer and multiplier vectors. The next (and final) result shows, in essence, that these are the weakest conditions implying this conclusion.

THEOREM 4.10. *Let $\bar{x}$ be a local minimizer for* $(\mathrm{Q}_{\bar{p}})$, *P open, $W \subset \tilde{C}$ a neighborhood of $\bar{w} = (\bar{x}, \bar{y}, \bar{z})$, $w(p) = (x(p), y(p), z(p))$ a $\underline{C}^1$ function: $P \to W$ such that $\bar{w} = w(\bar{p})$ and for each $p \in P$, $w(p)$ is the unique point in W satisfying*

$$(4.11) \qquad -\tau(w(p), p) \in N_{\tilde{C}}(w(p)).$$

Assume also that

(4.12) *the Jacobian at $\bar{p}$ of the map $p \to \nabla_w L(\bar{w}, p) \in R^r$ is of rank r.*

Then $\bar{w}$ satisfies (SSOC) *for* $(\mathrm{Q}_{\bar{p}})$.

(Note. Expression 4.11 is equivalent, by Lemma 2.1, to the assertion that $x(p)$ is feasible for (Q_p) and $w(p)$ satisfies the first-order conditions of Theorem 3.3 for (Q_p). To get a feeling for (4.12), the reader should interpret it for the case where $P = R^n \times R^I \times R^J$ and, for any $p = (v, s, t) \in P$, (Q_p) is the problem of minimizing $f(x) - x \cdot v$ subject to $g_i(x) \leqq s_i\ \forall i$, $h_j(x) = t_j$, $\forall j$ and $x \in C$).

Proof. Let $\bar{\zeta} = -\tau(\bar{w}, \bar{p})$. We will first show that $\bar{\zeta} \in \operatorname{relint} N_{\tilde{C}}(\bar{w})$. If not, then by a separation argument [10, 11.3], there is a vector μ such that

$$(4.13) \qquad 0 \neq \mu \in T_{\tilde{C}}(\bar{w}) \cap L_{\tilde{C}}(\bar{w})^{\perp}, \qquad \mu \cdot \bar{\zeta} = 0.$$

Fix any $\xi \in \operatorname{relint} N_{\tilde{C}}(\bar{w})$. Then (4.13) implies $\mu \cdot \xi < 0$. By (4.12) and the implicit function theorem, there is a $\underline{C}^1$ function $p(t)$ defined for t near $0 \in R$ such that

$$(4.14) \qquad p(0) = \bar{p}, \qquad \tau(\bar{w}, p(t)) + \bar{\zeta} + t\xi = 0.$$

For $t \geqq 0$, $-\tau(\bar{w}, p(t)) = \bar{\zeta} + t\xi \in N_{\tilde{C}}(\bar{w})$, so the uniqueness of $w(p(t))$ in (4.11) implies that $\bar{w} = w(p(t))$ for $t \geqq 0$. In particular,

$$(4.15) \qquad \left.\frac{d}{dt}\right|_{t=0} w(p(t)) = 0.$$

Let (G_α, H_β, U) be a l.r. for $\tilde{C}$ with $\bar{w} \in Z(A)$. By (2.1b), there is a number $M > 0$ such that for all (ζ, w) in a neighborhood of $(\bar{\zeta}, \bar{w})$,

$$(4.16) \qquad \begin{aligned} &\text{if } \zeta = \sum_A a_\alpha \nabla G_\alpha(w) + \sum_B b_\beta \nabla H_\beta(w) \text{ with } a \in R_+^A \\ &\text{and } b \in R^B, \text{ then } a_\alpha \leqq M \ (\forall \alpha) \text{ and } |b_\beta| \leqq M \ (\forall \beta). \end{aligned}$$

Define $\zeta(t) = -\tau(w(p(t)), p(t))$. Now,

$$\zeta(t) \cdot \mu = (\tau(\bar{w}, p(t)) - \tau(w(p(t)), p(t))) \cdot \mu - \tau(\bar{w}, p(t)) \cdot \mu.$$

So, by (4.15) and then by (4.13) and (4.14), we obtain

$$(4.17) \qquad \left.\frac{d}{dt}\right|_{t=0} (\zeta(t) \cdot \mu) = -\left.\frac{d}{dt}\right|_{t=0} (\tau(\bar{w}, p(t)) \cdot \mu) = \xi \cdot \mu < 0.$$

For each $\alpha \in A$ (resp., $\beta \in B$), define $\hat{G}_\alpha(t) = \mu \cdot \nabla G_\alpha(w(p(t)))$ (resp., $\hat{H}_\beta(t) = \mu \cdot \nabla H_\beta(w(p(t)))$). By (4.13) and (4.15),

$$(4.18) \qquad \hat{G}_\alpha(0) \leqq 0, \qquad \hat{H}_\beta(0) = 0, \qquad \hat{G}'_\alpha(0) = \hat{H}'_\beta(0) = 0 \quad \forall \alpha, \beta.$$

For some $a(t) \in R_+^A$, and $b(t) \in R^B$, we have by (4.11),

$$(4.19) \qquad \zeta(t) = \sum_A a_\alpha(t) \nabla G_\alpha(w(p(t))) + \sum_B b_\beta(t) \nabla H_\beta(w(p(t))) \in N_{\tilde{C}}(w(p(t))).$$

So by (4.16), for t near 0,

$$\mu \cdot \zeta(t) \leqq M\Big(\sum_A \max\{0, \hat{G}_\alpha(t)\} + \sum_B |\hat{H}_\beta(t)|\Big)$$

which is impossible in light of (4.13), (4.17) and (4.18). Thus

$$(4.20) \qquad \bar{\zeta} \in \operatorname{relint} N_{\tilde{C}}(\bar{w}).$$

Let $\tilde{G}$ be the face of $\tilde{C}$ containing $\bar{w}$. By Proposition 2.10, (4.19) and (4.20) imply $w(p) \in \tilde{G}$ for all p in a neighborhood of $\bar{p}$. Let $\tilde{\phi}: \tilde{U} \to \tilde{G}$ be a local parametrization for $\tilde{G}$ with $\tilde{\phi}(\bar{v}) = \bar{w}$, $\tilde{U} \subset R^c$ open ($c = \dim \tilde{G}$). Define $v(p) = \tilde{\phi}^{-1}(w(p))$. By (4.1) and (4.11), $\nabla_v L(\tilde{\phi}(v(p)), p) = 0$. Differentiating this with respect to p gives $HD + E = 0$ where H is the Hessian at $\bar{v}$ of the function $v \to L(\tilde{\phi}(v), p)$, D is the Jacobian of $v(p)$ at $\bar{p}$, and E is the Jacobian at $\bar{p}$ of the map $p \to \nabla_v L(\bar{w}, p)$. By (4.12), E has (full) rank c, so the $(c \times c)$-matrix H must be invertible. Thus $\bar{w}$ is a nondegenerate critical point for L on $\tilde{G}$. This and (4.20) show, by Theorem 3.5, that $\bar{w}$ satisfies (SSOC) for $(\mathrm{Q}_{\bar{p}})$. □

Acknowledgment. I wish to thank my thesis advisor, Professor R. T. Rockafellar, for his many helpful comments and suggestions, and a referee for a thoughtful job in editing the manuscript.

REFERENCES

[1] L. Auslander and R. MacKenzie, *Introduction to Differentiable Manifolds*, McGraw-Hill, New York, 1963.

[2] A. V. Fiacco, *Sensitivity analysis for nonlinear programming using penalty methods*, Math. Programming, 10 (1976), pp. 287–311.

[3] J. GAUVIN AND J. W. TOLLE, *Differential stability in nonlinear programming*, this Journal, 15 (1977), pp. 294–311.

[4] B. GRÜNBAUM, *Convex Polytopes*, Interscience, New York, 1967.

[5] M. R. HESTENES, *Optimization Theory, the Finite Dimensional Case*, John Wiley, New York, 1975.

[6] M. HIRSCH, *Differential Topology*, Springer-Verlag, New York, 1976.

[7] G. P. MCCORMICK, *Optimality criteria in nonlinear programming*, SIAM-AMS Proceedings 9 (1976), W. Cottle and C. E. Lemke, eds., American Mathematical Society, Providence, RI, 1976, pp. 27–38.

[8] J. W. MILNOR, *Topology from the Differentiable Viewpoint*, University Press of Virginia, 1965.

[9] S. M. ROBINSON, *Generalized equations and their solutions, Part I: Basic theory*, Tech. summary rep. 1812, Mathematics Research Center, Univ. of Wisconsin, Madison, WI, 1977.

[10] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1972.

[11] J. E. SPINGARN, *On optimality conditions for structured families of nonlinear programming problems*, forthcoming.

[12] ———, *Second-order optimality conditions that are necessary with probability one*, Proceedings, Symposium on Mathematical Programming with Data Perturbations, George Washington University, May 1979, to appear.

[13] ———, *Generic conditions for optimality in constrained minimization problems*, Dissertation, Dept. of Mathematics, Univ. of Washington, Seattle, WA., 1977.

[14] J. E. SPINGARN AND R. T. ROCKAFELLAR. *The generic nature of optimality conditions in nonlinear programming*, Math of O.R. 4 (1979).

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 3, May 1980

0363-0129/80/1803-0006$01.00/0

A NOTE ON STABILIZATION OF INFINITE DIMENSIONAL LINEAR OSCILLATORS BY COMPACT LINEAR FEEDBACK*

J. S. GIBSON†

Abstract. This note points out the fact that a linear oscillator in an infinite dimensional Hilbert space, with no uniform decay rate, cannot be given a uniform decay rate with compact linear feedback. The motivation for the analysis here is the use of a finite number of control elements to stabilize a system with an infinite number of modes of vibration, and the implications of the inability to produce a uniform decay rate are elaborated in regard to optimal regulation. For systems with and without inherent damping, the result is based on approximating a compact operator with a sequence of finite dimensional operators. The physical interpretation of this technique is discussed.

Introduction. A considerable amount of literature has been devoted to the problem of using a linear feedback control to stabilize a linear oscillator in an infinite dimensional Hilbert space. See, for example, [1], [10], [11], [12], [13]. In practice, there are two methods of implementing a stabilizing feedback: active control and passive control. For active control, one or more actuators, i.e., power sources, are attached to the system and activated according to an appropriate control law based on the observed state of the system, whereas, for the second case, passive control elements such as dampers and springs are attached to the system to dissipate energy. In either case a finite number of control elements are added to the original oscillating system, and thus, when the linear operator representing the feedback control in the system differential equation is bounded, it is compact.

While an oscillator can usually be stabilized—i.e., for any initial condition, all the energy can be dissipated—with a finite number of control elements, the purpose of this note is to point out the fact that, if the original system has an infinite number of modes with no uniform decay rate, then no compact linear feedback will yield a uniform decay rate. (The precise meaning of this statement will become clear.) Russell [10] proved this result for the case of a linear oscillator with only compact velocity feedback, using different reasoning than we use here.

The principle underlying the whole development here is the fact that a compact linear operator on a Hilbert space can be approximated uniformly by its projections on the members of any increasing sequence of finite dimensional subspaces whose union is dense. Although we will elaborate on the idea, the recollection of this approximation result makes the no-uniform-decay result almost obvious for an infinite dimensional oscillator for which a compact linear control produces the only energy dissipation. However, the extension to more realistic systems with some inherent damping, but no uniform decay rate, is a little more difficult.

Finally, some rather disquieting conclusions about optimal regulation of systems of the type considered here can be drawn from the two theorems of this note and the results of Datko [3] and the present author [4].

The originally undamped oscillator. Consider the differential equation

$$\ddot{x}(t)+A_0x(t)=B_1\dot{x}(t)+B_2x(t), \tag{1}$$

where $x(t)$ is in a real separable Hilbert space H, A_0 is a self-adjoint linear operator from $D(A_0)$, which is dense in H, to H, and B_1 and B_2 are compact elements of $\mathscr{L}(H, H)$.

* Received by the editors September 12, 1978, and in revised form May 21, 1979.

† Mechanics and Structures Department, School of Engineering and Applied Science, University of California, Los Angeles, California 90024.

Also, we assume that the further characteristics of the linear oscillator hold: there exists $c>0$ such that

$$(2) \qquad \langle A_0x, x\rangle_H \geqq c\|x\|_H^2, \qquad x\in D(A_0),$$

and A_0^{-1} is compact from H to H.

Assuming that H is infinite dimensional, we know that the spectrum of A_0 is an infinitely increasing sequence of positive real eigenvalues ω_n^2, each of finite multiplicity, and that the mutually orthogonal eigenvectors ϕ_n of A_0 form a complete basis in H. As usual, we define the "energy space" by $E=V\times H$, where $V=D(A^{1/2})$ has inner product $\langle v_1, v_2\rangle_V=\langle A_0^{1/2}v_1, A_0^{1/2}v_2\rangle_H$ and E has the energy inner product $\langle (v_1, h_1), (v_2, h_2)\rangle_E=\langle v_1, v_2\rangle_V+\langle h_1, h_2\rangle_H$. The eigenvectors of A_0 are also mutually orthogonal and complete in V, and the pairs $(\phi_n, 0)$ and $(0, \phi_n)$ are thus mutually orthogonal and complete in E.

Next, we define an operator A in E by

$$(3) \qquad A=\begin{bmatrix} 0 & I \\ -A_0 & 0\end{bmatrix}, \qquad D(A)=D(A_0)\times V.$$

This A generates a strongly continuous group $T(\cdot)$ on E, and we have conservation of energy:

$$(4) \qquad \|T(t)y\|_E=\|y\|_E, \qquad y\in E, \quad -\infty<t<\infty.$$

With $B\in\mathscr{L}(E, E)$ defined by

$$(5) \qquad B=\begin{bmatrix} 0 & 0 \\ B_2 & B_1\end{bmatrix},$$

$A+B$, with domain $D(A)$, generates a strongly continuous group $S(\cdot)$ on E, and we have the first of our two results:

THEOREM 1. *Let A be the operator of* (3) *and B the operator of* (5); *assume the hypotheses stated for A and B, including the compactness of B_1 and B_2. Then the group $S(\cdot)$ generated by $A+B$ is not uniformly exponentially stable.*[1]

Proof. Let $V_n=H_n=\operatorname{span}\{\phi_j\}_{j\leqq n}$, and let Λ_n be the projection operator from E to $V_n\times H_n$. Then, since B is compact, it can be approximated arbitrarily closely in $\mathscr{L}(E, E)$ by the sequence of operators $B_n=\Lambda_nB\Lambda_n$ (see [9, p. 204]). Now, if $S(\cdot)$ is uniformly exponentially stable, the group $S_n(\cdot)$ on E, generated $A+B_n$, is uniformly exponentially stable for $\|B-B_n\|_{\mathscr{L}(E,E)}$ sufficiently small, but not zero (see [5, p. 390], or [6, p. 498]). However, since

$$(6) \qquad \|S_n(t)(\phi_m, 0)\|_E=\|(\phi_m, 0)\|_E, \qquad -\infty<t<\infty, \quad 1\leqq n<m,$$

$S_n(\cdot)$ can never be uniformly exponentially stable, and therefore neither can $S(\cdot)$ be.

As already mentioned, Russell [10] proved this result for the case $B_2=0$, and it appears that his argument, though substantially different from the one here, can be extended easily to obtain Theorem 1 with $B_2\neq 0$. However, it does not appear that Russell's argument could yield anything resembling Theorem 2 of this note. We will refer to some of the positive results of [10] later.

[1] A group (or semi-group) $S(\cdot)$ is said to be uniformly exponentially stable if there are positive constants M and α such that $\|S(t)\|\leqq Me^{-\alpha t}$ for all $t\geqq 0$. This is equivalent to saying that $S(\cdot)$ has a uniform decay rate, i.e., there are constants $t_0>0$ and $r<1$ such that $\|S(t_0)\|\leqq r$.

Systems with inherent damping. Of course, physical systems always have some inherent damping, so a more realistic version of the uncontrolled oscillator should be

$$\ddot{x}(t)+B_0\dot{x}(t)+A_0x(t)=0, \tag{7}$$

where $B_0\in\mathscr{L}(H,H)$ and

$$\langle B_0x, x\rangle\geqq 0, \qquad x\in H. \tag{8}$$

The operator $\hat{A}$ defined by

$$\hat{A}=\begin{bmatrix} 0 & I \\ -A_0 & -B_0\end{bmatrix}, \qquad D(\hat{A})=D(A), \tag{9}$$

generates a strongly continuous group $\hat{T}(\cdot)$ on E, with

$$\|\hat{T}(t)\|_E\leqq 1, \qquad t\geqq 0. \tag{10}$$

As a matter of fact, in most physical systems there is complete damping, however slight; i.e.,

$$\|\hat{T}(t)y\|_E\to 0 \quad \text{as } t\to\infty,\ y\in E. \tag{11}$$

When (11) holds, we say that $\hat{T}(\cdot)$ is strongly stable.

We should note here that the notion of "weak stability" is sometimes used (see [2], [12], [13]): a semigroup $T(\cdot)$ on a Hilbert space E is said to be weakly stable if

$$\langle T(t)x, y\rangle_E\to 0 \quad \text{as } t\to\infty,\ \forall x, y\in E. \tag{12}$$

However, for the damped oscillator represented by the $\hat{T}(\cdot)$ generated by the $\hat{A}$ of (9), weak stability is equivalent to strong stability because, as Benchimol has shown (see the proof of Corollary 3.1 of [2]), a weakly stable C_0 semigroup whose generator has compact resolvent is also strongly stable. Now, as is well known, the compactness of A^{-1} and $\hat{A}^{-1}$ follow from the hypotheses already stated: since A_0^{-1} is compact from H to H, $A_0^{-1/2}$ is compact from H to H, and thus the injection from V into H is compact; since $A_0^{-1/2}$ is compact from H to H and bounded from H to V, $A_0^{-1/2}$ is compact from H to V and $A_0^{-1}B_0$ is compact from V to V. Hence A^{-1} and $\hat{A}^{-1}$ are compact operators on E.

It is natural to ask then whether we can use a compact feedback control to give a uniform decay rate to a system which is already strongly stable, but not uniformly exponentially stable. The answer is no, as the following theorem states.

THEOREM 2. *Let E be a Hilbert space and $T(\cdot)$ a strongly stable strongly continuous contraction semigroup on E, with generator A. Let B be a compact element of $\mathscr{L}(E,E)$ and let $S(\cdot)$ be the semigroup generated by $A+B$. Then, if $S(\cdot)$ is uniformly exponentially stable, so is $T(\cdot)$.*

Proof. Since B can be approximated in $\mathscr{L}(E,E)$ by its projection onto finite dimensional subspaces,[2] we need only show that, if E_n is a finite dimensional subspace of E, Λ_n is the projection operator from E to E_n, and $B_n=\Lambda_nB_n$, then the semigroup $S_n(\cdot)$ generated by $A+B_n$ cannot be uniformly exponentially stable.

We have

$$S_n(t)x=T(t)x+\int_0^t S_n(t-\eta)B_nT(\eta)x\,d\eta, \qquad t\geqq 0, \quad x\in E. \tag{13}$$

[2] It is not necessary to require E to be separable; see [9], p. 205.

Assuming there is a constant M such that

$$\|S_n(t)\| \leqq M, \qquad t \geqq 0, \tag{14}$$

we then have

$$\|S_n(t)x\| \geqq \|T(t)x\| - Mt \sup_{0 \leqq \eta \leqq t} \|B_n T(\eta)x\|, \qquad t \geqq 0, \quad x \in E. \tag{15}$$

Let $t_1 > 0$, $0 < \varepsilon < Mt_1\|B_n\|$, and $\delta = \varepsilon/(2Mt_1\|B_n\|)$. Since E_n is finite dimensional and $T(\cdot)$ is a strongly stable semigroup, we can choose $t_0 > 0$ such that

$$\sup_{\substack{x \in E_n, \\ \|x\|=1}} \|T(t)x\| \leqq \frac{\delta^2}{4}, \qquad t \geqq t_0. \tag{16}$$

Since $T(\cdot)$ is a contraction semigroup and not uniformly exponentially stable, there is an $x_0 \in E$, with $\|x_0\| = 1$, such that

$$\max\left\{1 - \frac{\delta^2}{4}, 1 - \frac{\varepsilon}{2}\right\} \leqq \|T(t_0 + t_1)x_0\| \leqq \|T(t_1)x_0\|. \tag{17}$$

For $t \geqq 0$, write $T(t)x_0 = y(t) + z(t)$, where $y(t) \in E_n$ and $z(t) \in E_n^\perp$. If, for some $\eta \in [0, t_1]$, $\|y(\eta)\| > \delta$, then $\|T(\eta)x_0\|^2 = \|y(\eta)\|^2 + \|z(\eta)\|^2 \leqq 1$ implies $\|z(\eta)\|^2 < 1 - \delta^2 \leqq (1 - \delta^2/2)^2$, and

$$\begin{aligned} \|T(t_0 + t_1)x_0\| &= \|T(t_0 + t_1 - \eta)T(\eta)x_0\| \\ &= \|T(t_0 + t_1 - \eta)(y(\eta) + z(\eta))\| < \frac{\delta^2}{4} + 1 - \frac{\delta^2}{2} = 1 - \frac{\delta^2}{4}. \end{aligned} \tag{18}$$

Hence

$$\|y(\eta)\| \leqq \delta, \qquad 0 \leqq \eta \leqq t_1, \tag{19}$$

and

$$\sup_{0 \leqq \eta \leqq t_1} \|B_n T(\eta)x_0\| \leqq \|B_n\|\delta. \tag{20}$$

Then (15), (17), and (20) show

$$\|S_n(t_1)x_0\| \geqq 1 - \varepsilon, \tag{21}$$

and the arbitrariness of t_1 and ε imply that $S_n(\cdot)$ is not uniformly exponentially stable.

Physical interpretation of the theorem proofs. In view of the current prominence of modal control techniques for flexible systems (see [1] and its references), it might be useful to elaborate briefly on the physical interpretation of the proofs of Theorems 1 and 2. The eigenvectors ϕ_n of A_0 are of course the natural modes of vibration for the undamped linear oscillator, and the proof of Theorem 1 relies on the fact that they are mutually orthogonal—i.e., the modes of free vibration are uncoupled. While the feedback control represented by $B_n = \Lambda_n B \Lambda_n$ is physically unrealistic because both observation and control action are restricted to the space $V_n \times H_n$, spanned by the first n modes, it is legitimate mathematically, and useful for the theorem because it keeps the first n modes uncoupled from the remaining modes.

Although Theorem 2 is not restricted to second order oscillators, both the theorem and its proof were motivated by the damped system described by (7). In general, the damping represented by the operator B_0 couples the modes of vibration, so that a more difficult proof is required for Theorem 2; however, the idea is not very different. Due to the possible coupling, we may not be able, as we were in the case of Theorem 1, to select an initial condition for which all the energy stays out of the first n modes; but, for any time $t_1 > 0$, we can choose an initial condition for which all but an arbitrarily small fraction of the initial energy stays out of the first n modes for $0 \leqq t \leqq t_1$.

Of course both theorems result from our ability to approximate uniformly (in norm and on bounded time intervals) the semigroup $S(\cdot)$, representing the system with compact feedback, by the sequence of semigroups $S_n(\cdot)$, representing systems in which both observation and control action are restricted to the first n modes. Actually, for Theorem 1, we could take either $B_n = B\Lambda_n$ or $B_n = \Lambda_n B$, while, for Theorem 2, we could take $B_n = B\Lambda_n$. $B_n = B\Lambda_n$ means that observation is restricted to the first n modes, while $B_n = \Lambda_n B$ means that control action is restricted to the first n modes.

Implications for optimal regulation. From the results just presented, we can draw some interesting, if frustrating, conclusions about the linear-quadratic regulator problem on the infinite interval, for infinite dimensional systems (see [3], [4], [7], [8]). The problem is stated as follows. Let E and U be Hilbert spaces, A generate a strongly continuous semigroup $T(\cdot)$ on E, and $B \in \mathscr{L}(U, E)$; the state vector is $x(t) \in E$, the control vector is $u(t)$, $u(\cdot) \in L_2((0, \infty); U)$, and the input-output relation is

$$x(t) = T(t)x(0) + \int_0^t T(t-\eta)Bu(\eta)\,d\eta, \qquad t \geqq 0, \quad x(0) \in E. \tag{22}$$

The performance index is

$$J(x(0), u) = \int_0^\infty (\langle Dx(\eta), x(\eta)\rangle_E + \langle Qu(\eta), u(\eta)\rangle_U)\,d\eta, \tag{23}$$

where D is a nonnegative, self-adjoint element of $\mathscr{L}(E, E)$ and Q is a positive definite, self-adjoint element of $\mathscr{L}(U, U)$. For the rest of this discussion, let us assume that D is also positive definite.

Of course, for a given initial condition $x(0)$, there may be no control $u(\cdot)$ for which $J(x(0), u) < \infty$. It is shown in [4] (see Theorem 4.11) that there exists a nonnegative, self-adjoint operator $P \in \mathscr{L}(E, E)$ satisfying the Riccati equation

$$A^*P + PA - PBQ^{-1}B^*P + D = 0 \tag{24}$$

if and only if, for each $x(0) \in E$, there is a $u(\cdot)$ such that $J(x(0), u) < \infty$. In this case P is unique and the optimal control is given by

$$u(t) = -Q^{-1}B^*Px(t), \qquad t \geqq 0. \tag{25}$$

This means that the optimal state is given by $x(t) = S(t)x(0)$, where $S(\cdot)$ is the semigroup generated by $A - BQ^{-1}B^*P$, and (see [3], [4]) $S(\cdot)$ is uniformly exponentially stable.

Now, if U is finite dimensional, as it always is in active control and usually is in passive control (see, for example, [1] and [11]), $BQ^{-1}B^*P$ is compact. So, if $T(\cdot)$ represents the undamped linear oscillator discussed here, or if it represents a damped system and satisfies the hypotheses of Theorem 2, we know that there must be an initial condition $x(0)$ for which there is no control $u(\cdot) \in L_2((0, \infty); U)$ such that $J(x(0), u) < \infty$. Furthermore, there can be no bounded nonnegative, self-adjoint solution of the Riccati equation; indeed the notion of an optimal closed loop control becomes unclear.

Various sufficient conditions have been given (see [2], [10], [11], [12], [13]) for linear oscillators to be stabilizable by linear feedback, and, for the system (1) with $B_2 = 0$ and under certain hypotheses on B_1, Russell [10] has derived decay rates that show that the performance index of (23) can be made finite with a velocity feedback, for initial conditions in $D(A)$. We might speculate then that we should seek a feedback control which minimizes $J(x(0), u)$ whenever it can be made finite, and, possibly, makes the system (1) strongly stable. Unfortunately, as we know from the discussion above, such a control cannot be based on a bounded solution of the Ricatti equation; however, the next best thing to a bounded operator is a closed operator, and this seems a definite possibility for determining an optimal control scheme for oscillators of the type considered here.

REFERENCES

[1] M. J. BALAS, *Modal control of certain flexible dynamic systems*, this Journal, 16 (1978), pp. 450–462.

[2] C. D. BENCHIMOL, *A note on weak stabilizability of contraction semigroups*, this Journal, 16, (1978), pp. 373–379.

[3] R. DATKO, *A linear control problem in abstract Hilbert space*, J. Differential Equations, 9 (1971), pp. 346–359.

[4] J. S. GIBSON, *The Riccati integral equations for optimal control problems on Hilbert spaces*, this Journal, 17 (1979), pp. 537–565.

[5] E. HILLE AND R. S. PHILLIPS, *Functional Analysis and Semigroups*, Colloquium Publications, Vol. 31, American Mathematical Society, Providence, 1957.

[6] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.

[7] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, 1971.

[8] D. L. LUKES AND D. L. RUSSELL, *The quadratic criterion for distributed systems*, this Journal, 7, (1969), pp. 101–121.

[9] F. RIESZ AND B. SZ.-NAGY, *Functional Analysis*, Ungar, New York, 1955.

[10] D. L. RUSSELL, *Decay rates for weakly damped systems in Hilbert space obtained with control-theoretic methods*, J. Differential Equations, 19 (1975), pp. 344–370.

[11] ———, *Linear stabilization of the linear oscillator in Hilbert space*, J. Math. Anal. Appl., 25 (1969), pp. 663–675.

[12] M. SLEMROD, *A note on complete controllability and stabilizability for linear control systems in Hilbert space*, this Journal, 12 (1974), pp. 500–508.

[13] ———, *The linear stabilization problem in Hilbert space*, J. Functional Analysis, 11 (1972), pp. 334–345.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 3, May 1980

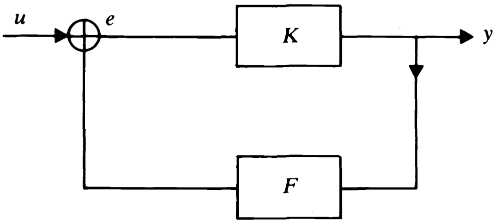
0363-0129/80/1803-0007$01.00/0

STRONG CAUSALITY CONDITIONS AND CAUSAL INVERTIBILITY*

AVRAHAM FEINTUCH†

Abstract. Various strong causality conditions for linear systems are studied and compared. The relationship between these conditions and the causal invertibility problem is considered.

1. Introduction. Let $(\mathscr{H}, \mathscr{E})$ be a Hilbert resolution space, and consider the linear feedback system



described by the equations

$$y = Ke, \qquad e = Fy + u,$$

where K and F are bounded linear causal operators on $(\mathscr{H}, \mathscr{E})$. The relationship between the well-posedness and stability of the above system and the existence of the operator $(I-KF)^{-1}$ as a bounded causal operator is well-known and has been studied by various authors ([1], [6], [10], [12], [13]). Unfortunately, even when $(I-KF)^{-1}$ exists as a bounded linear operator, it may not be causal. The simplest example of such a situation is as follows. Suppose $\mathscr{H} = L^2(-\infty, \infty; \mu)$, where μ is Lebesque measure and $\mathscr{E}$ is the usual family of truncation projections. Let S_t be the shift operator on $\mathscr{H}$ defined by

$$(S_tf)(a) = f(a-t), \qquad t > 0.$$

Then S_t is causal but its inverse is anticausal. Thus if $S_t = I - KF$, $(I-KF)^{-1}$ exists as a bounded linear operator but is not causal.

This gave rise to the causal invertibility problem: When is the inverse of a causal operator causal? The appearance of [2] has shown the importance of this problem in optimal control as well.

One of the ways to attack this problem is to strengthen the causality hypothesis in various ways, usually involving a delay type condition or Lipshitz condition in the hope that this will give positive results to the problem. Here we will discuss a number of such strengthenings of the causality hypothesis. In particular we will consider the strict causality of DeSantis, Porter, Saeks (see [10]) and the strong causality of Willems [12] and Zames [13]. We show that strong causality (while historically appearing first) is a natural extension of strict causality.

In the last part of the paper, we give a very general formulation for the physical notion of delay which seems to include the delay type strong causality conditions discussed above. Mathematically, this will mean that the operator has zero memoryless part. This is done by introducing a mapping from the algebra of all bounded linear operators to the algebra of memoryless bounded linear operators. The null-space of this mapping will contain the strongly causal operators.

* Received by the editors January 24, 1979, and in revised form October 10, 1979.

† Department of Mathematics, Ben Gurion University of the Negev, Beersheva, Israel. The results in this paper were obtained while the author was visiting at Centre de Recherche Mathematique, Université de Montreal and Department of Electrical Engineering, McGill University, Montreal, Quebec, Canada.

The main results of this paper are necessary and sufficient conditions for the inverse of $I-T$ to exist and be causal when T is either strictly or strongly causal.

There are a number of results that appear in this paper that are not new but have appeared in the operator theory literature. These play a transitionary role here, and are necessary for setting up the apparatus needed to study the systems theoretic notions. The appropriate references are mentioned when this is done.

2. Causality and strict causality. Let $\mathscr{H}$ be a complex separable Hilbert space and $\mathscr{E}=\{P^t : t\in\Gamma\}$ be a resolution of the identity on $\mathscr{H}$. $\mathscr{B}(\mathscr{H})$ will denote the algebra of bounded linear operators on $\mathscr{H}$. Then an operator $A\in\mathscr{B}(\mathscr{H})$ is causal if $P^tA=P^tAP^t$.

The concept of strict causality is described here very briefly. For a detailed discussion the reader is referred to [5], [10]. A partition $\mathscr{P}$ of $\mathscr{E}$ is a finite subset of $\mathscr{E}$ containing 0 and I. For any partition $\mathscr{P}=\{0<P^{t_1}<\cdots<P^{t_n}=I\}$ define $\Delta P_i=P^{t_i}-P^{t_{i-1}}$. Any $A\in\mathscr{B}(\mathscr{H})$ can then be written as

$$A=\sum_{i<j}\Delta P_iA\Delta P_j+\sum_{i=1}^{n}\Delta P_iA\Delta P_i+\sum_{j<i}\Delta P_iA\Delta P_j.$$

Simple manipulations allow us to rewrite this expression as

$$A=\sum_{i=1}^{n}P^{t_{i-1}}A\Delta P_i+\sum_{i=1}^{n}\Delta P_iA\Delta P_i+\sum_{i=1}^{n}\Delta P_iAP^{t_{i-1}}.$$

If A is causal, it is easy to see that $\sum_{i=1}^{n}p^{t_{i-1}}A\Delta P_i=0$, and thus

$$A=\sum_{i=1}^{n}\Delta P_iA\Delta P_i+\sum_{i=1}^{n}\Delta P_iAP^{t_{i-1}}.$$

DEFINITION 2.1. The causal operator $A\in\mathscr{B}(\mathscr{H})$ is *strictly causal* if, given $\varepsilon>0$, there exists a partition $\mathscr{P}$ such that for any refinement $\mathscr{P}_1$ of $\mathscr{P}$,

$$\left\|\sum_{i=1}^{n}\Delta P_iA\,\Delta P_i\right\|<\varepsilon,$$

where ΔP_i is constructed from the projections in $\mathscr{P}_1$.

This is generally denoted by $\int dPA\,dP=0$.

The strictly causal operators have been studied in detail in [5], [1], [10]. The main property that will be used here is that the family of strictly causal operators forms a uniformly closed two-sided ideal of quasinilpotent operators (T is quasinilpotent if its spectrum is $\{0\}$) in the algebra of causal operators which is maximal in the sense that it contains all two-sided ideals of quasinilpotent operators. Our first result will follow from this remark.

LEMMA 2.2. *Suppose A is strictly causal and B is causal. Then $I-B$ has a causal inverse if and only if $I-(A+B)$ does.*

Proof. Suppose $(I-B)^{-1}$ exists and is causal. Note that $(I-A-B)=(I-B)[I-(I-B)^{-1}A]$. Since A is strictly causal, by the above remarks $(I-B)^{-1}A$ is strictly causal, and therefore $\sigma([I-B]^{-1}A)=\{0\}$. Thus $[I-(I-B)^{-1}A]$ has a causal inverse and so does $(I-A-B)$. On the other hand, if $I-(A+B)$ has a causal inverse, then

$$I-B=[I-A-B]+A=[I-A-B](I+[I-A-B]^{-1}A),$$

and the same argument applies. This completes the proof.

Now suppose A is causal and consider $I-A$. It was shown in [1], [5] that if A is strictly causal, then $(I-A)^{-1}$ exists and is causal. Various generalizations of this were given in [10, Chap. 2]. Here we present a result that includes all those stated there and seems to be the best possible result involving strict causality considerations.

We recall that if A is causal and $\mathcal{P} = \{0 < P^{t_1} < \cdots < P^{t_n} = I\}$ is any partition, then

$$A = \sum_{i=1}^{n} \Delta P_i A P^{t_{i-1}} + \sum_{i=1}^{n} \Delta P_i A \,\Delta P_i.$$

The question: When do these sums converge, is a special case of the additive decomposition problem considered in [7], [10]. It is clear that if one of the sums converges, then so does the other. The case where A is strictly causal is the extreme case where the second sum converges to zero and the first sum converges to A. In fact, if $\int dPA\,dP$ exists, then it will be memoryless and $A - \int dPA\,dP$ is strictly causal. Thus A strictly causal means A has no memoryless part. A generalization of this idea will appear later. For the present we note that the existence of $\int dPA\,dP$ gives a unique decomposition of A into a strictly causal and memoryless part.

THEOREM 2.3. *Suppose $A = A_1 + A_2$, where A_1 is strictly causal and A_2 is memoryless. Then $I - A$ has a causal inverse if and only if $I - A_2$ is invertible.*

Proof. Suppose $I - A = I - A_1 - A_2$ has a causal inverse. Since A_1 is strictly causal, Lemma 2.2 implies that $I - A_2$ is invertible.

On the other hand if $I - A_2$ is invertible, it has a causal inverse. For $(I - A_2)$ is memoryless and, therefore, $(I - A_2)P^t = P^t(I - A_2)$ for all t. It is immediate that $(I - A_2)^{-1}P^t = P^t(I - A_2)^{-1}$. We now apply the other half of Lemma 2.2 and the proof is complete.

It is natural to ask when does $\int dPA\,dP$ converge. This can be answered by reinterpreting a result of Larson [14]. We first present a simple lemma which, we feel, has great significance, especially for state decompositions (see [2], [11]). This is actually implicit in [5, Thm. 9.6].

LEMMA 2.4. *If $T \in \mathcal{B}(\mathcal{H})$ and $P_t = I - P^t$, then $P_t T P^t$ is strictly causal.*

Proof. Note that $(P_t T P^t)^2 = 0$. We show that in fact the two-sided ideal generated by $P_t T P^t$ in the algebra of causal operators is nilpotent. Clearly $P_t T P^t$ is causal. If A is causal, then

$$(AP_t T P^t)^2 = AP_t T P^t A P_t T P^t = AP_t T P^t (P_t A P_t) T P^t = 0.$$

Here we used the fact that A causal implies $AP_t = P_t A P_t$.

In the same way one shows that $(P_t T P^t A)^2 = 0$. Since the strictly causal operators are the maximal quasinilpotent two-sided ideal in the algebra of causal operators, it follows that $P_t T P^t$ is strictly causal. This completes the proof.

Note that if T is causal, then $P_t T P^t = (I - P^t) T P^t = TP^t - P^t T P^t = TP^t - P^t T$. Let $\mathcal{R}$ denote the strongly closed algebra of operators generated by $\{P^t : t \in \Gamma\}$. The condition for the existence of $\int dPA\,dP$ can now be stated as follows.

THEOREM 2.5. *Suppose A is causal. Then $\int dPA\,dP$ exists if and only if for all $E \in \mathcal{R}$; $AE - EA$ is strictly causal.*

3. Strong causality. In this section, we consider the notion of strong causality presented in [12], [13]. While the discussion there is quite general, here we restrict ourselves to linear systems. We present, *for this case*, an equivalent formulation. This will make the relationship with strict causality more transparent. In fact we will see that the strictly causal operators are a proper subclass of the strongly causal ones. We then give a necessary and sufficient condition for the causal invertibility of $I - A$ when A is strongly causal.

From this point on we assume that $\Gamma \subset (-\infty, \infty)$.

DEFINITION 3.1. Suppose A is a causal operator on $(\mathcal{H}, \mathcal{C})$. A is *strongly causal* if for any $\varepsilon > 0$ there exists $\Delta t > 0$ such that for any $t' \in \Gamma$, $P^{t'}u = 0$ implies $\|P^{t'+\Delta t}Au\| \leqq \varepsilon \|P^{t'+\Delta t}u\|$.

This definition corresponds to the linear case of the definition of [12]. We reformulate it in the spirit of § 2. $\Delta P_{t'}$ will denote $P^{t'+\Delta t} - P^{t'}$.

LEMMA 3.2. *A is strongly causal if for any $\varepsilon > 0$ there exists $\Delta t > 0$ such that for all $t' \in \Gamma$,*

$$\|\Delta P_{t'} A \; \Delta P_{t'}\| < \varepsilon.$$

Proof. Fix $\varepsilon > 0$. Then there exists $\Delta t > 0$ such that if $t' \in \Gamma$ and $P^{t'}u = 0$,

$$\frac{\|P^{t'+\Delta t}Au\|}{\|P^{t'+\Delta t}u\|} < \varepsilon.$$

Since $P^{t'}u = 0$, we have $P^{t'+\Delta t}u = \Delta P_{t'}u$ and by causality,

$$P^{t'+\Delta t}A = P^{t'+\Delta t}AP^{t'+\Delta t}$$

and

$$P^{t'}A \; \Delta P_{t'} = 0.$$

Thus we may rewrite the above expression as

$$\frac{\|\Delta P_{t'} A \; \Delta P_{t'} u\|}{\|\Delta P_{t'} u\|} < \varepsilon.$$

Taking the supremum over all such $u \in \mathcal{H}$ gives $\|\Delta P_{t'} A \; \Delta P_{t'}\| < \varepsilon$.

We can now clarify the relationship between strict and strong causality. We assume that $\Gamma = [0, \infty)$. The case $\Gamma = (-\infty, \infty)$ can be handled in a similar way.

DEFINITION 3.3. A generalized partition of $\mathcal{C}$ is a *sequence* $\{P^{t_i}\}$ from $\mathcal{C}$ such that $\bigvee_i \Delta P_i = I$ (or equivalently $P^{t_i} \to I$ strongly as $t_i \to \infty$).

THEOREM 3.4. *A is strongly causal on $(\mathcal{H}, \mathcal{C})$ if and only if given $\varepsilon > 0$ there exists a generalized partition $\mathcal{P} = \{P^{t_i}\}$ in $\mathcal{C}$ such that $\|\Delta P_i A \; \Delta P_i\| < \varepsilon$ for all i.*

Proof. Suppose A is strongly causal and let $\varepsilon > 0$. Then there exists $\Delta t > 0$ such that for any $t' \in \Gamma$, $\|\Delta P_{t'} A \; \Delta P_{t'}\| < \varepsilon$. Then just choose the partition $\mathcal{P} = \{P^{i\Delta t} : i = 1, 2, \cdots\}$. This satisfies the requirements.

On the other hand, suppose A has a generalized partition $\mathcal{P} = \{P^{t_i}\}$ such that $\|\Delta P_i A \; \Delta P_i\| < \varepsilon$. Let $\Delta t = \min (t_i - t_{i-1})$. Since any refinement of $\mathcal{P}$ is easily seen to have the property $\|\Delta P_j A \; \Delta P_j\| < \varepsilon$, this Δt satisfies the requirements and A is strongly causal.

COROLLARY 3.5. *If A is strongly causal on $(\mathcal{H}, \mathcal{C})$, then P^tAP^t is strictly causal on $(P^t\mathcal{H}, P^t\mathcal{C})$ for any $t \in \Gamma$.*

Proof. Consider the operator P^tAP^t on $(P^t\mathcal{H}, P^t\mathcal{C})$. By the results of [4], P^tAP^t is strictly causal on $(P^t\mathcal{H}, P^t\mathcal{C})$ if given $\varepsilon > 0$, there exists a partition $\hat{\mathcal{P}} = \{0 < P^{t_1} < \cdots < P^{t_n} = P^t\}$ such that

$$\max_{1 \leqq i \leqq n} \|\Delta P_i P^tAP^t \; \Delta P_i\| < \varepsilon.$$

By strong causality, there exists a generalized partition $\mathcal{P} = \{P^{t_i}\}$ such that

$$\|\Delta P_i A \; \Delta P_i\| < \varepsilon$$

for all i. Let $\hat{\mathscr{P}} = P^t\mathscr{P} = \{P^tP^{t_i}\}$. Since

$$P^tP^{t_i} = P^{t_i}p^t = \begin{cases} P^{t_i}, & t_i \leqq t, \\ P^t, & t_i > t, \end{cases}$$

we have that

$$\Delta P_iP^t = \begin{cases} \Delta P_i, & t_i \leqq t, \\ P^t - P^{t_{i-1}}, & t_{i-1} \leqq t \leqq t_i, \\ 0, & t_{i-1} \geqq t. \end{cases}$$

If $t_{n-1} \leqq t \leqq t_n$, then it follows immediately that $\hat{\mathscr{P}}$ is a partition on $(P^t\mathscr{H}, P^t\mathscr{E})$ and that

$$\max_{1 \leqq i \leqq n} \|\Delta P_iP^tAP^t\ \Delta P_i\| < \varepsilon.$$

This completes the proof.

As is well-known [2], the bounded strongly causal operators form a uniformly closed two-sided ideal and the formulation of strong causality given in Theorem 3.4 together with Corollary 3.5 show that they properly contain the ideal of strictly causal operators.

We now turn to the invertibility problem. If A is strongly causal on $(\mathscr{H}, \mathscr{E})$, then P^tAP^t is strictly causal. Thus $(I - P^tAP^t)^{-1}$ exists and is causal.

THEOREM 3.6. *Suppose A is strongly causal on $(\mathscr{H}, \mathscr{E})$. Then $(I - A)^{-1}$ exists and is causal if and only if* $\sup_{t<\infty} \|(I - P^tAP^t)^{-1}\| < \infty$.

Proof. If $(I - A)^{-1}$ exists, then since $(I - P^tAP^t)(I - P^tAP^t)^{-1} = I$ and $\lim_{t\to\infty} (I - P^tAP^t) = (I - A)$ it follows that $(I - P^tAP^t)^{-1} \to (I - A)^{-1}$ strongly. Then by the uniform boundedness principle,

$$\|(I - P^tAP^t)^{-1}\| < \infty \quad \text{for all } t.$$

Now suppose the condition

$$\sup_{t<\infty} \|(P^t - P^tAP^t)^{-1}\| \leqq M < \infty$$

holds. We show that $(I - A)$ is invertible. Then the argument used above shows that $(I - P^tAP^t)^{-1} \to (I - A)^{-1}$ strongly. Since, by Corollary 3.5, P^tAP^t is strictly causal for all t, and thus $(I - P^tAP^t)^{-1}$ is causal, it will follow from the strong closure of the causal operators [5] that $(I - A)^{-1}$ is causal.

To show $I - A$ is invertible, it suffices to show that $\|(I - A)x\|$ and $\|(I - A)^*x\|$ are bounded below for $x \in \mathscr{H}$. We do this for the first term noting that a similar argument holds for the second.

Then

$$\begin{aligned} \|x\| &= \|(I - P^tAP^t)^{-1}(I - P^tAP^t)x\| \\ &\leqq \|(I - P^tAP^t)^{-1}\|\ \|(I - P^tAP^t)x\| \\ &\leqq M\|(I - P^tAP^t)x\|. \end{aligned}$$

Thus $\|(I - P^tAP^t)x\| \geqq (1/M)\|x\|$ for all $t \in \Gamma$.

Since as $t \to \infty$, $P^t \to I$ strongly, we obtain

$$\|(I - A)x\| \geqq M\|x\|.$$

This completes the proof.

Remark 3.7. The largeness of the class of strongly causal operators becomes clear by taking $\mathcal{H} = l^2(-\infty, \infty)$ with the usual orthonormal basis. The strongly causal operators are those whose matrix representations are strictly lower triangular.

4. The memoryless part of an operator. Here we present a mapping that gives for each operator in $\mathcal{B}(\mathcal{H})$ a memoryless operator which, in a natural way, will be the memoryless part of the operator. In the case where an additive decomposition exists, the two notions of memorylessness coincide. In fact, much more is true. The notions of strict and strong causality were delay conditions. This means that these operators have, in a certain sense, no memoryless part. This construction will show in what sense this is meant. For strongly causal operators, when operated on by this mapping, give zero memoryless part. Thus strongly causal operators will be in the null space of this mapping. The problem of characterizing the null space of this mapping seems to be important and open.

This construction is not new and is well-known in the study of operator algebras [9], [14]. However, since we feel it is quite important for the study of feedback systems, we go into great detail. Since the first version of this paper was written, we have made use of this construction in studying some problems in stochastic optimization [15].

It is a basic fact that $\mathcal{B}(\mathcal{H})$ is a Banach space and is the dual space of the Banach space $\mathcal{C}_1$ of trace class operators [7]. Thus every operator $A \in \mathcal{B}(\mathcal{H})$ can be looked upon as a bounded linear functional on $\mathcal{C}_1$. This is defined by

$$T \rightarrow \operatorname{Tr}(AT), \qquad T \in \mathcal{C}_1.$$

Details can be found in [7].

Now consider $\mathcal{E} = \{P^t : t \in \Gamma\}$. Since these projections commute with each other and are selfadjoint, they generate a strongly closed, Abelian, selfadjoint algebra which we denote by $\mathcal{R}$. Let U denote the set of unitary operators in $\mathcal{R}$ (i.e., $U \in \mathcal{U} \Leftrightarrow U \in \mathcal{R}$ and $UU^* = U^*U = I$). Then $\mathcal{U}$ is an Abelian group in $\mathcal{R}$, and generates $\mathcal{R}$ in the sense that every element in $\mathcal{R}$ can be written as a finite linear combination of elements of $\mathcal{U}$ [3].

DEFINITION 4.1. Let G be a group and denote by $\mathcal{B}(G)$ the algebra of bounded complex functions on G. A linear functional M on $\mathcal{B}(G)$ is called an *invariant mean* on G if:

(i) $f \in \mathcal{B}(G)$ is real valued, then
$\inf\{f(x) : x \in G\} \leqq M(f) \leqq \sup\{f(x) : x \in G\}$

(ii) for each $g \in G$, if $f_g(x) = f(gx)$, then $M(f_g) = M(f)$.

Note that (i) implies $M(f) = f$ for f constant. Not all groups have invariant means. However, they do exist for Abelian groups although they are not necessarily unique [8]. Thus $\mathcal{U}$ has an invariant mean which we will denote by M.

Now fix $A \in \mathcal{B}(\mathcal{H})$ and $T \in \mathcal{C}_1$, and consider the bounded complex function defined on $\mathcal{U}$ by

$$U \rightarrow \operatorname{Tr}(U^*AUT),$$

Our mapping will be defined by means of this function. For $A \in \mathcal{B}(\mathcal{H})$ define $\psi(A)$ as a bounded linear functional on $\mathcal{C}_1$ by

$$\langle T, \psi(A)\rangle = M[\operatorname{Tr}(U^*AUT)].$$

Since $\mathcal{B}(\mathcal{H})$ is the dual space of $\mathcal{C}_1$, the mapping ψ for each $A \in \mathcal{B}(\mathcal{H})$, defines an element of $\mathcal{B}(\mathcal{H})$. The properties of ψ have been studied in detail in [14], [15]. Here we list those useful for our discussion.

THEOREM 4.2. *Let ψ be the mapping on $\mathscr{B}(\mathscr{H})$ defined above. Then*

(i) *ψ is linear and $\|\psi\| = 1$.*

(ii) *If A is memoryless and $B \in \mathscr{B}(\mathscr{H})$, then $\psi(AB) = A\psi(B)$ and $\psi(BA) = \psi(B)A$.*

(iii) *For all $A \in \mathscr{B}(\mathscr{H})$, $\psi(A)$ is memoryless and $\psi[\psi(A)] = \psi(A)$.*

We note that property (iii) implies that ψ is a projection of $\mathscr{B}(\mathscr{H})$ onto the algebra of memoryless operators. That $\psi(A)$ can be looked upon as an extension of the memoryless part of A encountered in the additive decomposition can be seen in the next result. This is seen for different extensions of strict causality in [15].

THEOREM 4.3. *If A is strongly causal, then $\psi(A) = 0$.*

Proof. Fix $\varepsilon > 0$. Then by Theorem 3.5 there exists a generalized partition $\mathscr{P} = \{P^{t_i}\}$ such that $\|\Delta P_i A \, \Delta P_i\| < \varepsilon$ for all i. Since ψ has norm 1, it follows that $\|\psi(\Delta P_i A \, \Delta P_i)\| < \varepsilon$. But $\psi(\Delta P_i A \, \Delta P_i) = \Delta P_i \psi(A) \, \Delta P_i$ by Theorem 4.2 (ii) and since $\sum_i \Delta P_i = I$, we have

$$\psi(A) = \sum_i \psi(A) \, \Delta P_i = \sum_i \Delta P_i \psi(A) \, \Delta P_i$$

by memorylessness of $\psi(A)$. Since $\Delta P_i \perp \Delta P_j$ for $i \neq j$, it follows that $\|\psi(A)\| < \varepsilon$. Since ε was arbitrary, $\psi(A) = 0$.

Examples 4.4. (1) Let $\mathscr{H} = l^2(-\infty, \infty)$ with the standard resolution of the identity. Then the strictly lower triangular matrices have the property that $\psi(A) = 0$.

(2) Let K be a convolution operator on $L^2(0, \infty)$ with L^1 kernel. Then K is strongly causal and $\psi(K) = 0$.

Now suppose A is causal. Then

$$A = \psi(A) + (A - \psi(A))$$

is a decomposition of A into a memoryless part $\psi(A)$ and an operator whose memoryless part is zero. This "additive decomposition" is equivalent to the additive decomposition described in Theorem 2.3 when that exists. This raises the question of the uniqueness of the above decomposition in the following sense: if $A = B_1 + B_2$, where B_2 is memoryless and $\psi(B_1) = 0$, does this imply that $B_2 = \psi(A)$ and $B_1 = A - \psi(A)$? A partial answer is given in the next theorem.

THEOREM 4.5. *Suppose A is causal and $A - \psi(A)$ is strongly causal. If $A = B_1 + B_2$ with B_1 strongly causal and B_2 memoryless, then $B_1 = A - \psi(A)$ and $B_2 = \psi(A)$.*

Proof. Since $A = [A - \psi(A)] + \psi(A) = B_1 + B_2$, it follows that $[A - \psi(A)] - B_1 = B_2 - \psi(A)$. Since the strongly causal operators form an ideal, $A - \psi(A) - B_1$ is strongly causal. Thus, so is $\psi(A) - B_2$. It then follows from Theorem 4.3 that $\psi(\psi(A) - B_2) = 0$. But

$$\begin{aligned}\psi(\psi(A) - B_2) &= \psi(\psi(A)) - \psi(B_2) \\ &= \psi(A) - B_2.\end{aligned}$$

Thus, $\psi(A) = B_2$ and $A - \psi(A) = B_1$. This completes the proof.

We now raise some invertibility questions related to the fact that $\psi(A) = 0$ is a strong causality condition for a causal operator A.

(1) Suppose A is causal and $\psi(A) = 0$. When does $I - A$ have a causal inverse? The answer must clearly include Theorem 3.5.

(2) Suppose A is causal and $I - (A - \psi(A))$ is invertible with causal inverse. The argument used in Theorem 2.3 shows that $I - A$ is invertible with causal inverse if and only if $1 \notin \sigma(\psi(A))$. Can we determine $\sigma(\psi(A))$?

(3) We have noted that the null space of ψ contains the strongly causal operators. We would like to characterize this null space. In particular, are there operators which

are not strongly causal for which $\psi(A)=0$? This question becomes particularly important in light of the results of [15].

Acknowledgment. The author would like to thank Professor R. DeSantis, A. Manitius and G. Zames for making his visit at Université de Montreal and McGill University possible.

REFERENCES

[1] R. DeSantis, *Causality, strict causality, invertibility for systems in Hilbert resolution spaces*, this Journal, 12 (1974), pp. 536–553.

[2] R. DeSantis, R. Saeks and L. Tung, *Basic optimal estimation and control problems in Hilbert space*, Math. Systems Theory, to appear.

[3] J. Dixmier, *Les Algebres d'Operateurs dans l'Espace Hilbertien*, Gauthier-Villars, Paris, 1957.

[4] J. A. Erdos and W. E. Longstaff, *The convergence of triangular integrals of operators on Hilbert space*, Indiana Univ. Math. J., 22 (1973), pp. 929–938.

[5] A. Feintuch, *Strict and strong strict causality for operators*, SIAM J. Math. Anal., to appear.

[6] ———, *Causal invertibility in resolution spaces*, preprint.

[7] I. C. Gohberg and M. G. Krein, *Theory and applications of Volterra operators in Hilbert space*, Transl. Math. Mono. 24, American Mathematical Society, Providence, RI, 1970.

[8] F. Greenleaf, *Invariant Means on Topological Groups*, Van Nostrand, New York, 1969.

[9] B. Johnson and S. Parrott, *Operators commuting with a Von Neumann algebra modulo the set of compact operators*, J. Functional Analysis, 11 (1972), pp. 39–61.

[10] R. Saeks, *Resolution Space, Operators and Systems*, Lecture Notes in Economics and Math. Systems 82, Springer Verlag, New York, 1973.

[11] M. L. Steinberger, A. Schumitzky and L. Silverman, *Optimal causal feedback control of linear infinite dimensional systems*, this Journal, to appear.

[12] J. C. Willems, *Stability, instability, invertibility and causality*, this Journal, 7 (1969), pp. 645–671.

[13] G. Zames, *Realizability conditions for nonlinear feedback systems*, IEEE Trans. Circuit Theory, CT-11 (1964), pp. 186–194.

[14] D. Larson, *On the structures of certain reflexive operators*, J. Functional Analysis, to appear.

[15] A. Feintuch, R. Saeks and R. Neil, *A new performance measure for stochastic optimization in Hilbert space*, preprint.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 3, May 1980

0363-0129/80/1803-0008$01.00/0

ERRATUM: ON DECOMPOSITION OF GENERATORS*

JERZY ZABCZYK†

Let A and B be infinitesimal generators of linear C_0-semigroups on Banach spaces X and Y, respectively, and F a bounded linear operator from Y into X. The computations on page 525 of my paper imply that the operators

$$A_0 = \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix} \begin{pmatrix} I & -F \\ 0 & I \end{pmatrix} \quad \text{and} \quad A_1 = \begin{pmatrix} I & F \\ 0 & I \end{pmatrix} \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}$$

generate C_0-semigroups on the product space $X \times Y$, provided that the condition—FB has a continuous extension to the whole Y—is satisfied. Therefore, one should add this condition in the formulation of part b of Theorem 1. A different sufficient condition for operators A_0 and A_1 to be C_0-generators is that operator F transforms Y into $D(A)$, and this fact follows from part a of Theorem 1. Nevertheless, all applications of Theorem 1 considered in the paper are not affected by the correction; as in all of them, the operator B was bounded. That some conditions are in fact necessary for A_0 and A_1 to generate C_0-semigroups can be seen from an example contained in [1].

The author would like to thank Professor Goong Chen and Professor Ronald Grimmer for drawing his attention to the described incorrectness.

Another reference, recently discovered by the author, is relevant to the content of the paper; namely, an analogous proof to that given in the paper of the property, that A-bounded finite dimensional operator on a reflexive Banach space has A-bound zero (see Remark 3 of my paper), is contained in [2, pp. 195–196] together with the property itself.

REFERENCES

[1] G. CHEN AND R. GRIMMER, *Semigroups and integral equations*, J. Integral Equations, submitted.
[2] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.

* This Journal, 6 (1978), pp. 523–534. Received by the editors September 11, 1979.

† Institute of Mathematics, Polish Academy of Sciences, 00-950 Warsaw, Poland.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 4, July 1980

0363-0129/80/1804-0001 $01.00/0

NULL CONTROLLABILITY OF LINEAR SYSTEMS WITH CONSTRAINED CONTROLS*

W. E. SCHMITENDORF† AND B. R. BARMISH‡

Abstract. The paper considers the problem of steering the state of a linear time-varying system to the origin when the control is subject to magnitude constraints. Necessary and sufficient conditions are given for global constrained controllability as well as a necessary and sufficient condition for the existence of a control (satisfying the constraints) which steers the system to the origin from a specified initial epoch (x_0, t_0). The global result does not require zero to be an interior point of the control set Ω, and the theorem for constrained controllability at (x_0, t_0) only requires that Ω be compact, not that it contain zero. The results are compared to those available in the literature. Furthermore, numerical aspects of the problem are discussed as is a technique for determining a steering control.

1. Introduction and formulation. Consider the problem of steering the state of a linear system

$$\text{(S)} \qquad \dot{x}(t)=A(t)x(t)+B(t)u(t), \qquad t\in[t_0,\infty)$$

to the origin from a specified initial condition

$$x(t_0)=x_0$$

by choice of control function $u(\cdot)$. Here $x(t)\in R^n$, $u(t)\in R^m$, and $A(\cdot)$ and $B(\cdot)$ are continuous matrices[1] of appropriate dimension. Unlike the usual controllability problem, where the control values at each instant of time are unconstrained, we insist here that the control values at each instant of time belong to a prespecified set Ω in R^m.

Let $\mathcal{M}(\Omega)$ denote the set of functions from R into Ω that are measurable on $[t_0,\infty)$. Then any control $u(\cdot)\in\mathcal{M}(\Omega)$ is termed admissible. We now define three notions of constrained controllability or, more precisely, Ω-null controllability.

DEFINITION 1.1. The linear system (S) is *Ω-null controllable at* (x_0, t_0) if, given the initial condition $x(t_0)=x_0$, there exists a control $u(\cdot)\in\mathcal{M}(\Omega)$ such that the solution $x(\cdot)$ of (S) satisfies $x(t)=0$ for some $t\in[t_0,\infty)$.

DEFINITION 1.2. The linear system (S) is *globally Ω-null controllable at* t_0 if (S) is Ω-null controllable at (x_0, t_0) for all $x_0\in R^n$.

Our major result will pertain to the global type of controllability. To compare our results to those of previous researchers, we also need a local controllability concept.

DEFINITION 1.3. The linear system (S) is *locally Ω-null controllable at* t_0 if there exists an open set $V\subset R^n$, containing the origin, such that (S) is null controllable at (x_0, t_0) for all $x_0\in V$.

The majority of constrained controllability results are for autonomous systems, i.e., systems where A and B are constant. When $\Omega=R^m$, Kalman [1] showed that a necessary and sufficient condition for global R^m-null controllability is $\operatorname{rank}(Q)=n$, where $Q\triangleq[B, AB, \cdots, A^{n-1}B]$. Lee and Markus [2] considered constraint sets $\Omega\subset R^m$ which contain $u=0$, and showed that $\operatorname{rank}(Q)=n$ is a necessary and sufficient condition for (S) to be locally Ω-null controllable. Furthermore, if each eigenvalue λ of A satisfies $\operatorname{Re}(\lambda)<0$, then (S) is globally Ω-null controllable. This result is typical of the results available when Ω contains the origin.

* Received by the editors September 1, 1978, and in final revised form August 28, 1979.

† Technological Institute, Northwestern University, Evanston, Illinois 60201. The research of this author was supported by Air Force Office for Scientific Research under Grant 76-2923.

‡ Department of Electrical Engineering, University of Rochester, Rochester, New York 14627.

[1] This requirement can be weakened to local integrability.

Saperstone and Yorke [3] were the first to eliminate the assumption that zero is an interior point of Ω when they considered problems with $m=1$ and $\Omega=[0, 1]$. Their result states that, for these problems, (S) is locally Ω-null controllable if and only if $\operatorname{rank}(Q)=n$ and A has no real eigenvalues. They also extend this result to $m>1$ and consider the m-fold product set $\Omega=\prod_1^m[0, 1]$. Problems with more general constraint sets were studied by Brammer [4] who showed that if there exists a $u\in\Omega$ satisfying $Bu=0$, and the convex hull of Ω has a nonempty interior, then necessary and sufficient conditions for local Ω-null controllability are $\operatorname{rank}(Q)=n$ and the nonexistence of a real eigenvector v of A' satisfying $v'Bu\leqq 0$ for all $u\in\Omega$. In addition, if no eigenvalue of A has a positive real part, then the theorem becomes one for global Ω-null controllability. A similar result for global controllability when $\Omega=[0, 1]$ was obtained by Saperstone [5]. Friedman [6] considers a linear pursuit evasion problem, where the target is a closed convex set, and gives a sufficient condition for the existence of a pursuer control, based on the evader's control, which drives the system from a specified initial condition to the target.

For nonautonomous systems, the most familiar controllability result is that of Kalman [1] when $\Omega=R^m$. He showed that (S) is R^m-null controllable if and only if $W(t_0, t_1)$ is positive definite for some $t_1\in[t_0, \infty)$, where

$$W(t_0, t_1)\triangleq\int_{t_0}^{t_1}\phi(t_1, \tau)B(\tau)B'(\tau)\phi'(t_1, \tau)\, d\tau,$$

and $\phi(t, \tau)$ is the state transition matrix for (S). When the control is constrained, the major global results are those by Conti [7] and Pandolfi [8]. In [7], Conti describes a "divergent integral condition" which is necessary and sufficient for global Ω-null controllability when Ω is the closed unit ball. In order to make Conti's result more compatible with existing theory for time-invariant systems, Pandolfi in [8] defines the notion of pth characteristic exponent for time-varying systems. For the special case when the system is time-invariant, the characteristic exponent turns out to be the real part of some eigenvalue of A. Subsequently, controllability criteria are provided in terms of this exponent.

The Ω-null controllability problem is also studied in papers by Dauer [9], [10], Chukwu and Gronski [11] and Chukwu and Silliman [12]. In order to answer the question of Ω-controllability, one must test a certain *growth condition* which involves searching a function space. In contrast, the results given here are finite-dimensional in nature.

In [13], Grantham and Vincent consider the problem of steering a nonlinear system to a target. They present a technique for determining the boundary between the set of states which can be steered to the target and those which cannot. More recently, Murthy and Evans [14] obtained results comparable to [3]–[5] for discrete linear systems and Pachter and Jacobson [15] developed sufficient conditions for controllability for case where $A(\cdot)$ and $B(\cdot)$ are time-invariant and Ω is a closed convex cone containing the origin. A readable account of the state of the art is contained in the book by Jacobson [16, Chap. 5].

In contrast to much of the work of previous authors, this paper concentrates on the case where $A(\cdot)$ and $B(\cdot)$ are time-varying. Our results for global Ω-null controllability are for constraint sets Ω that are compact and contain zero (but not necessarily as an interior point). One of our main results on global Ω-null controllability is an extension of a theorem of Conti [7] and it degenerates to Conti's theorem when Ω is a unit ball.

Our results for Ω-null controllability at (x_0, t_0) have even wider applicability since they do not require $0 \in \Omega$. Neither do they require the existence of a $u \in \Omega$ such that $Bu = 0$ as in [3]–[5], [7]–[12]. Thus we can analyze controllability of a system with, for example, $m = 1$ and $\Omega = [1, 2]$, whereas, many of the presently available theorems do not apply. Furthermore, as will be illustrated by examples, there are autonomous systems (S) which are neither globally Ω-null controllable nor locally Ω-null controllable but nevertheless are Ω-null controllable at some (x_0, t_0). Our theorem can be used to decompose the state space into two sets. Initial states in one set can be steered to the origin while those in the other cannot be driven to the origin by an admissible control.

2. Main results. In order to describe our necessary and sufficient conditions for global Ω-null controllability, we make use of the *support function* $H_\Omega : R^m \to R \cup \{+\infty\}$ on Ω which for any $\alpha \in R^m$ is given by

$$H_\Omega(\alpha) \triangleq \sup \{\omega'\alpha : \omega \in \Omega\}.$$

Using this notation, we have the following theorem, which is proven in Appendix A.

THEOREM 2.1. *Suppose Ω is a compact set which contains zero.*[2] *Then,* (S) *is globally Ω-null controllable at t_0 if and only if*

$$\int_{t_0}^{\infty} H_\Omega(B'(\tau)z(\tau))\, d\tau = +\infty \tag{2.1}$$

for all nonzero solutions $z(\cdot)$ of the adjoint system

$$\dot{z}(t) = -A'(t)z(t), \qquad t \in [t_0, \infty), \tag{S'}$$

or equivalently, if and only if

$$\int_{t_0}^{\infty} \sup \{\omega' B'(\tau)\phi'(t_0, \tau)\lambda : \omega \in \Omega\}\, d\tau = +\infty$$

for all $\lambda \in R^n$, $\lambda \neq 0$.

We note that $H_\Omega(B'(\tau)z(\tau))$ can be viewed as the composition of a nonnegative Baire function with a measurable function. Hence, the integral in (2.1) is well-defined along all trajectories $z(\cdot)$ of (S).

In the following corollary, we examine the special case of Theorem 2.1 which arises under the strengthened hypothesis "zero is an interior point of Ω." As we might anticipate, for this special case, the structure of the set Ω will not matter other than the requirement that it contains zero in its interior.

COROLLARY 2.2 *Suppose there exists a compact set Ω such that*

(i) *zero is an interior point of Ω;*

(ii) (S) *is globally Ω-null controllable.*

Then (S) *is also globally Ω'-null controllable for any other set Ω'* (*not necessarily compact*) *which contains zero in its interior.*

See Appendix A for proof.

Our proof of Theorem 2.1 will make use of a more fundamental result (also proven in Appendix A) giving conditions for Ω-null controllability at a fixed initial epoch (x_0, t_0). To meet this end, we define the scalar function $J : R^n \times R \times R^n \to R$ by

$$J(x_0, T, \lambda) \triangleq x_0'\phi'(T, t_0)\lambda + \int_{t_0}^{T} H_\Omega(B'(\tau)\phi'(T, \tau)\lambda)\, d\tau. \tag{2.2}$$

[2] The theorem is also valid if the requirement "$0 \in \Omega$" is replaced with "there exists a $u \in \Omega$ such that $Bu = 0$". This type of assumption is used by Brammer [4].

We note that $J(x_0, T, \lambda)$ can be viewed as the support function on the so-called attainable set. This fact is used implicitly in the proof of the next theorem.

THEOREM 2.3. *Let Ω be a compact set. Pick any subset Λ of R^n which contains* 0 *as an interior point. Then* (S) *is Ω-null controllable at (x_0, t_0) if and only if*

$$\min\{J(x_0, T, \lambda): \lambda \in \Lambda\} = 0 \tag{2.3}$$

for some $T \in [t_0, \infty)$, or equivalently, if and only if

$$J(x_0, T, \lambda) \geqq 0 \quad \textit{for all } \lambda \in \Lambda \tag{2.4}$$

for some $T \in [t_0, \infty)$.

Comment. If Ω is also convex and A and B are constant, the sufficiency portion of this theorem is just a special case of Theorem 7.2.1 of [6]. Naturally, the smallest time T for which (2.3) holds will be the minimum *arrival time* at the origin.

Theorem 2.3 can also be stated in terms of the adjoint system (S′), i.e., if we take $\Lambda = R^n$ and notice that $z(t) = \phi'(t_0, t)z(t_0)$ is the response of the adjoint system (S′), then the following theorem is easily proven. (The proof is established by making the change of variables $z(t) \triangleq \phi'(T, t)\lambda$).

THEOREM 2.3′. *Let Ω satisfy the hypothesis of Theorem* 2.3. *Then* (S) *is Ω-null controllable at (x_0, t_0) if and only if there exists some $T \in [t_0, \infty)$ such that*

$$x_0'z(t_0) + \int_{t_0}^{T} H_\Omega(B'(\tau)z(\tau))\,d\tau \geqq 0 \tag{2.5}$$

for all solutions $z(\cdot)$ of (S′).

This theorem demonstrates that the question of Ω-null controllability at (x_0, t_0) can be answered by solving a finite dimensional optimization problem. Moreover, the question of global Ω-null controllability can also be answered via a finite dimensional optimization problem.

COROLLARY 2.4. *Let Ω and Λ be as in Theorem* 2.3. *Then* (S) *is globally Ω-null controllable at t_0 if and only if for every $x_0 \in R^n$ there is a time $T_{x_0} \in [t_0, \infty)$ such that*

$$\min\{J(x_0, T_{x_0}, \lambda): \lambda \in \Lambda\} = 0.$$

The proof of this corollary follows from Theorem 2.3 in conjunction with the definition of global Ω-null controllability.

There is one point worth noting. In using Theorem 2.1 to check for Ω-null controllability at t_0, Ω must be compact and contain 0. If Corollary 2.4 is used, only the compactness assumption must be satisfied.

Next, we present some examples to illustrate how our theorems can be applied and to compare our results to those of [3]–[5].

Example 1. Let $x(t)$ and $u(t)$ be scalars and suppose (S) is described by

$$\dot{x}(t) = x(t) + u(t), \qquad t \in [0, \infty).$$

This system is R^1-null controllable if $\Omega = R^1$. But suppose $\Omega = [0, 1]$. Then the system is *not* globally Ω-null controllable at $t_0 = 0$. This follows from Theorem 2.1 since, for $z_0 < 0$, $H_\Omega(B'(\tau)z(\tau)) = 0$, and thus $\int_0^\infty H_\Omega(B'(\tau)z(\tau))\,d\tau < +\infty$. Also, using [3] or [4], it can be shown that the system is *not* locally Ω-null controllable. Nevertheless, there do exist initial states x_0 from which it is possible to steer the system to the origin. Such states can be determined via Theorem 2.3.

For the above,

$$J(x_0, T, \lambda) = x_0 e^T \lambda + \int_0^T \sup\{\omega e^{T-\tau}\lambda : \omega \in [0, 1]\}\,d\tau.$$

When $\Lambda=[-1, 1]$, this becomes

$$J(x_0, T, \lambda)=\begin{cases} x_0 e^T\lambda, & \lambda \leqq 0, \\ x_0 e^T\lambda+\lambda(e^T-1), & \lambda>0, \end{cases}$$

and thus

$$\min\{J(x_0, T, \lambda): \lambda \in[-1, 1]\}=0$$

if and only if $x_0 \leqq 0$ and $x_0 \geqq e^{-T}-1$ for some $T \in[0, \infty)$, or equivalently, if and only if $-1<x_0 \leqq 0$. We conclude that even though (S) is not locally Ω-null controllable, it is Ω-null controllable at $(x_0, 0)$ whenever $-1<x_0 \leqq 0$.

If $\Omega=[1, 2]$, neither [3]–[5] nor Theorem 2.1 apply. However, we can use Theorem 2.3. Since

$$H_\Omega(B'(\tau)\phi'(T, \tau)\lambda)=\begin{cases} 2\lambda\, e^{(T-\tau)}, & \lambda>0, \\ \lambda\, e^{(T-\tau)}, & \lambda \leqq 0, \end{cases}$$

$J(x_0, T, \lambda)$ becomes

$$J(x_0, T, \lambda)=\begin{cases} x_0 e^T\lambda+2\lambda(e^T-1), & \lambda>0, \\ x_0 e^T\lambda+\lambda\,(e^T-1), & \lambda \leqq 0, \end{cases}$$

and

$$\min\{J(x_0, T, \lambda): \lambda \in[-1, 1]\}=0$$

if and only if $2(e^{-T}-1) \leqq x_0 \leqq e^{-T}-1$. Thus (S), with $\Omega=[1, 2]$, is Ω-null controllable at $(x_0, 0)$ whenever $-2<x_0 \leqq 0$.

As a final variation of this problem, suppose $\Omega=[-a, a]$. Then [4] or Theorem 2.1 shows that (S) is not globally Ω-null controllable. Using [4], it can be demonstrated that (S) is locally Ω-null controllable, while Theorem 2.3 not only tells us that (S) is locally Ω-null controllable but also that the states x_0 which can be steered to the origin are those satisfying $-a<x_0<a$.

Example 2. Our second example illustrates the application of Theorem 2.1 for a nonautonomous system. We consider the time-varying two-dimensional system (S) described by

$$\dot{x}_1(t)=u(t) \sin t,$$

$$\dot{x}_2(t)=-\frac{1}{(t+1)^2} x_1(t)+u(t) t \sin t, \qquad t \in[0, \infty).$$

The control constraint set is taken to be $\Omega=[0, 1]$. By a straightforward computation, the state transition matrix for the adjoint system (S′) is found to be

$$\phi_*(t, t_0)=\begin{bmatrix} 1 & \dfrac{t-t_0}{(t+1)(t_0+1)} \\ 0 & 1 \end{bmatrix}.$$

Hence, in accordance with Theorem 2.1, (S) is globally Ω-null controllable at $t_0=0$ if and only if

$$\int_0^\infty \sup_{\omega \in[0, 1]} \omega[\sin \tau \quad \tau \sin \tau]\begin{bmatrix} 1 & \dfrac{\tau}{\tau+1} \\ 0 & 1 \end{bmatrix}\begin{bmatrix} z_{01} \\ z_{02} \end{bmatrix} d\tau=+\infty$$

for *all* nonzero initial conditions $z_0 \triangleq [z_{01}\ z_{02}]'$. Evaluating above, this reduces to the requirement that

$$(2.6)\qquad \int_0^\infty I(\tau)\,d\tau \triangleq \int_0^\infty \max\left\{0, z_{01}\sin\tau + z_{02}\tau\sin\tau\left(1+\frac{1}{\tau+1}\right)\right\} d\tau = +\infty$$

for all $z_0 \neq 0$. We shall show that this condition is indeed satisfied.

Case 1. $z_{01} \neq 0$, $z_{02} = 0$. For this case, we have

$$\int_0^\infty I(\tau)\,d\tau = \int_0^\infty \max\{0, z_{01}\sin\tau\}\,d\tau$$
$$= \int_{\mathcal{T}_1} z_{01}\sin\tau\,d\tau,$$

where $\mathcal{T}_1 \triangleq \{\tau \geqq 0: z_{01}\sin\tau > 0\}$. Because the range set $\mathcal{T}_1$ of integration is the union of infinitely many intervals of length π, it follows that

$$\int_0^\infty I(\tau)\,d\tau = +\infty.$$

Case 2. $z_{01} =$ anything, $z_{02} \neq 0$. Let $T^* \triangleq (|z_{01}|+1)/|z_{02}|$. Then to verify (2.6), it suffices to show that

$$\int_{\mathcal{T}_2} I(\tau)\,d\tau = +\infty,$$

where $\mathcal{T}_2 = \{\tau \geqq T^*: z_{02}\sin\tau > 0\}$. (Recall that the integrand is nonnegative.) Now, for $\tau \in \mathcal{T}_2$, we notice that the integrand $I(\tau)$ can be bounded from below as follows:

$$\begin{aligned} z_{01}\sin\tau + z_{02}\tau\sin\tau\left(1+\frac{1}{\tau+1}\right) &\geqq |z_{02}|\,|\sin\tau|\,\tau\left(1+\frac{1}{\tau+1}\right) - |z_{01}|\,|\sin\tau| \\ &\geqq (|z_{02}|\tau - |z_{01}|)|\sin\tau| \\ &\geqq (|z_{02}|T^* - |z_{01}|)|\sin\tau| \\ &= |\sin\tau|. \end{aligned}$$

Hence,

$$\int_{\mathcal{T}_2} I(\tau)\,d\tau \geqq \int_{\mathcal{T}_2} |\sin\tau|\,d\tau = +\infty$$

because the range of integration is once again the union of infinitely many intervals of length π.

We conclude that (S) is globally Ω-null controllable.

3. Relationship with other controllability results. In this section, we compare our controllability results with those of Conti [7] and Brammer [4]. We also consider, as a limiting case of our theory, the usual controllability problem obtained when magnitude constraints are not present.

Result of Conti. An important special case of Theorem 2.1 occurs when Ω is a closed unit ball in R^m, i.e.,

$$\Omega = \{\omega \in R^m: \|\omega\| \leqq 1\},$$

where $\|\cdot\|$ is a prespecified norm on R^m. For this situation we have

$$H_\Omega(B'(\tau)z(\tau)) = \sup\{\omega' B'(\tau)z(\tau): \|\omega\| \leqq 1\} = \|B'(\tau)z(\tau)\|_*,$$

where $\|\cdot\|_*$ is the norm on R^m which is dual to $\|\cdot\|$. (For example $\|\cdot\|_*$ is the l^1 norm when $\|\cdot\|$ is the l^∞ norm; $\|\cdot\|$ and $\|\cdot\|_*$ coincide when $\|\cdot\|$ is the usual l^2 (Euclidean) norm.)

By Theorem 2.1, we conclude that (S) is globally Ω-null controllable at t_0 if and only if

$$\int_{t_0}^{\infty} \|B'(\tau)z(\tau)\|_* \, d\tau = +\infty \tag{3.1}$$

for all nonzero solutions $z(\cdot)$ of (S'). This result is established independently in Conti [7] and also discussed in Pandolfi [8]. This result, in conjunction with Corollary 2.2 leads immediately to the following proposition.

PROPOSITION 3.1. *Let Ω be any set containing zero in its interior. Then* (3.1) *is a necessary and sufficient condition for global Ω-null controllability.*

Thus, Conti's condition is a necessary and sufficient condition for global Ω-null controllability for any set Ω containing zero in its interior, not just when Ω is the closed unit ball.

Result of Brammer. Consider the case when $A(t)\equiv A$ and $B(t)\equiv B$ are time-invariant. For these autonomous problems, the following necessary conditions can be obtained directly from Theorem 2.1. Recall that $Q=[B, AB, \cdots, A^{n-1}B]$.

THEOREM 3.2. *Assume $A(t)\equiv A$ and $B(t)\equiv B$ are time-invariant and that Ω is a compact set which contains the origin. If* (S) *is globally Ω-null controllable then*

(i) rank $(Q)=n$;

(ii) *there is no real eigenvector v of A' satisfying $v'B\omega \leqq 0$ for all $\omega \in \Omega$;*

(iii) *no eigenvalue of A has a positive real part.*

The proof of this result is in Appendix B.

In [4], Brammer has obtained the same result using a different method of proof. There, he also shows that the above three conditions are also sufficient for global Ω-null controllability in the time invariant case if it is also assumed that the convex hull of Ω has a nonempty interior. Alternative proofs of the sufficiency results have been given by Heymann and Stern [25] and Hajek. The latter proof is in [5].

We note that the system of Example 1 of § 2 does not satisfy these three conditions. Nevertheless, it is Ω-null controllable at $(x_0, 0)$ for some initial states x_0.

The Case $\Omega=R^m$. When $\Omega=R^m$, it is well-known [17, p. 171] that the time-varying system (S) is completely controllable (globally R^m-null controllable at t_0 in our notation) if and only if the rows of $\phi(t_0,\cdot)B(\cdot)$ are linearly independent on some bounded interval $[t_0, T]$. Here we show that when $\Omega=R^m$, (2.1) is a necessary and sufficient condition for global R^m-null controllability. This is accomplished by showing that (2.1) is equivalent to the rows of $\phi(t_0,\cdot)B(\cdot)$ being linearly independent on some bounded interval $[t_0, T]$.

PROPOSITION 3.3. (S) *is globally R^m-null controllable if and only if*

$$\int_{t_0}^{\infty} H_{R^m}(B'(\tau)z(\tau))\, d\tau = +\infty$$

for all nonzero solutions $z(\cdot)$ of (S').

The proof of this result is in Appendix B.

4. Some computational aspects. In a large number of problems, one may have to resort to the computer to check whether or not a system is Ω-null controllable. When using (2.3), a solution of the minimization problem $\min\{J(x_0, T, \lambda)\colon \lambda \in \Lambda\}$ is needed. Direct application of so-called gradient or descent algorithms to compute $\min\{J(x_0, T, \lambda)\colon \lambda \in \Lambda\}$ is precluded by the fact that $J(x_0, T, \lambda)$ is, in general, not differentiable in λ. This fact is a consequence of the sup-operation involved in the definition of $H_\Omega(B'(\tau)\phi'(T,\tau)\lambda)$. Fortunately, however, numerical computation of $\min\{J(x_0,T,\lambda): \lambda\in\Lambda\}$ is feasible if "generalized steepest descent" schemes are used. These schemes rely on subdifferential[3] rather than gradient information. The next two lemmas develop a description of the subdifferential of $J(x_0,T,\lambda)$. The proofs are given in Appendix C.

LEMMA 4.1. *For fixed* $(x_0,T)\in R^n\times R$, $J(x_0,T,\lambda)$ *is a lower semicontinuous convex function of* λ.

LEMMA 4.2. *For fixed* $(x_0,T)\in R^n\times R$, *the subdifferential of* $J(x_0,T,\cdot)$ *at* $\lambda\in R^n$ *consists of all vectors* $\lambda_*\in R^n$ *of the form*

$$\lambda_*=\phi(T,t_0)x_0+\int_{t_0}^{T}\phi(T,\tau)B(\tau)\omega_*(\tau)\,d\tau, \tag{4.1}$$

where

$$\begin{aligned}\omega_*(\tau)&\in\arg\max\{\omega'B'(\tau)\phi'(T,\tau)\lambda:\omega\in\Omega\}\\&=\{\omega\in\Omega:\omega'B'(\tau)\phi'(T,\tau)\lambda\geqq\eta'B'(\tau)\phi'(T,\tau)\lambda\ \forall\eta\in\Omega\}\end{aligned} \tag{4.2}$$

for almost all $\tau\in[0,T]$.

Remark. Since $J(x_0,T,\lambda)$ is the support function on the attainable set (see discussion preceding Theorem 2.3), a geometric interpretation of the subdifferential at λ is available: This set consists of all vectors in the normal cone to the attainable set at λ. (See Goodman [24, p. 285].)

Formulae (4.1) and (4.2) hold for arbitrary compact-convex Ω. Often, however, more structural information is known about Ω. In such cases, (4.1) and (4.2) may simplify. To illustrate, suppose

$$\Omega=[-M_1,M_1]\times[-M_2,M_2]\times\cdots\times[-M_m,M_m],\qquad M_i>0.$$

Then, the maximum in (4.2) is achieved in the ith component by

$$[\omega_*(\tau)]_i\in M_i\operatorname{sgn}[B'(\tau)\phi'(T,\tau)\lambda]_i,\qquad i=1,2,\cdots,m,$$

where $\operatorname{sgn}x\triangleq1$ if $x>0$; $\operatorname{sgn}x\triangleq-1$ if $x<0$; $\operatorname{sgn}0\triangleq[-1,1]$. Consequently, for this case, we can substitute into (4.1), and show that the subdifferential $\partial J(x_0,T,\lambda)$ consists of all vectors $\lambda_*\in R^n$ of the form

$$\lambda_*=\phi(T,0)x_0+\int_0^T\sum_{i=1}^{m}M_ih_i(T,\tau)\operatorname{sgn}\lambda'h_i(T,\tau)\,d\tau, \tag{4.3}$$

where $h_i(T,\tau)$ is the ith column of $H(T,\tau)\triangleq\phi(T,\tau)B(\tau)$. This description of the subdifferentials of $J(x_0,T,\cdot)$ can be used in conjunction with the generalized steepest descent algorithms to compute $\min\{J(x_0,T,\lambda):\lambda\in\Lambda\}$.

We also note that λ_* is uniquely specified by (4.3) if

$$\text{measure}\{\tau:\lambda'h_i(T,\tau)=0\}=0\quad \textit{for } i=1,2,\cdots,m.$$

For such λ, $\partial J(x_0,T,\lambda)$ is precisely $\nabla_\lambda J(x_0,T,\lambda)$, the gradient of $J(x_0,T,\cdot)$ at λ.

5. The steering control. Using the results of § 2, we can determine if (S) is Ω-null controllable. However, those results do not give a method for determining a *steering control* $u_*(\cdot)\in\mathcal{M}(\Omega)$ which accomplishes this objective.

[3] $\lambda_*\in\partial J(x_0,T,\lambda)$, the subdifferential of $J(x_0,T,\cdot)$ at λ, if and only if

$$J(x_0,T,z)\geqq J(x_0,T,\lambda)+(z-\lambda)'\lambda_*\quad\text{for all } z\in R^n.$$

One method of determining an appropriate $u_*(\cdot)$ is to solve the time optimal control problem, i.e., find $u_*(\cdot)\in\mathcal{M}(\Omega)$ which steers (S) from given (x_0, t_0) to the origin and does so in minimum time. If there is a control which steers the system to the origin, then there is a time optimal one [2]. Hence, in principle, a steering control can be numerically computed using any of a wide variety of algorithms which are available for solution of the time optimal control problem.

Since the solution of the time optimal problem is determined by solving a two point boundary value problem, it can be quite difficult to obtain the steering control this way. In this section, a "simpler" alternative method for generating a steering control is presented. This technique does not involve a two point boundary value problem and leads to a control which steers the system arbitrarily close to the origin. Our result is obtained from the following minimum norm problem:[4] Given initial point (x_0, t_0) and a final time T, find $u(\cdot)\in\mathcal{M}(\Omega)$ which leads to the smallest value of $\|x(T)\|$. The solution of this minimum norm problem is characterized in the next theorem.

THEOREM 5.1. *Let (x_0, t_0) and T be given. Suppose that $\lambda_*\in R^n$ achieves the minimum of $J(x_0, T, \lambda)$ over the closed unit ball. Then any solution of the minimum norm problem satisfies*

$$(5.1)\qquad u_*(\tau)\in\arg\max\{\omega' B'(\tau)\phi'(T,\tau)\lambda_*: \omega\in\Omega\}$$

for almost all $\tau\in[t_0, T]$.

See Appendix D for proof.

We note that condition (5.1) will uniquely determine $u_*(\cdot)$ whenever the minimum of $\omega' B'(\tau)\phi'(T,\tau)\lambda_*$ is uniquely achieved. For example, suppose

$$\Omega=[-M_1, M_1]\times[-M_2, M_2]\times\cdots\times[-M_m, M_m],\qquad M_i>0.$$

Then (5.1) requires

$$(5.2)\qquad [u_*(\tau)]_i\in M_i \operatorname{sgn}[B'(\tau)\phi'(T,\tau)\lambda_*]_i,\qquad i=1,2,\cdots,m.$$

For the case when the minimum of $\|x(T)\|=0$, $\lambda_*=0$ and (5.1) will not determine a control which steers (S) to the origin. The following heuristic procedure can be used to determine a control which steers (S) arbitrarily close to the origin: Choose a T such that the minimum of $\|x(T)\|$ is nonzero. As T is increased, the minimum of $\|x(T)\|$ approaches zero and the corresponding solution $u_*(\cdot)$, generated via (5.2), of the minimum norm problem results in a control which steers the system progressively closer to the origin.

In our next theorem, we provide another useful characterization of steering controls. For fixed $T\in[0,\infty)$, $x_0\in R^n$, we define the functional $V_T: R^n\times\mathcal{M}(\Omega)\to R$ by

$$V_T(\lambda, u(\cdot))=\lambda'\phi(T,0)x_0+\int_0^T \lambda'\phi(T,\tau)B(\tau)u(\tau)\,d\tau.$$

THEOREM 5.2. *Pick any compact convex set Λ containing zero as an interior point. Then $V_T(\lambda, u(\cdot))$ possesses at least one saddle point $(\lambda_*, u_*(\cdot))\in\Lambda\times\mathcal{M}(\Omega)$. Moreover, $u_*(\cdot)$ steers x_0 to zero at time T if and only if $V_T(\lambda_*, u_*(\cdot))=0$.*

See Appendix D for proof.

6. Additional applications. In this section, we use our results to obtain an existence theorem for the time optimal control problem and also apply our results to a pursuit game.

[4] Here the linear system (S) is required to be R^m-null controllable.

6.1. Existence of time optimal controls. Consider the following time optimal control problem: Find $u(\cdot)\in \mathcal{M}(\Omega)$ which drives the state $x(\cdot)$ of (S) from an initial position $x(t_0)=x_0$ to the origin and minimizes

$$C(u(\cdot))=\int_{t_0}^{t_f} dt; \qquad t_f = \text{arrival time at the origin.}$$

The classical theorem for existence of a time optimal control (e.g., Lee and Markus [2]) requires that there is at least one control which transfers the state $x(\cdot)$ of (S) to the origin. Combining the result of [2] with our Theorem 2.3, we obtain the following existence lemma.

LEMMA 6.1. *There exists a solution to the time optimal control problem if and only if there is some finite $t_f\in[t_0,\infty)$ such that*

$$\min\{J(x_0,t_f,\lambda):\lambda\in\Lambda\}=0.$$

Furthermore, the time optimal cost is given by

$$C^*(u_*(\cdot))=\min\{t_f:\min[J(x_0,t_f,\lambda):\lambda\in\Lambda]=0\}.$$

6.2. Pursuit Games. Next, we consider the pursuit game studied by Hájek [18]. The system is described by

$$\dot{x}(t)=Ax(t)-p(t)+q(t), \qquad p(t)\in P,\quad q(t)\in Q,\quad x(t_0)=x_0, \tag{6.1}$$

where P and Q are compact convex subsets of R^n. The pursuer $p(\cdot)$ seeks a strategy $\sigma: Q\times[t_0,\infty)\to P$ which steers $x(\cdot)$ to the origin for all possible quarry controls $q(\cdot):[t_0,\infty)\to Q$. A quarry control is admissible if it is measurable and a strategy is admissible if $\sigma(\cdot)$ preserves measurability.

In [18], a solution to this problem is obtained in terms of the *associated control system*

$$\dot{y}(t)=Ay(t)-u(t), \qquad u(t)\in P\overset{*}{-}Q, \qquad y(t_0)=x_0, \tag{6.2}$$

where $P\overset{*}{-}Q$ is the Pontryagin difference; i.e.,

$$P\overset{*}{-}Q\triangleq\{x\in R^n: x+Q\subseteq P\}.$$

Admissible controls $u(\cdot)$ above must be measurable.

Simply put, Hájek's result says that the state $x(\cdot)$ of (6.1) can be forced to the origin, for all admissible $q(\cdot)$, if and only if the state $y(\cdot)$ of (6.2) can be steered to the origin. More precisely, the following theorem is available.

FIRST RECIPROCITY THEOREM [18]. *Initial position x_0 in* (6.1) *can be* (*stroboscopically*) *forced to the origin at time $T\geqq t_0$ by a strategy $\sigma(\cdot)$ if and only if x_0 in* (6.2) *can be steered to the origin at time T by an admissible control $u(\cdot)$. Furthermore, $\sigma(\cdot)$ and $u(\cdot)$ are related by*

$$\sigma(q,t)=u(t)+q. \tag{6.3}$$

By applying Theorem 2.3 to (6.2), we obtain another condition for determining if (6.1) can be forced to the origin.

LEMMA 6.2. *Assume $P\overset{*}{-}Q$ compact. Pick any subset Λ of R^n containing zero as an interior point. Then x_0 in* (6.1) *can be forced to the origin at time $T\geqq t_0$ by a strategy $\sigma(\cdot)$ if and only if*

$$\min\{K(x_0,T,\lambda):\lambda\in\Lambda\}=0,$$

where

$$K(x_0, T, \lambda) \triangleq x_0' e^{A'(T-t_0)}\lambda + \int_{t_0}^{T} H_{P\pm Q}\,(e^{A'(T-\tau)}\lambda)\,d\tau.$$

It should be pointed out that in addition to pursuit game interpretation of (6.1), (6.1) can also be viewed as a problem of steering a system with disturbances to the origin if $q(\cdot)$ is thought of as a disturbance. Also, the results apply to systems described by

$$\dot{x}(t) = Ax(t) + Bp(t) + Cq(t), \qquad p(t) \in P, \quad q(t) \in Q$$

if one replaces $Bp(t)$ by $p'(t)$, $Cq(t)$ by $-q'(t)$, P by BP, and Q by CQ.

Appendix A. Proof of Theorems 2.1, 2.3 and Corollary 2.2. Since Theorem 2.3 is used in the proof of Theorem 2.1, we first present the proof of Theorem 2.3. There are many ways to prove Theorem 2.3; our proof exploits the convexity of the attainable set in conjunction with a measurable selection theorem. We note that a proof of the sufficiency part of the theorem is given in [6, Thm. 7.2.1]. To simplify our notation, we henceforth take $t_0 = 0$ without loss of generality. This will apply to subsequent appendices as well.

Proof of Theorem 2.3. Let $A_T(x_0)$ be the set of states which can be attained from x_0 at time T, i.e.,

$$A_T(x_0) = \left\{\phi(T, 0)x_0 + \int_0^T \phi(T, \tau)B(\tau)u(\tau)\,d\tau \colon u(\cdot) \in \mathcal{M}(\Omega)\right\}. \tag{A.1}$$

The set $A_T(x_0)$ is convex and compact [2]. From Definition 1.1, it follows that x_0 can be steered to 0 at time T if and only if $0 \in A_T(x_0)$ or, equivalently, by the separating hyperplane theorem [21],

$$0 \leqq \sup\{\lambda' a \colon a \in A_T(x_0)\} \tag{A.2}$$

for all vectors $\lambda \in R^n$. Using (A.1), requirement (A.2) becomes

$$\lambda'\phi(T, 0)x_0 + \sup\left\{\int_0^T \lambda'\phi(T, \tau)B(\tau)u(\tau)\,d\tau \colon u(\cdot) \in \mathcal{M}(\Omega)\right\} \geqq 0 \tag{A.3}$$

for all $\lambda \in R^n$. As a consequence of the measurable selection theory of [19], we can commute the supremum and integral operations in (A.3)[5]. Thus, $0 \in A_T(x_0)$ if and only if

$$0 \leqq \lambda'\phi(T, 0)x_0 + \int_0^T H_\Omega(B'(\tau)\phi'(T, \tau)\lambda)\,d\tau = J(x_0, T, \lambda) \tag{A.4}$$

for all $\lambda \in R^n$. Since $J(x_0, T, \lambda)$ is positively homogeneous in λ, we can restrict λ to Λ in (A.4), Theorem 2.3 now follows. □

Next, we present the proof of Theorem 2.1. In the proof, Theorem 2.3 is used.

Proof of Theorem 2.1 (*Necessity*). We suppose that (S) is globally Ω-null controllable at $t_0 = 0$. Let $z(\cdot)$ be any nonzero solution of (S'); we must prove that

$$\int_0^\infty H_\Omega(B'(\tau)z(\tau))\,d\tau = +\infty. \tag{A.5}$$

[5] $\phi(T, \tau)\ B(\tau)$ being a Carthéodory function enables us to apply the results of [19].

Proceeding by contradiction, suppose there is a nonzero solution $\hat{z}(\cdot)$ such that

$$\int_0^\infty H_\Omega(B'(\tau)\hat{z}(\tau))\,d\tau=\alpha,\qquad \alpha<\infty.$$

Then there is a positive constant $\beta<\infty$ such that

$$\int_0^\infty H_\Omega(B'(\tau)\hat{z}(\tau))\,d\tau<\beta.$$

Define

$$x_0^*\triangleq\frac{-2\beta\hat{z}(0)}{\hat{z}'(0)\hat{z}(0)},\qquad x_0^*\neq 0.$$

We now claim that x_0^* *cannot* be steered to zero by an admissible control $u(\cdot)\in\mathcal{M}(\Omega)$. To prove our claim, for each $t\in[0,\infty)$, define

$$\lambda_t\triangleq\phi'(0,t)\hat{z}(0),\qquad \lambda_t\neq 0.$$

Now, given any $t\in[0,\infty)$,

$$\begin{aligned}J(x_0^*,t,\lambda_t)&=x_0^{*\prime}\phi'(t,0)\lambda_t+\int_0^t H_\Omega(B'(\tau)\phi'(t,\tau)\lambda_t)\,d\tau\\&=x_0^{*\prime}\hat{z}(0)+\int_0^t H_\Omega(B'(\tau)\hat{z}(\tau))\,d\tau\\&\leqq-2\beta+\beta\\&<0.\end{aligned}$$

Taking $\Lambda=R^n$ in Theorem 2.3, it follows that

$$\min\{J(x_0^*,t,\lambda):\lambda\in\Lambda\}\leqq J(x_0^*,t,\lambda_t)<0$$

for all $t\in[0,\infty)$. By Theorem 2.3, (S) is *not* Ω-null controllable at $(x_0^*,0)$. □

(*Sufficiency*). Now, we assume that (A.5) holds. Again, we proceed by contradiction, i.e., suppose (S) is *not* globally Ω-null controllable at $t_0=0$. Hence, there exists an initial condition $x_0^*\neq 0$ which cannot be steered to zero. By Theorem 2.3 (with $\Lambda=R^n$), we can find a sequence of times $\langle t_k\rangle_{k=1}^\infty$ and a sequence of vectors $\langle\lambda_k\rangle_{k=1}^\infty$ having the following properties:

(P1) $$\lim_{k\to\infty} t_k=+\infty,$$

(P2) $$J(x_0^*,t_k,\lambda_k)<0\quad\text{for } k=1,2,3\cdots.$$

We are going to construct an initial condition $\tilde{z}_0\neq 0$ for (S') which makes the integral in (A.5) finite. To meet this end, let

$$z_k=\frac{\phi'(t_k,0)\lambda_k}{\|\phi'(t_k,0)\lambda_k\|},\qquad k=1,2,\cdots.$$

We note that each z_k above is nonzero because $\lambda_k\neq 0$, and $\phi(t_k,0)$ is invertible. Then $\langle z_k\rangle_{k=1}^\infty$ is a sequence in R^n belonging to the set

$$S\triangleq\{z\in R^n:\|z\|=1\}.$$

Since S is compact, we can extract a subsequence $\langle z_{k_j}\rangle_{j=1}^\infty$ which converges to some vector $\tilde{z}_0 \in S$. We will now show that $\tilde{z}_0$ is the initial condition which we seek. Let $\tilde{z}(\cdot)$ be the trajectory of (S') generated by $z(0) \triangleq \tilde{z}_0$; let $\langle t_{k_j}\rangle_{j=1}^{\infty}$ denote the subsequence of times corresponding to $\langle z_{k_j}\rangle_{j=1}^{\infty}$. By (P1), we have

$$\lim_{j \to \infty} t_{k_j} = +\infty,$$

and by (P2), it follows that

$$x_0^{*\prime}\phi'(t_{k_j}, 0)\lambda_{k_j} + \int_0^{t_{k_j}} H_\Omega(B'(\tau)\phi'(t_{k_j}, \tau)\lambda_{k_j})\, d\tau < 0 \quad \textit{for } j = 1, 2, 3, \cdots .$$

Dividing by $\|\phi'(t_{k_j}, 0)\lambda_{k_j}\|$ and noting that H_Ω is positively homogeneous, we obtain

$$\begin{aligned} \int_0^{t_{k_j}} H_\Omega(B'(\tau)\phi'(0, \tau)z_{k_j})\, d\tau &\leqq \|x_0^*\|\,\|z_{k_j}\| \quad \textit{for } j = 1, 2, 3, \cdots, \\ &\leqq \|x_0^*\| \qquad\quad \textit{for } j = 1, 2, 3, \cdots . \end{aligned}$$

We would like to obtain an inequality involving $\tilde{z}_0$ with an infinite upper limit on this integral. To accomplish this, we define

$$f_{k_j}(\tau) \triangleq \begin{cases} H_\Omega(B'(\tau)\phi'(0, \tau)z_{k_j}) & \text{if } \tau \in [0, t_{k_j}], \\ 0 & \text{otherwise,} \end{cases} \qquad j = 1, 2, 3, \cdots,$$

$$f(\tau) \triangleq H_\Omega(B'(\tau)\phi'(0, \tau)\tilde{z}_0), \qquad \tau \in [0, \infty),$$

and make the following observations.

(i) $\int_0^\infty f_{k_j}(\tau)\, d\tau$ is bounded (by $\|x_0^*\|$) for $j = 1, 2, 3, \cdots$.

(ii) $f_{k_j}(\tau)$ converges pointwise to $f(\tau)$ on $[0, \infty)$. This observation is proven using the facts that $z_{k_j} \to \tilde{z}_0$, $t_{k_j} \to +\infty$, and H_Ω depends continuously on its argument.

Applying Fatou's lemma [20, p. 83], we have

$$\begin{aligned} \int_0^\infty f(\tau)\, d\tau &\leqq \liminf_{j \to \infty} \int_0^\infty f_{k_j}(\tau)\, d\tau \\ &\leqq \limsup_{j \to \infty} \int_0^\infty f_{k_j}(\tau)\, d\tau \\ &\leqq \|x_0^*\|. \end{aligned}$$

Substitution for $f(\tau)$ above gives

$$\int_0^\infty H_\Omega(B'(\tau)\phi'(0, \tau)\tilde{z}_0)\, d\tau \leqq \|x_0^*\|,$$

i.e.,

$$\begin{aligned} \int_0^\infty H_\Omega(B'(\tau)\tilde{z}(\tau))\, d\tau &\leqq \|x_0^*\| \\ &< \infty, \end{aligned}$$

which is the contradiction that we seek. This completes the proof of the theorem. □

Proof of Corollary 2.2. Suppose Ω and Ω' satisfy the hypotheses of the corollary. We are going to show that (S) is globally Ω'-null controllable. To prove this, it is sufficient to find a subset $\Omega'_\delta \subseteq \Omega'$ such that (S) is globally Ω'_δ-null controllable: Pick $\delta > 0$ such that

$$\Omega'_\delta \triangleq \{\omega : \|\omega\| \leqq \delta\} \subseteq \Omega'.$$

(This can be accomplished because zero is interior to Ω'.) Now, to prove that Ω'_δ has the desired property, we pick $R > 0$ such that

$$\Omega_R \triangleq \{\omega : \|\omega\| \leqq R\} \supseteq \Omega.$$

(This can also be done since Ω is compact, hence bounded.) Let $z(\cdot)$ be any nonzero solution of (S'). Then we have

$$\begin{aligned}\int_0^\infty H_{\Omega'_\delta}(B'(\tau)z(\tau))\,d\tau &= \int_0^\infty \sup\{\omega' B'(\tau)z(\tau) : \|\omega\| \leqq \delta\}\,d\tau\\ &= \delta \int_0^\infty \|B'(\tau)z(\tau)\|\,d\tau\\ &= \frac{\delta}{R}\int_0^\infty R\|B'(\tau)z(\tau)\|\,d\tau\\ &= \frac{\delta}{R}\int_0^\infty \sup\{\omega' B'(\tau)z(\tau) : \|\omega\| \leqq R\}\,d\tau\\ &= \frac{\delta}{R}\int_0^\infty H_{\Omega_R}(B'(\tau)z(\tau))\,d\tau\\ &= +\infty\end{aligned}$$

since (S) is globally Ω_R-null controllable. (Ω_R-null controllability follows from Ω-null controllability in conjunction with the fact that $\Omega_R \supseteq \Omega$.) By Theorem 2.1, we conclude that (S) must be globally Ω'_δ-null controllable and hence Ω'-null controllable. □

Appendix B.

Proof of Theorem 3.2. (*i*) This condition follows immediately from the fact that global R^m-null controllability is necessary for global Ω-null controllability.

It is also possible to prove (i) directly from Theorem 2.1. Suppose (S) is globally Ω-null controllable but rank $(Q) < n$. Then there exists a $v \in R^n$, $v \neq 0$, such that $B' e^{-A't}v = 0$ for all $t \geqq 0$. Let $z(0) = v$. Then $z(\tau) = e^{-A't}v$ and

$$\int_0^\infty \sup_{\omega\in\Omega}(\omega' B' z(\tau))\,d\tau = \int_0^\infty \sup_{\omega\in\Omega}(\omega' B' e^{-A't}v)\,d\tau = 0$$

which contradicts Theorem 2.1.

(ii) Suppose (S) is globally Ω-null controllable but there exists a real eigenvector v of A' satisfying $\omega' B' v \leqq 0$ for all $\omega \in \Omega$. Denoting by λ the real eigenvalue associated with v, we have $e^{-A't}v = e^{-\lambda t}v$. With $z(0) = v$, $z(\tau) = e^{-A'\tau}v = e^{-\lambda\tau}v$, and

$$\begin{aligned}\int_0^\infty \sup_{\omega\in\Omega}(\omega' B' z(\tau)\,d\tau &= \int_0^\infty \sup_{\omega\in\Omega}(\omega' B' e^{-\lambda\tau}v)\,d\tau\\ &= \int_0^\infty e^{-\lambda\tau}\sup_{\omega\in\Omega}(\omega' B' v)\,d\tau.\end{aligned}$$

Now this integral is less than or equal to zero since $\sup\{\omega' B' v : \omega \in \Omega\} \leqq 0$ and $e^{-\lambda t} \geqq 0$. This contradicts Theorem 2.1.

(iii) Again the proof is by contradiction. Assume (S) is globally Ω-null controllable but A has an eigenvalue λ with a positive real part. Then λ is also an eigenvalue of A' so that $A'v = \lambda v$, w'ere v is an eigenvector corresponding to A'. Let $\bar{\lambda}$ and $\bar{v}$ denote the complex conjugate of λ and v. They satisfy $A\bar{v} = \bar{\lambda}\bar{v}$. Hence,

$$e^{-A't}v = e^{\lambda t}v \quad \text{and} \quad e^{-A't}\bar{v} = e^{\bar{\lambda}t}\bar{v}.$$

Consider the solution of the adjoint equation corresponding to the initial condition $z(0) = v + \bar{v}$. (Note that $z(0)$ is real.) For this $z(0)$,

$$\begin{aligned}\sup_{\omega\in\Omega}(\omega' B' z(\tau)) &= \sup_{\omega\in\Omega}(\omega' B' \, e^{-A'\tau}(v+\bar{v})) \\ &= \sup_{\omega\in\Omega}[\omega' B'(e^{-\lambda t}v + e^{-\bar{\lambda}t}\bar{v})] \\ &= \sup_{\omega\in\Omega}\{\omega' B' \, e^{-at}[2m \cos bt + 2n \sin bt]\},\end{aligned}$$

where a and b are the real part and imaginary part of λ and n and m are the real part and imaginary part of v. Let $M \triangleq \sup\{\sup[\omega' B'(2n \cos bt + 2n \sin bt) : \omega \in \Omega] : t \geqq 0\}$. M is finite since Ω is compact, i.e., $M \leqq 2 \max\{|n|, |m|\} \, \|B\| \sup\{\|\omega\| : \omega \in \Omega\}$. Thus

$$\sup_{\omega\in\Omega}(\omega' B' z(\tau)) \leqq M e^{-at},$$

and

$$\int_0^\infty \sup_{\omega\in\Omega}(\omega' B' z(\tau))\, d\tau \leqq M \int_0^\infty e^{-at}\, dt.$$

The integral on the right is finite since $a > 0$ and we have a contradiction to Theorem 2.1. □

Proof of Proposition 3.3. (*Necessity*). Suppose (S) is globally R^m-null controllable. Then there is a finite interval $[0, T]$ on which the rows of $\phi(0, \cdot)B(\cdot)$ are linearly independent. Thus, for every nonzero vector $z_0 \in R^n$, it follows that $B'(t)\phi'(0, t)z_0 \neq 0$ for some $t \in [0, T]$. Since $B'(\cdot)\phi'(0, \cdot)z_0$ is continuous, there must be an interval $I = [t - \delta, t + \delta]$ on which $B'(\tau)\phi'(0, \tau)z_0 \neq 0$ for all $\tau \in I$. On this interval, we have

$$\sup\{\omega B'(\tau)\phi'(0, \tau)z_0 \colon \omega \in R^m\} = +\infty.$$

Hence, using the nonnegativity of $H_\Omega(\cdot)$, we conclude that

$$\begin{aligned}\int_0^\infty H_{R^m}(B'(\tau)z(\tau))\, d\tau &\geqq \int_I H_{R^m}(B'(\tau)\phi'(0, \tau)z_0)\, d\tau \\ &= \int_I \sup\{\omega' B'(\tau)\phi'(0, \tau)z_0 \colon \omega \in R^m\}\, d\tau \\ &= +\infty.\end{aligned}$$

(*Sufficiency*). Proceeding by contradiction, we suppose that for all nonzero solutions $z(\cdot)$ of (S′), we have

$$\int_0^\infty H_{R^m}(B'(\tau)z(\tau))\, d\tau = +\infty,$$

but the columns of $B'(\cdot)\phi'(0, \cdot)$ are linearly dependent on every bounded interval $[0, T]$. Let $\langle T_n\rangle_{n=1}^\infty$ be a monotone increasing sequence of times such that $T_n \to \infty$. Then, for each n, we can find a nonzero vector $\tilde{z}_n$ such that $B'(\tau)\phi'(0, \tau)\tilde{z}_n \equiv 0$ on $[0, T_n]$. Let

$$z_n \triangleq \frac{\tilde{z}_n}{\|\tilde{z}_n\|} \quad \text{for } n = 1, 2, \cdots,$$

Then, $\langle z_n \rangle_{n=1}^{\infty}$ is a sequence in the (compact) unit ball. Hence, we can extract a subsequence z_{n_j} converging to some $\hat{z}_0$, $\|\hat{z}_0\| = 1$. We notice that the corresponding subsequence of times T_{n_j} still converges to $+\infty$. Furthermore, for each fixed $\tau \in [0, \infty)$, we have

$$\begin{aligned} B'(\tau)\phi'(0, \tau)\hat{z}_0 &= \lim_{j \to \infty} B'(\tau)\phi'(0, \tau)z_{n_j} \\ &= 0. \end{aligned}$$

Consequently, if $\hat{z}(\tau)$ is the trajectory mate of $\hat{z}_0$,

$$\int_0^{\infty} H_{R^m}(B'(\tau)\hat{z}(\tau))\, d\tau = \int_0^{\infty} \sup\{\omega' B'(\tau)\phi'(0, \tau)\hat{z}_0 : \omega \in R^m\}\, d\tau = 0$$

which contradicts the assumed hypothesis. □

Appendix C.

Proof of Lemma 4.1. For (x_0, T) fixed, $J(x_0, T, \lambda)$ can be expressed as

$$J(x_0, T, \lambda) = \sup\{H_\omega(\lambda) : \omega(\cdot) \in \mathcal{M}(\Omega)\},$$

where

$$H_\omega(\lambda) \triangleq \lambda'\phi(T, 0)x_0 + \int_0^T \lambda'\phi(T, \tau)B(\tau)\omega(\tau)\, d\tau.$$

Consequently, $J(x_0, T, \cdot)$ is the pointwise supremum over an indexed collection of continuous linear (hence convex) functions. Hence $J(x_0, T, \cdot)$ itself must be convex and at least lower semicontinuous (in fact, continuous). □

Proof of Lemma 4.2. We prove this lemma using some of the standard properties of subdifferentials given in Rockafellar [21], [22]. Since both functions in the definition of $J(x_0, T, \lambda)$ are finite and convex, $\lambda_* \in \partial J(x_0, T, \lambda)$ if and only if

$$\begin{aligned} \lambda_* &\in \partial(x_0'\phi'(T, 0)\lambda) + \partial \int_0^T H_\Omega(B'(\tau)\phi'(T, \tau)\lambda)\, d\tau \qquad \text{(by Theorem 23.8 of [22]).} \\ &= \phi(T, 0)x_0 + \int_0^T \partial H_\Omega(B'(\tau)\phi'(T, \tau)\lambda)\, d\tau \qquad \text{(by Theorem 23 of [22])} \\ &= \phi(T, 0)x_0 + \int_0^T \phi(T, \tau)B(\tau) \cdot \partial H_\Omega(\hat{\omega}(\tau))\big|_{\hat{\omega}(\tau) = B'(\tau)\phi'(T, \tau)\lambda}\, d\tau \\ &\qquad \text{(by Theorem 23.9 of [21]).} \end{aligned}$$

Now, by Corollary 23.5.3 of [21], $\omega_*(\tau) \in \partial H_\Omega(\hat{\omega}(\tau))$ if and only if $\omega_*(\tau) \in \arg\max\{\omega'\hat{\omega}(\tau) : \omega \in \Omega\}$. Substituting the required form for $\hat{\omega}$ above, we obtain our desired representation for λ_*. □

Appendix D.

Sketch of a proof of Theorem 5.1. Let $f: L^1(0, T; R^m) \to R$, $g: R^n \to R$, $\Lambda_T: L^1(0, T; R^m) \to R^n$ be given by

$$f(u) \triangleq \begin{cases} 0 & \text{if } u(\cdot) \in \mathcal{M}(\Omega) \\ +\infty & \text{otherwise,} \end{cases}$$

$$g(z) \triangleq -\|\phi(T, 0)x_0 + z\|, \qquad z \in R^n,$$

$$\Lambda_T u \triangleq \int_0^T \phi(T, \tau)B(\tau)u(\tau)\, d\tau.$$

Then, using the notation above

$$\begin{aligned} \inf(MN) &\triangleq \inf\{\|x(T)\|: u(\cdot) \in \mathcal{M}(\Omega)\} \\ &= \inf\{f(u) - g(\Lambda_T u): u \in L^1(0, T; R^m)\}. \end{aligned}$$

Written in this way, inf (MN) is in the standard form for application of Rockafellar's extension of Fenchel's duality theorem (cf. [23, Thm.1]). The functionals f and g are, respectively, proper convex and concave functions; it can be easily shown that inf (MN) is "stably set"—a technical precondition for Rockafellar's theorem.

By carrying out the computations involved in Theorem 1 of [23], it can be shown that the problem

$$\min(MN)^* \triangleq \min\{J(x_0, T, \lambda): \lambda \in \Lambda\}$$

is dual to inf (MN) in the following sense:

$$\inf(MN) + \min(MN)^* = 0.$$

The "extremality condition" in Rockafellar's theorem provides a necessary condition which must be satisfied by all solution pairs λ_* solving $(MN)^*$ and $u_*(\cdot)$ solving (MN). This extremality condition requires

$$\Lambda_T^* \lambda_* \in \partial f(u_*),$$

where Λ_T^* is the adjoint of Λ_T, and $\partial f(u_*)$ is the subdifferential of f at u_*. For our choice of f, this necessary condition particularizes to

$$\lambda'_* \phi(T, \tau)B(\tau) \in (\text{normal cone of } \mathcal{M}(\Omega) \text{ at } u_*(\cdot)).$$

We denote this normal cone at u_* by $N_c(u_*)$. By definition of the normal cone, we have $v(\cdot) \in N_c(u^*)$ if and only if

$$\int_0^T u'_*(\tau)B'(\tau)\phi'(T, \tau)\lambda_* d\tau = \int_0^T \sup\{\omega' B'(\tau)\phi'(T, \tau)\lambda_*: \omega \in \Omega\}\, d\tau.$$

This is possible only if $\omega = u_*(\tau)$ achieves the supremum of $\omega' B'(\tau)\phi'(T, \tau)\lambda_*$ for almost all $\tau \in [0, T]$. Equivalently, we must have

$$u_*(\tau) \in \arg\max\{\omega' B'(\tau)\phi'(T, \tau)\lambda_*: \omega \in \Omega\}$$

for almost all $\tau \in [0, T]$.

Proof of Theorem 5.2. As in the proof of Theorem 2.3, let $A_T(x_0)$ be the set of states which can be attained from x_0 at time T. We recall that this set is compact and convex. Define $W_T: \Lambda \times A_T(x_0) \to R$ by

(D.1) $$W_T(\lambda, \xi) \triangleq \lambda' \xi.$$

In accordance with Proposition 2.3 of [19, p. 171], $W_T(\lambda, \xi)$ will possess a saddle point because the following conditions are satisfied:

(D.2.1) For all $\lambda \in \Lambda$, $W(\lambda, \cdot)$ is concave and upper semicontinuous.

(D.2.2) For all $\xi \in A_T(x_0)$, $W(\cdot, \xi)$ is convex and lower semicontinuous.

We note that

$$\min_{\lambda \in \Lambda} \max_{u(\cdot) \in \mathcal{M}(\Omega)} V_T(\lambda, u(\cdot)) = \min_{\lambda \in \Lambda} \max_{\xi \in A_T(x_0)} W_T(\lambda, \xi).$$

Furthermore,

$$\max_{u(\cdot) \in \mathcal{M}(\Omega)} \min_{\lambda \in \Lambda} V_T(\lambda, u(\cdot)) = \max_{\xi \in A_T(x_0)} \min_{\lambda \in \Lambda} W_T(\lambda, \xi).$$

These equalities, in conjunction with the fact that W_T possesses a saddle point, imply that V_T also has a saddle point.

To prove the last part of the theorem, we take $(\lambda_*, u_*(\cdot))$ to be a given saddle point of $V_T(\lambda, u(\cdot))$. Hence we have

$$\text{(D.3)} \qquad V_T(\lambda_*, u_*(\cdot)) = \min_{\lambda \in \Lambda} \max_{u(\cdot) \in \mathcal{M}(\Omega)} V_T(\lambda, u(\cdot)).$$

Using a measurable selection argument, as in the proof of Theorem 2.3, it is also apparent that

$$\text{(D.4)} \qquad \min_{\lambda \in \Lambda} \max_{u(\cdot) \in \mathcal{M}(\Omega)} V_T(\lambda, u(\cdot)) = \min_{\lambda \in \Lambda} J(x_0, T, \lambda).$$

From (D.3) and (D.4) we conclude that

$$\text{(D.5)} \qquad V_T(\lambda_*, u_*(\cdot)) = \min_{\lambda \in \Lambda} J(x_0, T, \lambda).$$

From Theorem 2.3 and the comments following the theorem, we know that x_0 can be steered to zero at time T if and only if

$$\begin{aligned} 0 &= \min_{\lambda \in \Lambda} J(x_0, T, \lambda) \\ &= V_T(\lambda_*, u_*(\cdot)) \qquad \text{(by (D.5)).} \end{aligned}$$

To complete the proof, we must show that if $V_T(\lambda_*, u_*(\cdot)) = 0$, then $u^*(\cdot)$ steers x_0 to 0. Now

$$0 = V_T(\lambda_*, u_*(\cdot)) \leqq V_T(\lambda, u_*(\cdot)) \quad \text{for all } \lambda \in \Lambda$$

or

$$0 \leqq \lambda' \left[\phi(T, 0)x_0 + \int_0^T \phi(T, \tau)B(\tau)u_*(\tau)\, d\tau \right] \quad \text{for all } \lambda \in \Lambda.$$

Thus

$$\text{(D.6)} \qquad 0 \leqq \lambda' x(T, x_0, u_*(\cdot)) \quad \text{for all } \lambda \in \Lambda.$$

Since 0 is an interior point of the convex, compact set Λ, (D.6) implies $x(T, x_0, u_*(\cdot)) = 0$ and $u_*(\cdot)$ is a steering control. □

Acknowledgment. The authors wish to thank the reviewers for their thorough reading of the paper. Their comments and suggestions led to many simplifications in the proofs.

REFERENCES

[1] R. E. KALMAN, *Mathematical description of linear dynamical system*, this Journal, 1 (1963), pp. 152–192.

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

[3] S. SAPERSTONE AND J. YORKE, *Controllability of linear oscillatory systems using positive controls*, this Journal, 9 (1971), pp. 253–262.

[4] R. BRAMMER, *Controllability in linear autonomous systems with positive controllers*, this Journal, 10 (1972), pp. 339–353.

[5] S. SAPERSTONE, *Global controllability of linear systems with positive controls*, this Journal, 11 (1973), pp. 417–423.

[6] A. FRIEDMAN, *Differential Games*, Wiley, New York, 1971.

[7] R. CONTI, *Teoria del Controllo e del Controllo Ottimo*, UTET, Torino, Italy, 1974.

[8] L. PANDOLFI, *Linear control systems: Controllability with constrained controls*, J. Optimization Theory Appl., 19 (1976), pp. 577–585.

[9] J. DAUER, *Controllability of nonlinear systems using a growth condition*, Ibid., 9 (1972), pp. 90–98.

[10] J. DAUER, *Controllability on nonlinear systems with restrained controls*, Ibid., 14 (1974), pp. 251–262.

[11] E. CHUKWU AND J. GRONSKI, *Controllability of nonlinear systems with restrained controls to closed convex sets*, Dept. of Mathematics Rep. CSUMD 45, Cleveland State Univ., Cleveland, OH, 1976.

[12] E. N. CHUCKWU AND S. D. SILLIMAN, *Complete controllability to a closed target set*, J. Optimization Theory Appl., 21 (1977), pp. 369–383.

[13] W. J. GRANTHAM AND T. L. VINCENT, *A controllability minimum principle*, Ibid., 17 (1975), pp. 93–114.

[14] M. E. EVANS AND D. N. P. MURTHY, *Controllability of discrete-time systems with positive controls*, IEEE Trans. Automatic Control, AC-22 (1977), pp. 942–945.

[15] M. PACHTER AND D. H. JACOBSON, *Control with conic constraint set*, J. Optimization Theory Appl., to appear.

[16] D. H. JACOBSON, *Extension of Linear-Quadratic Control, Optimization and Matrix Theory*, Academic Press, New York, 1977.

[17] C. T. CHEN, *Introduction to Linear System Theory*, Holt, Rinehart and Winston, New York, 1970.

[18] O. HÁJEK, *Pursuit Games*, Academic Press, New York, 1975.

[19] I. EKELAND AND R. TEMAN, *Convex Analysis and Variational Problems*, North-Holland, Amsterdam, 1976.

[20] H. L. ROYDON, *Real Analysis*, MacMillan, New York, 1968.

[21] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1972.

[22] ———, *Conjugate duality and optimization*, CBMS Regional Conference Series in Applied Mathematics No. 16, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1974.

[23] ———, *Duality and stability in extremum problems involving convex function*, Pacific J. Math., 21 (1967), pp. 167–187.

[24] G. S. GOODMAN, *Support Functions and the Integration of Set-Valued Mappings*, International Atomic Energy Agency, 2 (1976), pp. 281–296.

[25] M. HEYMANN AND R. J. STERN, *Controllability of linear systems with positive controls: Geometric considerations*, J. Math. Anal. Appl., 52 (1975), pp. 36–41.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 4, July 1980

0363-0129/80/1804-0002 $01.00/0

TRUNCATION ERROR BOUNDS AND CONVERGENCE OF LEAST SQUARES ESTIMATES*

YORAM BARAM†

Abstract. The error resulting from truncating a data record in least squares estimation of a Gaussian process is shown to be bounded in the mean square. The bounds are shown to be easily computable for linear processes, and the truncation error is shown to be strongly diminishing under a stability condition. These results have direct implications on filtering and prediction errors, data reduction and asymptotic analysis of parameter estimates.

1. Introduction. Consider the problem of estimating present or future values of a stochastic process using information which would normally consist of past values of the process itself or another related process. A realistic estimation procedure must be initialized at some finite time, which implies that any preceding information will be discarded. Furthermore, in many situations it may be desirable to discard some past information due to data storage and processing limitations.

This paper is concerned with the error arising in least squares estimation when a (possibly infinite) portion of the data record is discarded by truncation. In particular, bounds on this truncation error are of interest. A bound for general Gaussian processes is obtained, employing a general property of product measures. Specializing to linear processes, two bounds are obtained and shown to be computable by simple procedures. It is then shown that under a stability condition, the truncation error is strongly (almost surely and in the mean square) diminishing and not merely, as well known, in distribution. It is this strong convergence, which is significant in the asymptotic analysis of process estimation and parameter identification techniques.

2. Gaussian processes. Consider two stochastic processes, $x_t \in \mathbb{R}^n$ and $y_t \in \mathbb{R}^m$, where the time parameter t may take continuous or discrete values on different intervals of the real line. Let $\mathcal{F}_t$ and $\mathcal{B}_s$ denote the Borel fields generated by $(y_p, -\infty < p \leqq t)$ and $(y_p, 0 \leqq p \leqq s)$, respectively. Let $E^{\mathcal{F}_s}x_t$ and $E^{\mathcal{B}_s}x_t$ denote the conditional expectations or the least-squares estimates of x_t given $\mathcal{F}_s$ and $\mathcal{B}_s$, respectively. We denote the truncation error by $e^x_{t,s} = E^{\mathcal{F}_s}x_t - E^{\mathcal{B}_s}x_t$.

In the sequel, we shall restrict the discussion to the case where (x_t) and (y_t) are Gaussian processes, i.e., for each $t \in R$ any finite number of vectors from the set $(x_p, -\infty < p \leqq t)$ or from the set $(y_p, -\infty < p \leqq t)$ have a joint Gaussian distribution on the respective spaces. We shall denote by $\langle x_1, x_2 \rangle$ and $\|x\|$ the standard inner product and norm on $\mathbb{R}^n$.

THEOREM 2.1. *For the processes* (x_t) *and* (y_t) *defined above and for any* $t \geqq s > 0$ *we have*

$$E\|e^x_{t,s}\| \leqq E\|E^{\mathcal{F}_0}x_t\|. \tag{2.1}$$

Proof. x_t can be written as

$$x_t = E^{\mathcal{F}_0}x_t + \mu_t, \tag{2.2}$$

where

$$\mu_t = x_t - E^{\mathcal{F}_0}x_t \tag{2.3}$$

* Received by the editors December 13, 1978, and in final revised form September 24, 1979.

† Department of Electrical Engineering, Tel Aviv University, Ramat Aviv, Tel Aviv, Israel.

is Gaussian (e.g., [1, p. 13]) and uncorrelated with $(y_p, -\infty < p \leqq 0)$, (and with any $\mathcal{F}_0$-measurable random variable). It follows that μ_t is independent of $\mathcal{F}_0$ and thus, so is $E^{\mathcal{F}_s}\mu_t$. Now, since $\mathcal{F}_s$ is the smallest σ-field containing the Cartesian product of $\mathcal{F}_0$ and $\mathcal{B}_s$, in symbol

$$\mathcal{F}_s = \sigma(\mathcal{F}_0 \times \mathcal{B}_s), \tag{2.4}$$

and since $E^{\mathcal{F}_s}\mu_t$ is $\mathcal{F}_s$ measurable, then it is $\mathcal{B}_s$ measurable (see, e.g., [2, p. 96, Lemma 2]). It follows that

$$\begin{aligned} E\langle E^{\mathcal{F}_s}x_t - E^{\mathcal{B}_s}x_t, E^{\mathcal{B}_s}x_t - E^{\mathcal{F}_s}\mu_t\rangle &= EE^{\mathcal{B}_s}\langle E^{\mathcal{F}_s}x_t - E^{\mathcal{B}_s}x_t, E^{\mathcal{B}_s}x_t - E^{\mathcal{F}_s}\mu_t\rangle \\ &= E\langle E^{\mathcal{B}_s}x_t - E^{\mathcal{B}_s}x_t, E^{\mathcal{B}_s}x_t - E^{\mathcal{F}_s}\mu_t\rangle = 0, \end{aligned} \tag{2.5}$$

where we have used the $\mathcal{F}_s$-measurability of $E^{\mathcal{B}_s}$ and the smoothing property of the expectation. Hence,

$$\begin{aligned} E\|E^{\mathcal{F}_s}x_t - E^{\mathcal{B}_s}x_t\| \leqq E\|E^{\mathcal{F}_s}x_t - E^{\mathcal{B}_s}x_t\| + 2E\langle E^{\mathcal{F}_s}x_t - E^{\mathcal{B}_s}x_t, E^{\mathcal{B}_s}x_t - E^{\mathcal{F}_s}\mu_t\rangle \\ + E\|E^{\mathcal{B}_s}x_t - E^{\mathcal{F}_s}\mu_t\| \\ = E\|E^{\mathcal{F}_s}x_t - E^{\mathcal{F}_s}\mu_t\| = EE^{\mathcal{F}_s}\|x_t - \mu_t\| \\ = E\|E^{\mathcal{F}_0}x_t\| \end{aligned} \tag{2.6}$$

where the last equality follows from (2.2) and from the smoothing property of the expectation. The assertion is thus proven.

COROLLARY 2.2. *Suppose that*

$$\lim_{t\to\infty} E\|E^{\mathcal{F}_0}x_t\| = 0, \tag{2.7}$$

then the truncation error $e^x_{t,s}$ vanishes in the mean-square as $t \to \infty$.

Proof. The proof follows immediately from Theorem 2.1.

Remarks. 1) A case of interest is $s = t - T$ where T is a constant time lag.

2) Note that no particular relationship need be assumed between the processes (x_t) and (y_t). Normally, the fields generated by (y_t) will contain information on the process (x_t). A special case is $(y_t) = (x_t)$.

3. Linear processes. We now consider the case where the process (x_t) is generated by linear differential and difference equations with time varying coefficients. First consider

$$dx_t = A_t x_t\, dt + B_t\, dw_t, \qquad x_t \in \mathbb{R}^n, \tag{3.1}$$

where w_t is a zero-mean Gaussian process on $\mathbb{R}^p$ with uncorrelated increments (a Wiener process), uncorrelated with x_r for all $r \leqq t$. As before, let $\mathcal{F}_s$ and $\mathcal{B}_s$ be the Borel fields generated respectively by $(y_t, -\infty < t \leqq s)$ and $(y_t, 0 \leqq t \leqq s)$, where (y_t) is some Gaussian process on $\mathbb{R}^m$. We shall assume that w_t and y_r are uncorrelated for any $r \leqq t$; however, no particular relationship between (y_t) and (x_t) is assumed at this point. As in the previous section we denote $e^x_{t,s} \equiv E^{\mathcal{F}_s}x_t - E^{\mathcal{B}_s}x_t$. We shall denote by $\Phi_{t,0}$ the transition matrix corresponding to (3.1). For any $n \times n$ matrix A, we define the matrix norm as

$$\|A\| = \sup_{\|x\|=1} \|Ax\|, \qquad x \in \mathbb{R}^n. \tag{3.2}$$

$\|A\|$ is equal to the square root of the maximum eigenvalue of AA^T (e.g., [3 pp. 576, 577]).

3.1. Truncation error bounds.

THEOREM 3.1.1. *Let the process* (x_t) *be given by* (3.1) *and let the process* (y_t), *generating* $(\mathcal{F}_t)$ *be Gaussian. Then the truncation error* $e^x_{t,s}$ *is bounded as*

$$E\|e^x_{t,s}\| \leqq \|\Phi_{t,0}\| E\|E^{\mathcal{F}_0}x_0\| \leqq \|\Phi_{t,0}\| E\|x_0\|. \tag{3.3}$$

Proof. We have for any $t>0$

$$x_t = \Phi_{t,0}x_0 + \int_0^t \Phi_{t,\tau}B_\tau \, dw_\tau. \tag{3.4}$$

Since w_t and y_p are uncorrelated for all $p \leqq 0$, then, almost surely

$$E^{\mathcal{F}_0}x_t = \Phi_{t,0}\, E^{\mathcal{F}_0}x_0; \tag{3.5}$$

thus,

$$E\|E^{\mathcal{F}_0}x_t\| = E\|\Phi_{t,0}E^{\mathcal{F}_0}x_0\|, \tag{3.6}$$

yielding

$$E\|E^{\mathcal{F}_0}x_t\| \leqq \|\Phi_{t,0}\| E\|E^{\mathcal{F}_0}x_0\| \leqq \|\Phi_{t,0}\| EE^{\mathcal{F}_0}\|x_0\| = \|\Phi_{t,0}\| E\|x_0\|, \tag{3.7}$$

where the first inequality follows from (3.6) and (3.2) and the second follows from the convexity of the norm and from Jensen's inequality (see, e.g., [4 p. 76] for the vector case). It now follows from (3.7) and from Lemma 2.1 that

$$E\|e^x_{t,s}\| \leqq \|\Phi_{t,0}\| E\|E^{\mathcal{F}_0}x_0\| \leqq \|\Phi_{t,0}\| E\|x_0\|, \tag{3.8}$$

completing the proof.

Equation (3.3) presents two bounds for the truncation error. Naturally, the computation of the tighter bound is more laborious, but it is performed by a rather simple procedure, as shown in the sequel.

A case of particular interest is where $\mathcal{F}_t$ is generated by the process

$$y_t = C_t x_t + v_t \tag{3.9}$$

where x_t is given by (3.1), C_t is a time varying linear transformation and v_t is an uncorrelated Gaussian process, uncorrelated with x_t.

THEOREM 3.1.2. *Let* x_t *be given by* (3.1) *and let* y_t *be given by* (3.9). *Then the truncation error is bounded as*

$$e^y_{t,s} \equiv E^{\mathcal{F}_s}y_t - E^{\mathcal{B}_s}y_t, \qquad s<t; \tag{3.10}$$

$$\|e^y_{t,s}\| \leqq \|C_t\| \|\Phi_{t,0}\| E\|E^{\mathcal{F}_0}x_0\| \leqq \|C_t\| \|\Phi_{t,0}\| E\|x_0\|. \tag{3.11}$$

Proof. We have, almost surely,

$$e^y_{t,s} = C_t e^x_{t,s}. \tag{3.12}$$

Hence

$$\|e^y_{t,s}\| \leqq \|C_t\| \|e^x_{t,s}\|. \tag{3.13}$$

The proof then follows immediately from Theorem 3.1.1.

Suppose that it is desired to estimate present or predict future values of the process (x_t) given by (3.1), using values of the process (y_t), given by (3.9). Suppose that the data record $(y_p, 0 > r \leqq p \leqq t > 0)$ is truncated, so that only the portion $(y_p, 0 \leqq p \leqq t)$ is used.

Defining $\tilde{x}_t = [E^{\mathcal{F}_t}x_t^T, x_t^T]^T$ and $\psi_t = E\{\tilde{x}_t, \tilde{x}_t^T\}$, we have by Jensen's inequality

$$E\|E^{\mathcal{F}_0}x_0\| \leqq [\operatorname{tr}\{[I, 0]\psi_0[I, 0]^T\}]^{1/2} \tag{3.14}$$

where ψ_0 is obtained by integrating the Lyapunov equation (e.g., [5, p. 76])

$$\frac{d}{d\tau}\psi_\tau = F_\tau\psi_\tau + \psi_\tau F_\tau^T + G_\tau\Gamma_\tau G_\tau^T, \qquad r < \tau \leqq 0, \tag{3.15}$$

initialized at

$$\psi_r = \begin{bmatrix} P_r & 0 \\ 0 & P_r \end{bmatrix}, \tag{3.16}$$

where $P_r = \operatorname{cov}(x_r)$. Also in (3.15),

$$F_\tau \equiv \begin{bmatrix} A_\tau - K_\tau C_\tau & K_\tau C_\tau \\ 0 & K_\tau C_\tau \end{bmatrix}, \qquad G_\tau \equiv \begin{bmatrix} K_\tau & 0 \\ 0 & B_\tau \end{bmatrix}, \qquad \Gamma_\tau = \begin{bmatrix} R_\tau & 0 \\ 0 & Q_\tau \end{bmatrix}, \tag{3.17}$$

where $K_\tau \equiv P_\tau C_\tau^T R_\tau^{-1}$, $R_\tau = \operatorname{cov}(v_\tau)$, $Q_\tau = \operatorname{cov}(\theta_\tau)$, where θ_τ is the white noise process associated with $w_\tau \cdot P_\tau = E\{(E^{F_\tau}x_\tau - x_\tau)(E^{F_\tau}x_\tau - x_\tau)^T\}$ is obtained by integrating the Riccati equation (e.g., [5, p. 122]),

$$\frac{d}{d\tau}P_\tau = A_\tau P_\tau + P_\tau A_\tau^T + B_\tau Q_\tau B_\tau^T - P_\tau C_\tau^T R_\tau^{-1} C_\tau P_\tau \tag{3.18}$$

initialized at P_r. The first bound in (3.3) is then obtained as

$$E\|e_{t,s}^x\| \leqq \|\Phi_{t,0}\|[\operatorname{tr}\{[I \;\; 0]\psi_0[I \;\; 0]^T\}]^{1/2}. \tag{3.19}$$

The second bound is obtained as

$$E\|e_{t,s}^x\| \leqq \|\Phi_{t,0}\|[\operatorname{tr}\{\operatorname{cov}(x_0)\}]^{1/2}. \tag{3.20}$$

Equations (3.1) and (3.9) together constitute a linear system. Also of interest are the cases where x_t and y_t are discrete-time processes, and where x_t is a continuous-time process and y_t is a discrete-time process. In both cases we have truncation error bounds analogous to those shown above. The derivation is similar and will not be repeated here. Note that the integration of (3.15) is common in estimation and control practices.

3.2. Strong convergence of the truncation error. Convergence to zero of the truncation error has several implications, particularly in the context of asymptotic analysis of estimation procedures. Mean square and almost sure convergence of the truncation error for linear processes follow under a stability condition from the results of the previous section.

Recall that the process x_t given by (3.1) is said to be asymptotically stable if $\|\Phi_{t,0}\|$ is uniformly bounded and if

$$\lim_{t\to\infty} \|\Phi_{t,0}\| = 0. \tag{3.21}$$

THEOREM 3.2.1. *Let the process x_t given by* (3.1) *be asymptotically stable, let $E\|x_0\| < \infty$ and let the process y_t, generating $\mathcal{F}_t$ be Gaussian. Then the truncation error $e_{t,s}^x$ vanishes in the mean square. If, in addition, $t \in N_+ = (0, 1, \cdots)$ and if*

$$\sum_{t=1}^{\infty} \|\Phi_{t,0}\| < \infty, \tag{3.22}$$

then the truncation error vanishes also with probability 1.

Proof. The mean square convergence of $e^x_{t,s}$ follows from (3.21) and (3.3). To show convergence w.p.1. under (3.22), we note that (3.22) implies

$$\sum_{t=1}^{\infty} E\|e^x_{t,s}\| < \infty \tag{3.23}$$

which, by Chebyshev's inequality, implies that for any $\varepsilon > 0$,

$$\sum_{t=1}^{\infty} P\{\|e^x_{t,s}\| < \varepsilon\} < \infty. \tag{3.24}$$

Now since $e^x_{t,s}$ is an uncorrelated Gaussian, hence, independent sequence, we have by the Borel–Contelli lemma that for any $\varepsilon > 0$,

$$\lim \| e^x_{t,s}\| < \varepsilon \qquad \text{w.p.1;} \tag{3.25}$$

implying

$$\lim \| e^x_{t,s}\| = 0 \qquad \text{w.p.1,} \tag{3.26}$$

completing the proof.

THEOREM 3.2.2. *Let x_t, given by* (3.1) *be asymptotically stable with $E\|x_0\| < \infty$ and let the process y^t be given by* (3.9). *Suppose that $\|C_t\|$ is uniformly bounded for all $t \geqq 0$. Then the truncation error $e^y_{t,s}$ is strongly diminishing.*

Proof. The proof follows immediately from (3.13) and from Theorem 3.2.1.

Convergence results analogous to Theorems 3.2.1 and 3.2.2 readily follow for the cases where x_t or y_t are discrete time processes.

4. Discussion. It is well known that in the case of linear Gaussian processes the error covariance of the optimal estimator will converge under certain conditions (see [6], [7]) to a finite limit value. However, this convergence only implies that the estimates based on the truncated data will converge to the estimates based on the entire data in distribution. This weak form of convergence proves to be insufficient for inferring on properties of the one process from those of the other, while the strong convergence shown above implies that the two processes will almost surely possess the same probabilistic properties in the limit. Rissanen and Caines [8], [9] proved the consistency of maximum likelihood estimates of the parameters of stationary, Gaussian, autoregressive, moving average sequences, dispensing with the common but nonrealistic assumption that the estimator is operating at steady state. Their result was facilitated by showing that the filtered observation process strongly converges to the stationary process obtained from a "steady state" estimator. A similar analysis may now be performed for state space models, employing the above convergence results. Note, however, that the strong convergence of the truncation error holds for more general situations, i.e., for time varying, discrete and continuous time systems.

The estimation error resulting from a change in the initial conditions of a linear system has been shown by Ljung and Kailath [10] to be

$$\hat{x}(t) - \hat{x}_0(t) = H(t, 0)[m - m_0 + (\pi - \pi_0)\pi^{-1}(\hat{x}_0(0|t) - m_0)],$$

where (m_0, π_0) and (m, π) are the two sets of initial conditions, $H(t, 0)$ is the transition matrix of the Kalman filter corresponding to (m, π), $\hat{x}_0(0|t)$ is the smoothed estimate and $\hat{x}(t)$ and $\hat{x}_0(t)$ are the state estimates corresponding to the initial conditions. This expression can be related to our bounds by taking $m = m_0 = 0$ and by taking π_0 to be the initial condition for the truncated record and π to be the estimation covariance of the full record at $t = 0$ (note however that in computing our second bound, in (3.3), neither

π, nor the length of the truncated portion of the data are required). When convergence of the truncation error is concerned, here we see explicitly that filter stability, i.e., system observability, is sufficient for the error to vanish in the mean square (and w.p.1, if (3.22) holds for $H(t, 0)$).

As commented previously, no relationship between the process x_t and the data y_t need be assumed in order to obtain the bound on the truncation error $e^x_{t,s}$ and its strong convergence to zero. The bound follows from the normality condition on x_t and y_t while the convergence follows from the stability of x_t. Note that linearity is not necessary in order to obtain the general results, and stability is only required in the sense that $E^{\mathcal{F}_s}x_t$ converges to zero. This, of course, may hold for nonlinear processes. In the linear case, however, the bound assumes particularly simple forms which are computable by simple and familiar procedures.

Acknowledgment. I thank Professor Peter E. Caines of Harvard University for calling my attention to the analysis in [8], which has motivated the present study. I also thank the reviewers for their very useful comments and constructive suggestions.

REFERENCES

[1] R. S. LIPSTER AND A. N. SHIRYAYEV, *Statistics of Random Processes, Part II*, Springer Verlag, New York, 1977.

[2] I. I. GIKHMAN AND A. V. SKOROKHOD, *Introduction to the Theory of Random Processes*, W. B. Saunders, Philadelphia, PA, 1969.

[3] L. A. ZADEH AND C. A. DESOER, *Linear System Theory*, McGraw-Hill, New York, 1963.

[4] T. S. FERGUSON, *Mathematical Statistics, A Decision Theoretic Approach*, Academic Press, New York, 1967.

[5] A. GELB, ed., *Applied Optimal Estimation*, M.I.T. Press, Cambridge, MA, 1974.

[6] R. E. KALMAN, *New methods in Wiener filtering theory*, Proc. First Symposium on Engineering Application of Random Function Theory and Probability, John Wiley, New York 1963, pp. 270–388.

[7] W. M. WONHAM, *On a matrix Riccati equation of stochastic control*, this journal, (1968), pp. 681–697.

[8] J. RISSANEN AND P. E. CAINES, *Consistency of maximum likelihood estimation for ARMA processes*, Ann. Statist, 7 (1979), pp. 297–314.

[9] P. E. CAINES AND J. RISSANEN, *Maximum likelihood estimation of parameters in multivariate Gaussian stochastic processes*, IEEE Trans. Information Theory, IT-20 (1974), pp. 102–104.

[10] L. LJUNG AND T. KAILATH, *Efficient change of initial conditions, dual Chandrasekhar equations and some applications*, IEEE Trans. Automatic Control, AC-22, (1977), pp. 443–447.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 4, July 1980

0363-0129/80/1804-0003 $01.00/0

ON THE SYNTHESIS OF A STABILIZING FEEDBACK CONTROL VIA LIE ALGEBRAIC METHODS*

H. HERMES†

Abstract. Let the n-dimensional system of differential equations $dx/dt = X(x(t))$ have $p \in \mathbb{R}^n$ as a rest solution, i.e., $X(p) = 0$. Even in cases when this rest solution is unstable, one can often induce a strong stability (asymptotic stability) by the inclusion of one or more controls, e.g., via a controlled system (a) $dx/dt = X(x) + uY(x)$, where, say, $|u| \leqq 1$. Lie theory gives a computable, sufficient, condition to determine when, by use of the control u, one can steer a full n-dimensional neighborhood of p to p by solutions of (a). This condition is assumed to hold. One prefers a feedback control, i.e., that $u = u(x)$. The main result in this paper is an algorithm which determines a "modified" stabilizing feedback control. Specifically, for given $\varepsilon > 0$, one measures the current state q and the algorithm determines $u(t; q)$, $0 \leqq t \leqq \varepsilon$, such that the solution $x(\cdot\,; u)$ of (a) initiating from q and corresponding to this control u, satisfies distance $|x(\varepsilon; u) - p| < |q - p|$. In fact, iterates are theoretically shown to converge to p. Numerical examples computed via a simple FORTRAN program are included. These substantiate the strong stability achieved via such a modified feedback control.

Introduction. We consider the control system, on $\mathbb{R}^n$,

$$\dot{x}(t) = X(x) + uY(x), \tag{1}$$

where X, Y are smooth (actually C^n suffices) vector fields, $p \in \mathbb{R}^n$ is a rest solution of the uncontrolled system (i.e., $X(p) = 0$) and the control u may assume values in $[-1, 1]$. Let (adX, Y) denote the Lie product $[X, Y]$, inductively $(ad^kX, Y) = [X, (ad^{k-1}X, Y)]$; $\mathcal{S}^1 = \{(ad^jX, Y): j = 0, 1, 2, \cdots\}$ and $\mathcal{S}^1(p)$ denotes the elements of $\mathcal{S}^1$ evaluated at p. It is well-known (see [1], [2]) that dim span $\mathcal{S}^1(p) = n$ is a sufficient (but not necessary) condition that all points in some neighborhood of p can be attained by solutions of (1) initiating from p. Since reversing time merely introduces a minus sign in certain elements of $\mathcal{S}^1(p)$ but does not change the dimension of its span, it follows that dim span $\mathcal{S}^1(p) = n$ is a sufficient condition that all points in some neighborhood of p can be controlled (steered) to p in finite time. This is, therefore, a sufficient condition for "controlled stability" of the rest solution p of the uncontrolled system, even though p may not be a stable solution of $\dot{x} = X(x)$.

In practice one is not only interested in knowing that the ability to steer from a neighborhood of p to p is possible, but what is needed (preferred) is a method to construct a feedback control to do the task. One would like to merely measure the state x and have an algorithm which then provides the value $u(x)$. Our method, here, is not quite this. In applications one rarely has the ability to measure the state continuously; hence, we assume that measurements are made at time intervals of length $\varepsilon_1 > 0$. Suppose the first measured state is q^1. We give a constructive algorithm to generate a control $u(t; q^1)$, $0 \leqq t \leqq \varepsilon_1$. Let q^2 denote the solution of (1) at time $t = \varepsilon_1$ which initiates from q^1 at time $t = 0$ and corresponds to this control $u(\cdot\,; q^1)$. The algorithm may then be used to generate a new control $u(t, q^2)$, $0 \leqq t \leqq \varepsilon_1$, etc. The controls generated in this manner are such that if $\varepsilon_1 > 0$ is sufficiently small and the initial point q^1 sufficiently near p, then the sequence $q^1, q^2, \cdots$ converges to p. In this sense, the algorithm may be considered as generating a stabilizing feedback control.

For W a vector field on $\mathbb{R}^n$, we let $(\exp tW)(p)$ denote the solution, at time t, of $\dot{x}(t) = W(x(t))$, $x(0) = p$. The algorithm is based on the fact that by composing at most 2^k maps of the form $(\exp \varepsilon(X + u_i(\varepsilon)Y))p$, one can form maps, denoted $q_k^{\pm}(\varepsilon)p$, such that

* Received by the editors March 19, 1979, and in final revised form October 19, 1979.

† Department of Mathematics, University of Colorado, Boulder, Colorado 80309. This research was supported in part by the National Science Foundation under Grant MCS 76-04419 AO1.

$q_k^{\pm}(0)p = p$ and $d/d\varepsilon\, q_k^{\pm}(\varepsilon)p|_{\varepsilon=0} = \pm(ad^k X, Y)(p)$. Furthermore, the maps $q_k^{\pm}(\varepsilon)p$ correspond to admissible trajectories of system (1) resulting from controls with at most $2^k - 1$ switches. Now let $(\alpha_1, \cdots, \alpha_n) \in \mathbb{R}^n$ and consider the composition

$$q(\varepsilon)p = q_{n-1}^{e_{n-1}}(|\alpha_n|\varepsilon) \circ \cdots \circ q_0^{e_0}(|\alpha_1|\varepsilon)p,$$

where

$$e_i = \begin{cases} + & \text{if } \alpha_i \geqq 0, \\ - & \text{if } \alpha_i < 0. \end{cases}$$

Then for $\varepsilon \geqq 0$, $q(\varepsilon)p$ corresponds to a trajectory of (1) resulting from a control with at most $\sum_{i=0}^{n-1}(2^i - 1) = 2^n - n$ switches and $d/d\varepsilon\, q(\varepsilon)p|_{\varepsilon=0} = \sum_{i=1}^{n} \alpha_i (ad^{i-1}X, Y)(p)$. In particular, if span $\mathcal{S}^1(p) = n$ and we are at a point q^1 near p with $q^1 - p = \sum_{i=1}^{n} \alpha_i (ad^{i-1}X, Y)(p)$, one can prescribe a control over the interval $[0, \varepsilon_1]$, having at most $2^n - n$ switches, and which drives the solution in the direction $-\sum_{i=1}^{n} \alpha_i (ad^{i-1}X, Y)(p)$. This is the essence of the idea behind the algorithm.

It should be remarked that methods for "piecing together" trajectories of system (1) corresponding to controls having values ±1 or 0 to generate solutions, whose derivatives with respect to variation parameters are $(ad^k X, Y)(p)$, are by no means unique. An elegant approach to do this (which however, requires a modification of the formula on p. 271 to be correct) may be found in [3]. The method of [2] is theoretically simple and useful, but seemingly does not lend itself as well to numerical computations as the method used here.

While the theory insures convergence of the sequence $q^1, q^2, \cdots$ to p, numerical computations based on the algorithm produce a sequence $q^1, Q^2, Q^3, \cdots$ which does not always behave as the theoretical sequence. In the last section, we give numerical results for several examples which were computed via a simple FORTRAN program based on this algorithm. With some flexibility allowed in the control bounds and time of integration, i.e., ε_1, the results were excellent.

For the sake of exposition, we have chosen to present both the theory and numerical results for systems of the form (1). Very little effort is necessary to extend these first-order methods to systems of the form $\dot{x} = X(x) + \sum_{i=1}^{m} u_i Y^i$ as studied in [1], [2]. The basic idea also can be used to compute "stabilizing feedback controls" when first-order tests fail but higher order tests yield sufficient conditions for local controllability; see [3], [4], [5]. The numerical algorithm can be improved, for example, by compensating for the "drift" caused by the vector field X, by solving for the coefficients α_i which determine the direction to the rest point in terms of the basis $(ad^j X, Y)(q)$, $j = 0, \cdots, n-1$, rather than $(ad^j X, Y)(p)$, etc. These technical points for numerical procedures have been carried out, but omitted in this paper which presents the theory.

1. Basic theory. We consider the n-dimensional control system (1) on $\mathbb{R}^n$. One could, instead, consider X, Y as tangent vector fields on an n-manifold M but since our goal is a local theory, $\mathbb{R}^n$ suffices.

Let A denote the Jacobian matrix of partial derivatives of X with respect to x, evaluated at p, i.e., $A = X_x(p)$. Then since we assume $X(p) = 0$, $(ad^j X, Y)(p) = A^j Y(p)$, and from the linear theory we have

LEMMA 1. *Let $X(p) = 0$ and $\mathcal{S}^1(p) = \{(ad^\nu X, Y)(p) : \nu = 0, 1, \cdots\}$. Then if* dim span $\mathcal{S}^1(p) = k \leqq n$, *$Y(p)$, $(adX, Y)(p), \cdots, (ad^{k-1}X, Y)(p)$ are linearly independent.*

In more familiar language, the assumption dim span $\mathcal{S}^1(p) = n$ is equivalent to rank $\{Y(p), AY(p), \cdots, A^{n-1}Y(p)\} = n$, where $A = X_x(p)$; i.e., that the linearized system associated with (1) is controllable, and hence, system (1) is locally controllable at p.

Lemma 1 merely shows that with our assumptions, we need only consider the first n elements, i.e., $(ad^\nu X, Y)(p)$, $\nu = 0, \cdots, n-1$, of $\mathscr{S}^1(p)$. Loosely speaking, what we next show is how to obtain these "directions" via compositions of solutions of (1).

The ensuing computations are based on the Campbell–Baker–Hausdorff formula. Specifically, let $V = \varepsilon X - \varepsilon^k Y$ and $W = \varepsilon X + \varepsilon^k Y$. Then (see [6, pp. 116–118]) $(\exp V) \circ (\exp W)(p) = (\exp \sum_{m=1}^{\infty} c_m(V, W))(p)$, where $c_1 = V + W = 2\varepsilon X$, $c_2 = \varepsilon^k[X, Y]$, $c_3 = \frac{1}{3}\varepsilon^{2k+1}(ad^2 Y, X)$ and $c_4 = -\frac{1}{24}\varepsilon^{k+3}(ad^3 X, Y) - \frac{1}{24}\varepsilon^{3k+1}(ad^3 Y, X)$, etc.

Throughout, $\varepsilon \geqq 0$ and k, m are nonnegative integers. We define

$$q_0^+(\varepsilon, k)p = (\exp \varepsilon(X + \varepsilon^k Y))p, \qquad q_0^-(\varepsilon, k)p = (\exp \varepsilon(X - \varepsilon^k Y))p, \tag{2}$$

which we can interpret either as the solution at time 1, of $\dot{x} = \varepsilon(X + \varepsilon^k Y)$, or the solution, at time ε, of $\dot{x} = X + \varepsilon^k Y$, $x(0) = p$. Then

$$d/d\varepsilon\, q_0^+(\varepsilon, 0)p|_{\varepsilon=0} = Y(p), \qquad d/d\varepsilon\, q_0^-(\varepsilon, 0)p|_{\varepsilon=0} = -Y(p).$$

Next, we define (an exception to the general definition)

$$\begin{aligned} q_1^+(\varepsilon, k)p &= q_0^-(\varepsilon, k-1) \circ q_0^+(\varepsilon, k-1)p \\ &= \exp(\varepsilon(X - \varepsilon^{k-1}Y)) \circ \exp(\varepsilon(X + \varepsilon^{k-1}Y))p \\ &= \exp(2\varepsilon X + \varepsilon^{k+1}[X, Y] + \tfrac{1}{3}\varepsilon^{2k+1}(ad^2 Y, X) - \tfrac{1}{24}\varepsilon^{k+3}(ad^3 X, Y) \\ &\qquad - (\tfrac{1}{24})\varepsilon^{3k+1}(ad^3 Y, X) + o(\varepsilon^{k+3}))p \end{aligned} \tag{i}$$

In particular

$$\begin{aligned} &q_1^+(\varepsilon, 1)p = \exp(2\varepsilon X + \varepsilon^2[X, Y] + o(\varepsilon^3))p, \\ &d/d\varepsilon\, q_1^+(\varepsilon, 1)p|_{\varepsilon=0} = 0, \\ &d^2/d\varepsilon^2\, q_1^+(\varepsilon, 1)p|_{\varepsilon=0} = 2[X, Y](p). \end{aligned}$$

Similarly, we define

$$q_1^-(\varepsilon, k)p = q_0^+(\varepsilon, k-1) \circ q^-(\varepsilon, k-1)p,$$

and note that

$$d/d\varepsilon\, q_1^-(\varepsilon, 1)p|_{\varepsilon=0} = 0, \qquad d^2/d\varepsilon^2\, q_1^-(\varepsilon, 1)p|_{\varepsilon=0} = -2[X, Y](p).$$

For the moment, we will continue by generating only positive multiples of $(ad^m X, Y)(p)$; i.e., we deal only with $q_m^+(\varepsilon, m)$. Define

$$\begin{aligned} q_2^+(\varepsilon, k)p &= q_1^-(\varepsilon, k) \circ q_1^+(\varepsilon, k)p \\ &= \exp(\varepsilon(X + \varepsilon^{k-1}Y)) \circ \exp(2\varepsilon(X - \varepsilon^{k-1}Y)) \circ \exp(\varepsilon(X + \varepsilon^{k-1}Y))p \\ &= \exp(\varepsilon(X + \varepsilon^{k-1}Y)) \circ \exp(\varepsilon(X - \varepsilon^{k-1}Y)) \\ &\quad \circ \exp(\varepsilon(X - \varepsilon^{k-1}Y)) \circ \exp(\varepsilon(X + \varepsilon^{k-1}Y))p \\ &= \exp(2\varepsilon X - \varepsilon^{k+1}[X, Y] + (\varepsilon^{2k+1}/3)(ad^2 Y, X) \\ &\qquad + (\tfrac{1}{24})\varepsilon^{k+3}(ad^3 X, Y) + o(\varepsilon^M)) \circ \exp(2\varepsilon X + \varepsilon^{k+1}[X, Y] \\ &\qquad + (\varepsilon^{2k+1}/3)(ad^2 Y, X) - \tfrac{1}{24}\varepsilon^{k+3}(ad^3 X, Y) + o(\varepsilon^M))p \\ &= \exp(2^2\varepsilon X + 2\varepsilon^{k+2}(ad^2 X, Y) + \tfrac{2}{3}\varepsilon^{2k+1}(ad^2 Y, X) + o(\varepsilon^{2k+1}))p, \end{aligned} \tag{ii}$$

where $M = \max(2k+1), (k+3)$. In particular,

$$q_2^+(\varepsilon, 2)p = \exp(2^2\varepsilon X + 2\varepsilon^4(ad^2X, Y) + o(\varepsilon^4))p.$$

Remark 1. One notes that $q_1^-(\varepsilon, k)p$ is formed from $q_1^+(\varepsilon, k)p$, replacing Y with $-Y$. Thus each term in $q_1^-(\varepsilon, k)p$, which contains one factor Y (e.g., the terms $-\varepsilon^k[X, Y], \frac{1}{24}\varepsilon^{k+1}(ad^3X, Y)$, etc.) are "cancelled" by terms $\varepsilon^k[X, Y]$, $-\frac{1}{24}\varepsilon^{k+1}(ad^3X, Y)$ in $q_1^+(\varepsilon, k)p$, when the composition is formed to define $q_2^+(\varepsilon, k)p$. The terms with two (or an even number of) factors Y do not cancel, but instead sum, e.g., the terms $\frac{1}{3}\varepsilon^{2k-1}(ad^2Y, X)$, etc. The effect of these can be made small by increasing k, due to the exponent $2k$. This pattern will continue when we define $q_m^+(\varepsilon, k)$ and is essential in keeping track of the order of "remainders".

From (ii), we note that

$$d^j/d\varepsilon^j\, q_2^+(\varepsilon, 2)p|_{\varepsilon=0} = \begin{cases} 0 & \text{if } 1 \leqq j \leqq 3, \\ 2(4!)(ad^2X, Y)(p) & \text{if } j = 4. \end{cases}$$

Again, $q_2^-(\varepsilon, k)p = q_1^+(\varepsilon, k) \circ q_1^-(\varepsilon, k)$.

Next

$$q_3^+(\varepsilon, k)p = q_2^-(\varepsilon, k) \circ q_2^+(\varepsilon, k)p,$$

$$\text{(iii)} \qquad q_3^+(\varepsilon, 3)p = \exp(2^3\varepsilon X + 2^3\varepsilon^6(ad^3X, Y) + o(\varepsilon^6))p,$$

$$d^j/d\varepsilon^j\, q_3^+(\varepsilon, 3)p = \begin{cases} 0 & \text{if } 1 \leqq j \leqq 5, \\ 2^3(6!)(ad^3X, Y)(p) & \text{if } j = 6. \end{cases}$$

The inductive definitions of $q_m^{\pm}(\varepsilon, k)p$ should now be evident. We state the result as

LEMMA 2. *Let $q_0^{\pm}(\varepsilon, k)p, q_1^{\pm}(\varepsilon, k)$ be defined as in* (2) *and* (i) *above, and inductively, for $m = 2, \cdots$,*

$$q_m^{\pm}(\varepsilon, k)p = q_{m-1}^{\mp}(\varepsilon, k) \circ q_{m-1}^{\pm}(\varepsilon, k)p.$$

Then

$$\text{(3)} \qquad q_m^{\pm}(\varepsilon, m)p = \exp(2^m\varepsilon X \pm a_m\varepsilon^{2m}(ad^mX, Y) + o(\varepsilon^{2m}))p,$$

where $a_0 = 1$, $a_m = 2^{m-1}a_{m-1}$, so

$$a_m = 2^{m(m-1)/2}$$

and

$$\text{(4)} \qquad d^j/d\varepsilon^j\, q_m^{\pm}(\varepsilon, m)p|_{\varepsilon=0} = \begin{cases} 0 & \textit{if } 1 \leqq j \leqq 2m-1, \\ \pm a_m(2m)!(ad^mX, Y)(p) & \textit{if } j = 2m. \end{cases}$$

Remark 2. The bound $2^k - 1$ on the number of control switches needed to produce $q_k^{\pm}(\varepsilon, k)p$, and hence generate $\pm(ad^kX, Y)(p)$, is crude. Indeed, one may note that in defining $q_2^+(\varepsilon, k)p$ (see (ii)) the second exponential in $q_1^-(\varepsilon, k)p$ is the same as the first in $q_1^+(\varepsilon, k)p$, and hence, these combine to give $\exp(2\varepsilon(X - \varepsilon^{k-1}Y))$. Thus $q_2^{\pm}(\varepsilon, 2)p$ requires only two control switches.

Remark 3. For illustration, we again refer to $q_2^+(\varepsilon, k)p$. The reason one chooses $k = 2$ (i.e., $q_2^+(\varepsilon, 2)p$) is clearly to make sure the order of the coefficient $\frac{2}{3}\varepsilon^{2k+1}$ of (ad^2Y, X) is of higher order than the coefficient $2\varepsilon^{k+2}$ of (ad^2X, Y). If, however, we were to deal with an example (as we later shall) in which $(ad^2Y, X)(p) = 0$, we could generate $(ad^2X, Y)(p)$ from $q_2^+(\varepsilon, 1)p$.

It is preferable, for our purposes, to obtain the directions $\pm(ad^jX, Y)(p), j = 0, \cdots, (n-1)$ as first derivatives. This is readily accomplished by reparametrization.

COROLLARY 1. *Let* $q_0^{\pm}(\varepsilon)p = q_0^{\pm}(\varepsilon, 0)p$ *and for* $m = 1, 2, \cdots,$

$$q_m^{\pm}(\varepsilon)p = q_m^{\pm}((\varepsilon/a_m)^{1/2m}, m)p. \tag{5}$$

Then

$$d/d\varepsilon\, q_m^{\pm}(\varepsilon)p|_{\varepsilon=0} = \pm(ad^mX, Y)(p).$$

Proof. $q_m^{\pm}(a_m\varepsilon^{2m})p \equiv q_m^{\pm}(\varepsilon, m)p$, hence, by differentiating both sides of this identity $2m$ times, using (4), and letting $q^{\pm\prime}(\varepsilon)p$ denote the first derivative of $q^{\pm}(\varepsilon)p$ with respect to its argument, we obtain

$$a_m(2m)!q_m^{\pm\prime}(a_m\varepsilon^{2m})p|_{\varepsilon=0} = d^{2m}/d\varepsilon^{2m}q_m^{+}(\varepsilon, m)p|_{\varepsilon=0} = a_m(2m)!(ad^mX, Y)(p).$$

It follows that $q_m^{\pm}(0)p = \pm(ad^mX, Y)(p)$. □

For computing purposes, it is useful to have an explicit list of the first several functions $q_m^{\pm}(\varepsilon)p$. These are

$$\begin{aligned} q_0^{\pm}(\varepsilon)p &= \exp(\varepsilon(X \pm Y))p,\\ q_1^{\pm}(\varepsilon)p &= (\exp\sqrt{\varepsilon}(X \mp Y)) \circ (\exp\sqrt{\varepsilon}(X \pm Y))p,\\ q_2^{\pm}(\varepsilon)p &= \exp((\varepsilon/2)^{1/4}(X \pm (\varepsilon/2)^{1/4}Y)) \circ \exp(2(\varepsilon/2)^{1/4}(X \mp (\varepsilon/2)^{1/4}Y))\\ &\qquad \circ \exp((\varepsilon/2)^{1/4}(X \pm (\varepsilon/2)^{1/4}Y))p,\\ q_3^{\pm}(\varepsilon)p &= q_2^{\mp}((\varepsilon/8)^{1/6}, 3) \circ q_2^{\pm}((\varepsilon/8)^{1/6}, 3)p, \end{aligned}$$

which can now easily be constructed from $q_2^{\pm}(\varepsilon, 3)p$.

The maps $q_m^{\pm}(\varepsilon)p$ have been defined for $\varepsilon \geqq 0$. Since $d/d\varepsilon\, q_m^{\pm}(\varepsilon)p|_{\varepsilon=0} = \pm(ad^mX, Y)(p)$, if we define

$$q_m(\varepsilon)p = \begin{cases} q_m^{+}(\varepsilon)p & \text{if } \varepsilon \geqq 0,\\ q_m^{-}(|\varepsilon|)p & \text{if } \varepsilon < 0, \end{cases} \tag{6}$$

$m = 0, \cdots, n-1$, then $q_m(\cdot)p$ is defined, and continuously differentiable, for ε in a neighborhood of zero. Let $\mathcal{U}$ be a neighborhood of zero in $\mathbb{R}^n$. Then the map $(\alpha_1, \cdots, \alpha_n) \to q_{n-1}(\alpha_n) \circ \cdots \circ q_1(\alpha_2) \circ q_0(\alpha_1)p$ of $\mathcal{U}$ into $\mathbb{R}^n$ takes zero to p and has the Jacobian $n \times n$ matrix with columns $(ad^jX, Y)(p), j = 0, \cdots, (n-1)$. The assumption that $\dim \operatorname{span} \mathcal{S}^1(\mathrm{p}) = n$, together with Lemma 1, then implies this map covers a neighborhood of p. Furthermore, since $q_m^{\pm}(\varepsilon)p$, $\varepsilon \geqq o$ is the endpoint of an admissible trajectory of (1) corresponding to a control with at most $2^m - 1$ switches, the composition $q_{n-1}(\alpha_n) \circ \cdots \circ q_0(\alpha_1)p$ is the endpoint of an admissible trajectory of (1) corresponding to a control with at most $\Sigma_{i=0}^{n-1}(2^i - 1) = 2^n - n$ switches. We summarize this as

THEOREM 1. *Consider the system* $\dot{x} = X(x) + u(t)Y(x)$, $x(0) = p$ *on* $\mathbb{R}^n$ *with* $X(p) = 0$, $|u(t)| \leqq 1$ *and* $\dim \operatorname{span} \mathcal{S}^1(p) = n$. *Let* $q_m^{\pm}(\varepsilon)p$ *and* $q_m(\varepsilon)p$ *be defined as in* (5), (6), *above. Then every point in some neighborhood of p can be attained from p by a trajectory of* (1) *of the form* $q_{n-1}(\alpha_n) \circ \cdots \circ q_1(\alpha_2) \circ q_0(\alpha_1)p$ *which corresponds to a control having at most* $2^n - n$ *switches. Similarly, p can be attained from any point* p^1 *in some neighborhood of p by a trajectory of* (1) *of the form* $q_{n-1}(\alpha_n) \circ \cdots \circ q_0(\alpha_1)p^1$.

2. The feedback control algorithm. Again our assumptions for the system (1) are that $X(p) = 0$ and $\dim \operatorname{span} \mathcal{S}^1(p) = n$. For any $|\varepsilon| > 0$, let $\mathcal{U}(p, \varepsilon)$ denote a disc centered at p such that $x \in \mathcal{U}(p, \varepsilon)$ implies

$$|X(x)| \leqq \varepsilon, \qquad |(ad^jX, Y)(x) - (ad^jX, Y)(p)| \leqq \varepsilon, \quad j = 0, \cdots, (n-1). \tag{7}$$

ALGORITHM. Given $q^j \in \mathcal{U}(p, \varepsilon)$, express $q^j - p$ as a linear combination of $Y(p), \cdots, (ad^{n-1}X, Y)(p)$, i.e.,

$$\text{(a)} \qquad q^j - p = \textstyle\sum_{i=1}^n \alpha_i(q^j)(ad^{i-1}X, Y)(p)$$

and define

$$\text{(b)} \qquad q^{j+1} = q_{n-1}(-\alpha_n(q^j)\varepsilon) \circ \cdots \circ q_0(-\alpha_1(q^j)\varepsilon)q^j.$$

We will show that q^{j+1} again belongs to $\mathcal{U}(p, \varepsilon)$, and for $q^1 \in \mathcal{U}(p, \varepsilon)$ with $|\varepsilon| > 0$ and sufficiently small, the sequence $q^1, q^2, \cdots$ generated by the algorithm converges to p.

To gain insight into why the algorithm can be expected to work, we note that

$$q_{n-1}(-\alpha_n(q^1)\varepsilon) \circ \cdots \circ q_0(-\alpha_1(q^1)\varepsilon)q^1 - q^1 = -\varepsilon \sum_{i=1}^n \alpha_i(q^1)(ad^{i-1}X, Y)(q^1) + o(\varepsilon),$$

a direction which is "nearly" the negative of $q^1 - p$. The nearly is because $(ad^iX, Y)(q^1)$ and $(ad^iX, Y)(p)$ are close but not necessarily equal for $q^1 \in \mathcal{U}(\varepsilon, p)$.

THEOREM 2. *Let $\mathcal{U}(p, \varepsilon)$ be as above, $q^1 \in \mathcal{U}(p, \varepsilon)$ and $q^2, q^3, \cdots$ be the sequence generated by the steps* (a), (b) *of the above algorithm. Then for $\varepsilon > 0$ and sufficiently small, $q^i \to p$ as $i \to \infty$.*

Proof. For $x \in \mathbb{R}^n$, define

$$f(\varepsilon, x) = q_{n-1}(-\alpha_n(x)\varepsilon) \circ \cdots \circ q_0(-\alpha_1(x)\varepsilon)x,$$

where, as before,

$$x - p = \sum_{i=1}^n \alpha_i(x)(ad^{i-1}X, Y)(p).$$

Then

$$\text{(i)} \qquad f(0, x) = x,$$

$$\text{(ii)} \qquad f(\varepsilon, p) = p \quad \text{since } \alpha_i(p) = 0, i = 1, \cdots, n.$$

Now consider the expansion of f in a Taylor series, with remainder, about the point $(0, p)$, i.e.,

$$f(\varepsilon, x) = f(0, p) + \frac{\partial f(0, p)}{\partial \varepsilon}\varepsilon + \frac{\partial f(0, p)}{\partial x}(x - p) + \frac{\partial^2 f(0, p)}{\partial \varepsilon^2}\varepsilon^2$$
$$+ \varepsilon \frac{\partial^2 f(0, p)}{\partial \varepsilon \partial x} \cdot (x - p) + \text{higher order terms.}$$

From (ii), above, we see $\partial^k f(0, p)/\partial \varepsilon^k = 0$, $k = 1, 2, \cdots$ which implies all nonzero terms in the remainder (higher order terms) are of the order $|\varepsilon|^k |x - p|^m$ with $k, m \geqq 2$. Thus since $\partial f(0, p)/\partial x = id$, from (i),

$$(8) \qquad f(\varepsilon, x) = p + (x - p) + (x - p) + \varepsilon \frac{\partial^2 f(0, p)}{\partial \varepsilon\, \partial x} \cdot (x - p) + o(|\varepsilon|\, |x - p|).$$

We next wish to obtain a sharper estimate which includes $\partial^2 f(0, p)/\partial \varepsilon\, \partial x$ explicitly calculated.

From (3) and (5), we obtain the estimate (with $\varepsilon \geqq 0$)

$$q_k(-\alpha_{k+1}(x)\varepsilon)x = \left(\exp\left(2k\left(\frac{\varepsilon|\alpha_{k+1}(x)|}{a_k}\right)^{1/2k} X - \varepsilon\alpha_{k+1}(x)(ad^kX, Y) + o(\varepsilon)\right)\right)x.$$

Thus

$$\begin{aligned}|f(\varepsilon, x)-p| &= \Big|x-\varepsilon \sum_{k=1}^{n} \alpha_k(x)(ad^{k-1}X, Y)(x)+\sum_{k=0}^{n-1} 2^k\left(\frac{\varepsilon|\alpha_{k+1}(x)|}{a_k}\right)^{1/2k} X(x)\\ &\quad+\varepsilon \sum_{k=1}^{n} \alpha_k(x)(ad^{k-1}X, Y)(p)-\varepsilon \sum_{k=1}^{n} \alpha_k(x)(ad^{k-1}X, Y)(p)-p+o(\varepsilon)\Big|\\ &\leqq (1-\varepsilon)|x-p|+\varepsilon\left|\sum_{k=1}^{n} \alpha_k(x)((ad^{k-1}X, Y)(p)-(ad^{k-1}X, Y)(x))\right| \qquad (9)\\ &\quad+\left|\sum_{k=0}^{n-1} 2^k\left(\frac{\varepsilon|\alpha_{k+1}(x)|}{a_k}\right)^{1/2k} X(x)+o(\varepsilon)\right|\\ &\leqq (1-\varepsilon)|x-p|+o(\varepsilon) \quad \text{if } x \in \mathcal{U}(\varepsilon, p) \text{ by (7).}\end{aligned}$$

Equation (9) shows that the term $\varepsilon(\partial^2 f(0, p)/\partial\varepsilon\, \partial x)\cdot(x-p)$ in (8) is, in essence, actually $-\varepsilon(x-p)$. Using this in (8) yields

$$|f(\varepsilon, x)-p|=(1-\varepsilon)|x-p|+o(|\varepsilon|\,|x-p|) \tag{10}$$

if $x \in \mathcal{U}(\varepsilon, 'p)$. Now choose $|\varepsilon|>0$ and sufficiently small so that if $o \leqq \varepsilon_1 \leqq |\varepsilon|$, then

$$o(|\varepsilon_1|\,|x-p|) \leqq (\varepsilon_{1/2})|x-p|, \qquad x \in \mu(\varepsilon, p).$$

For such a choice of ε, and $x \in \mathcal{U}(\varepsilon, p)$,

$$|f(\varepsilon, x)-p| \leqq (1-\varepsilon/2)|x-p|. \tag{11}$$

Returning to the algorithm, and noting that $q^2=f(\varepsilon, q^1)$, etc., we see if $q^1 \in \mathcal{U}(\varepsilon, p)$, then $q^2 \in \mathcal{U}(\varepsilon, p)$ and

$$\begin{aligned}|q^2-p| &= |f(\varepsilon, q^1)-p| \leqq (1-\varepsilon/2)|q^1-p|,\\ |q^3-p| &= |f(\varepsilon, q^2)-p| \leqq (1-\varepsilon/2)|f(\varepsilon, q^1)-p| \leqq (1-\varepsilon/2)^2|q^1-p|,\end{aligned}$$

and inductively, $|q^m-p| \leqq (1-\varepsilon/2)^{m-1}|q^1-p|$ showing $q^m \to p$. □

3. Numerical results. This section will present the results of computations, using the feedback control algorithm, on two examples. The first is nonlinear and two dimensional; the second is nonlinear and three dimensional. In each case, the rest point p of the uncontrolled equation (i.e., of the vector field X) is taken to be the origin (as can always be accomplished via a change of coordinates). The symbol D^i denotes the Euclidean distance from q^i to the origin.

As the theory shows, for dimension $n \leqq 2$ we need only generate $Y(p)$ and $[X, Y](p)$, i.e., use $q_0^{\pm}(\varepsilon, 0)p$ and $q_1^{\pm}(\varepsilon, 1)p$. Both of these use only trajectories of the form $(\exp \varepsilon(X \pm Y))p$, i.e., no positive power of ε multiplies Y; hence, the control bounds (taken to be ± 1 initially) have significance. In dimension three, we must also generate $(ad^2X, Y)(p)$ by using $q_2^{\pm}(\varepsilon, 2)p$ which is composed of trajectories of the form $(\exp \varepsilon(X \pm \varepsilon Y))p$. Here, in general, the control bound is relatively insignificant in view of the factor ε multiplying Y. However, as indicated in Remark 3, if $(ad^2 Y, X)(p)=0$, we may generate $(ad^2X, Y)(p)$ by use of $q_2^{\pm}(\varepsilon, 1)p$ which is composed of trajectories of the form $(\exp \varepsilon(X \pm Y))p$. Our second example is of this form, hence the control bounds are significant. In both examples the control bounds initially began at ± 1, but the program automatically doubled the bound after any step in which $D^{i+1} \geqq D^i$. The time of integration (related directly to ε) was also modified whenever $D^{i+1} \geqq D^i$. The integration subroutine used was RKF45, by H. A. Watts and L. F. Shampine; see [7].

Example 1 (*Unstable pendulum.*) We take, as the equation of motion, $\ddot{x}+\sin x=u$, or as a first order system

$$\dot{x}_1=x_2,$$
$$\dot{x}_2=-\sin x_1+u.$$

We are interested in controlling so as to "stabilize" the above system in a neighborhood of the unstable rest solution $x_1(t)\equiv\pi$, $x_2(t)\equiv 0$; i.e., the pendulum points vertically upward. Letting $y_1=x_1-\pi$, $y_2=x_2$ to shift the rest solution to the origin gives the system

$$\dot{y}_1=y_2,$$
$$\dot{y}_2=\sin y_1+u$$

with, now, $p=(0,0)$ being the unstable rest-solution of the uncontrolled system.

For the sample run, shown in Table 1, the initial data given the program was $\varepsilon=.05$, $q^1=(0.1,-.007)$ with the computation to end after 20 steps (i.e., the computation of q^{20}) or if $D\leqq 0.0001$. The initial bounds on the control were ±4. To save space, we list here only the values q^i, D^i, control magnitude, and switching sequence to go from q^i to q^{i+1}.

TABLE 1

step	q coordinates	D	control bounds	switching sequence
1	(.1, −.007)	.1002	±4	4, −4, 4
2	(.08, .007)	.0804	±4	−4, −4, 4
3	(.06, .015)	.0671	±4	−4, −4, 4
4	(.05, .019)	.0573	±4	−4, −4, 4
5	(.045, .020)	.0495	±4	−4, −4, 4
6	(.038, .020)	.0429	±4	−4, −4, 4
7	(.032, .019)	.0372	±4	−4, −4, 4
8	(.026, .017)	.0322	±4	−4, −4, 4
9	(.022, .016)	.0277	±4	−4, −4, 4
10	(.019, .014)	.0237	±4	−4, −4, 4
11	(.015, .012)	.0202	±4	−4, −4, 4
12	(.013, .011)	.0172	±4	−4, −4, 4
13	(.011, .009)	.0145	±4	−4, −4, 4
14	(.009, .008)	.0122	±4	−4, −4, 4
15	(.007, .006)	.0102	±4	−4, −4, 4
16	(.006, .005)	.0085	±4	−4, −4, 4
17	(.005, .004)	.0071	±4	−4, −4, 4
18	(.004, .0039)	.0058	±4	−4, −4, 4
19	(.003, .0032)	.0048	±4	−4, −4, 4
20	(.0029, .0027)	.0040	±4	

As can be seen, the control switching sequence was such as to initially make the position and velocity agree in sign (for this initial data, both positive), and thereafter, both position and velocity decreased monotonically towards zero.

Example 2. This is a three dimensional, nonlinear system, of the form (1). For notational ease, all vectors will be written as row vectors. We take $X(x)=(\sin x_2, x_3, x_1x_2)$, $Y=(0,1,1)$. Computation shows $(adX, Y)(x)=(\cos x_2, 1, x_1)$, $(ad^2X, Y)(x)=(\cos x_2-x_3\sin x_3, x_1, x_2\cos x_2+x_1-\sin x_2)$ so $Y(p)$, $(adX, Y)(p)$, $(ad^2X, Y)(p)$ are linearly independent. Also, $(ad^2Y, X)(p)=0$.

In several runs, with the control bound fixed at 1 and initial value of D approximately 0.5, the sequence generated did not converge to zero (in fact divergence occurred). However, with the control bound allowed to automatically double when $D^{i+1} \geqq D^i$, rapid convergence occurred in all sample runs. A typical example (omitting switching times and switching sequences) is given in Table 2. The initial value of ε was 0.01 with the computation to stop if $D \leqq 10^{-5}$, which occurred.

TABLE 2

step	q coordinates	D	control bounds
1	(.06, .04, −.08)	.1077	±1
2	(.069, .006, −.081)	.1073	±1
3	(.067, −.017, −.082)	.1079	±1
4	(.059, −.032, −.080)	.1054	±2
5	(.043, −.056, −.078)	.1057	±2
6	(.006, −.076, −.067)	.1018	±4
7	(−.025, −.086, −.056)	.1058	±4
8	(−.025, −.038, −.021)	.0507	±8
9	(−.011, −.017, −.008)	.0221	±8
10	(−.004, −.007, −.003)	.0089	±8
11	(−.002, −.003, −.001)	.0034	±8
12	(−.0006, −.0010, −.0003)	.0012	±8
13	(−.0002, −.0003, −.0001)	.0004	±8
14	$(-9\times10^{-5}, -.0001, -5\times10^{-5})$	.00017	±8
15	$(-3\times10^{-5}, -5\times10^{-5}, -2\times10^{-5})$	6×10^{-5}	±8
16	$(-1\times10^{-5}, -2\times10^{-5}, -1\times10^{-5})$	2×10^{-5}	±8
17	$(0, -10^{-5}, 0)$	10^{-5}	±8

The print format contained 5 decimal accuracy; the relative error in the integration of the differential equations was kept at 10^{-8}. The above run used approximately $\frac{1}{2}$ second of central processing unit time on a CDC 6400, mainly on the 102 integrations of the three dimensional, nonlinear, system of ordinary differential equations.

Concluding Remarks. In two dimensions, and even three dimensions, this method of generating Y, (adX, Y) and (ad^2X, Y) is computationally feasible. For dimensions four, or more, difficulties in convergence have been experienced. This can be expected from the theory. Specifically, suppose that in (3), the terms $o(\varepsilon^{2m})$ began with elements in the Lie algebra generated by X and Y (denoted $L(X, Y)$) having coefficients with a factor ε^{2m+1}, i.e., suppose

$$q_m^{\pm}(\varepsilon, m)p = (\exp(2^m\varepsilon X \pm a_m\varepsilon^{2m}(ad^mX, Y) + \varepsilon^{2m+1}W))p$$

for some $W \in L(X, Y)$. Then, from (5),

$$q_m^{\pm}(\varepsilon)p = (\exp(2^m(\varepsilon/a_m)^{1/2m}X \pm \varepsilon(ad^mX, Y) + (\varepsilon/a_m)^{1+1/2m}W))p.$$

While, in theory, $(\varepsilon/a_m)^{1+1/2m} = o(\varepsilon)$ and is considered inessential, with m large (say $m \geqq 3$), this term can hardly be neglected for computational purposes. Either a better algorithm for generating $\pm(ad^mX, Y)(p)$, with m large, is needed, or in practice one should design the system to have a sufficient number of control components incorporated in such a way that high order brackets (ad^iX, Y^i) are unnecessary.

REFERENCES

[1] H. HERMES, *On local and global controllability*, this Journal, 12 (1974), 252–261.

[2] ——, *Local controllability and sufficient conditions in singular problems*, J. Differential Equations 20 (1976), pp. 213–232.

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

[4] H. HERMES, *Local controllability and sufficient conditions in singular problems, II*, this Journal 14 (1976), pp. 1049–1062.

[5] ——, *Controlled stability*, Ann. Mat. Pura Appl. 114 (1977), pp. 103–119.

[6] V. S. VARADARAGAN, *Lie Groups. Lie Algebras, and their Representations*, Prentice-Hall, Englewood Cliffs, NJ, (1974).

[7] G. FORSYTHE, M. MALCOLM AND C. MOLER, *Computer Methods for Mathematical Computations*, Prentice-Hall, Englewood Cliffs NJ 1977.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 4, July 1980

0363-0129/80/1804-0004 $01.00/0

OPTIMAL INFINITE-HORIZON UNDISCOUNTED CONTROL OF FINITE PROBABILISTIC SYSTEMS*

LOREN K. PLATZMAN†

Abstract. A finite-input, finite-state, finite-output stochastic control problem with imperfect state observation and classical information pattern is shown to be meaningful as the horizon increases without bound and the discount rate approaches unity. The plant model, a finite probabilistic system, includes the Markov decision and partially-observed Markov decision problems as special cases. Under conditions resembling controllability and observability in linear systems we show that: an optimal strategy exists, it may be realized by a stationary policy on the state estimate, its performance does not depend on the initial state distribution, and convergence rates for its finite-horizon and discounted performances are readily established.

1. Introduction. In this paper, we study a stochastic control problem in which

1) The decision-maker does not know the current state value, and acts on the basis of past observations and decisions, as well as an initial state distribution.
2) Performance is measured by averaging incremental rewards, with equal weight, over an indefinite time interval.

Such problems are characterized by various paradoxes. The performance may depend on how the limits (as the horizon increases without bound or the discount approaches unity) are taken [13]. In so-called "dual control" problems [4], [12] (where the decision-maker must choose between actions that improve short term performance and those that improve state information for the sake of performance in the long run) the infinite-horizon limit may be meaningless.

These difficulties vanish in the well-known Markov decision problem (MDP), where the state set is finite and the decision-maker always knows the current state value [5], [10], [14], [15], [16], [19], [25]. The object of this paper is to show how difficulties may be avoided in finite-state problems with imperfect state observation. Thus, we study the problem of optimally controlling a finite probabilistic system (FPS), a stationary discrete-time controlled stochastic process whose input, output, and (internal) state sets are finite.

Our principal results are:

1) A pair of conditions, one on the interaction of inputs and states (reachability) and one on the interaction of states and outputs (detectability), that together imply well-posedness of the infinite-horizon undiscounted problem (Theorem 3).
2) Bounds on convergence rates of finite-horizon and discounted performances as the infinite-horizon undiscounted limit is approached (Theorem 1).

This work is inspired by well-known (or, at least, generally assumed) properties of the infinite-horizon, linear-quadratic-Gaussian (LQG) control problem, (see, e.g., Kushner [18] or Athans [3]).

The expression "finite probabilistic system" is used in accordance with a classification of systems by Kalman, Falb, and Arbib [17]. Our representation of an FPS, borrowed from probabilistic automata theory, is that of Paz [20]. The FPS control

* Received by the editors, May 31, 1979, and in final revised form July 20, 1979.

† Bell Telephone Laboratories, Naperville, Illinois 60540. Now at Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, Michigan 48109. This research is based primarily on research conducted at the Decision and Control Sciences Group of the MIT Electronic Systems Laboratory, Cambridge, Massachusetts, supported in part by Air Force Office for Scientific Research under Contracts 72-2273 and 77-3281, and reported in the author's doctoral dissertation.

problem is a slight generalization of the partially-observed Markov decision problem (POMDP), independently conceived by Drake [11] and Astrom [1], among others, who showed that the problem is equivalent to a (nondenumerable state) MDP. Such MDP's have been studied by Ross [24].

The structure of our problem, however, makes it considerably more tractable than a general-state MDP. The value function is convex [2], and in some cases piecewise-linear as well [27]. Piecewise-linearity of the value function, discovered by Drake [11], but extensively developed by Sondik [28], occurs when the optimal strategy has a finite-memory realization. An example is given in § 5.

This paper is organized according to the following plan. The FPS model is introduced in § 2, along with standard state-estimation and dynamic programming terminology. The infinite-horizon undiscounted problem formulation and its consequences are presented in § 3. Conditions implying well-posedness of the problem are given in § 4. Section 5 is devoted to illustrative examples.

2. The model.

2.1. The plant. A *finite probabilistic* (*dynamical*) *system* (FPS) is a 4-tuple $\mathbf{S}=(U, Y, S, \{P(y|u): y\in Y, u\in U\})$ where:

(i) U is a finite nonempty set of *input values* (or *decisions*);

(ii) Y is a finite nonempty set of *output values* (or *observations*);

(iii) $S=\{1,\cdots,N\}$ is a finite nonempty set of (internal) *state values*;

(iv) Each $P(y|u)$ is an $N\times N$ substochastic matrix of state *transition probabilities*, and

$$P(u)=\sum_{y\in Y} P(y|u)$$

is a stochastic matrix $\forall u\in U$.

Let Π denote the simplex of horizontal stochastic N-vectors

$$\Pi=\{\pi: \pi^t\in\mathcal{R}^N,\ \pi_i\geqq 0\ \ \forall i\in S,\ \sum_{i\in S}\pi_i=1\},\tag{2.1}$$

and also define:

$$Z^*=\text{the free monoid generated by } U\times Y\text{; i.e., the set of finite strings of input-output pairs.}\tag{2.2}$$

$$\Gamma=U^{Z^*}=\text{the set of mappings from } Z^* \text{ to } U.\tag{2.3}$$

An FPS is studied in conjunction with an *initial state probability* (ISP) $\pi\in\Pi$ and a *control strategy* (CS) $\gamma\in\Gamma$. Given $(\mathbf{S},\pi,\gamma)$, we construct an *input process* $\{u(k)\in U\}_{k=0}^{\infty}$, a *state process* $\{s(k)\in S\}_{k=0}^{\infty}$, an *output process* $\{y(k)\in Y\}_{k=1}^{\infty}$, and an *information process* $\{z(k)\in Z^*\}_{k=0}^{\infty}$ according to:

$$\Pr[s(0)=i]=\pi_i,\qquad i\in S,\tag{2.4}$$

$$\begin{aligned}&\Pr[y(k+1)=y, s(k+1)=j|s(k)=i, u(k)=u, \{u(k')\}_{k'=0}^{k-1},\\ &\qquad\qquad \{s(k')\}_{k'=0}^{k-1},\{y(k')\}_{k'=1}^{k}]=P_{ij}(y|u),\\ &\qquad\qquad u\in U,\quad i,j\in S,\quad y\in Y,\quad k=0,1,2,\cdots,\end{aligned}\tag{2.5}$$

$$z(k)=\begin{cases}(u(0),y(1))(u(1),y(2))\cdots(u(k-1),y(k)), & k=1,2,\cdots,\\ \text{empty string}, & k=0,\end{cases}\tag{2.6}$$

$$u(k)=\gamma(z(k)),\qquad k=0,1,2,\cdots.\tag{2.7}$$

Thus $\{s(k), u(k)\}$ is an MDP and $\{[z(k), s(k)]\}$ is a denumerable state Markov chain.

Clearly, $s(k)$, $u(k)$, $y(k+1)$, $z(k)$, may be viewed as random variables on a probability space $\mathcal{P}_{\pi,\gamma}$ that depends on the ISP π and the CS γ. $P_{\pi,\gamma}$ and $E_{\pi,\gamma}$ will denote the probability measure and the expectation operator (respectively) associated with $\mathcal{P}_{\pi,\gamma}$.

The plant model includes a number of special cases, notably:

1) The MDP, when $Y=S$ and $P_{ij}(j|u)=P_{ij}(u)$, $i, j \in S$, $u \in U$.
2) The MDP with state observation delay [7] when $Y=S$ and $P_{ij}(i|u)=P_{ij}(u)$, $i, j \in S$, $u \in U$.
3) Certain statistical decision problems [9], when $P(u)$ does not depend on u.
4) The POMDP, when $y(k)$ is a "random function" of $s(k-1)$ and $u(k-1)$, i.e., when $P_{ij}(y|u)=P_{ij}(u)Q_{iy}(u)$. An alternate formulation of the POMDP expresses $y(k)$ as a "random function" of $u(k-1)$ and $s(k)$; then $P_{ij}(y|u)=P_{ij}(u)Q_{jy}(u)$.

2.2. The performance indices. Consider a bounded real-valued *reward function* R on $S\times U\times Y\times S$, and define:

$$r(k)=R[s(k), u(k), y(k+1), s(k+1)], \tag{2.8}$$

$$g(K)=K^{-1}\cdot \sum_{k=0}^{K-1} r(k), \qquad K=1,2,\cdots, \tag{2.9}$$

$$\tilde{g}(\beta)=(1-\beta)\cdot \sum_{k=0}^{\infty} \beta^k r(k), \qquad 0\leqq \beta<1. \tag{2.10}$$

We call $r(k)$ an *incremental reward*; $g(K)$ is the *actual finite-horizon average performance* for *horizon* K; and $\tilde{g}(\beta)$ is the *actual discounted average performance* for *discount* β. Each is a random variable on $\mathcal{P}_{\pi,\gamma}$.

Following convention, we assume that the decision-maker cannot directly observe the reward process $\{r(k)\}$. If it is desired to make this information available to the decision-maker, we may easily incorporate it into the observation process.

2.3. State-estimation of FPS's. In order to apply dynamic programming techniques to our problem, we must first devise a process of sufficient statistics [30], [31]. Following [1], [11], [27]–[29], the process we seek is one consisting of horizontal stochastic N-vectors $\{\eta(k)\in \Pi\}$ having entries:

$$\eta_i(k)=P_{\pi,\gamma}[s(k)=i|z(k)], \qquad i\in S. \tag{2.11}$$

We may readily verify that it satisfies the recursive form:

$$\eta(0)=\pi, \tag{2.12}$$

$$\eta(k+1)=T(\eta(k),(u(k),y(k+1))), \qquad k=0,1,2,\cdots, \tag{2.13}$$

where

$$T(\pi,(u,y))=\pi P(y|u)/\pi P(y|u)\nu, \tag{2.14}$$

and ν is a vertical N-vector whose entries all equal unity.

A multiple-step version of (2.14) will also be required. For $z=(u_1,y_1)\cdot(u_2,y_2)\cdots(u_l,y_l)\in Z^*$, define the matrix product $P(z)=P(y_1|u_1)\cdot P(y_2|u_2)\cdots P(y_l|u_l)$. Then let

$$T(\pi,z)=\pi P(z)/\pi P(z)\nu. \tag{2.15}$$

Now (2.12)–(2.13) may be expressed as:

(2.16) $$\eta(k)=T(\pi, z(k)).$$

2.4. Dynamic programming equations. Still following [1], [11], [27]–[29], define:

(2.17) V is the vector space of bounded real-valued functions on Π.

(2.18) $q(u)$ is the *expected incremental reward vector*, a vertical N-vector with entries

$$q_i(u)=\sum_{j\in S}\sum_{y\in Y} P_{ij}(y|u)R(i, u, y, j), \qquad i\in S, \quad u\in U.$$

(2.19) $\tilde{f}_\beta : V\rightarrow V$ is the *discounted dynamic programming operator*

$$[\tilde{f}_\beta v](\pi)=\max_{u\in U}\{\pi q(u)+\beta \sum_{y\in Y}(\pi P(y|u)\nu)v(T(\pi, (u, y)))\}.$$

(2.20) $f: V\rightarrow V$ is the *undiscounted dynamic programming operator* given by $f=\tilde{f}_1$.

The finite horizon (undiscounted) FPS control problem is solved [1], [11] by the iterative procedure:

(2.21) $$v_K=f v_{K-1}, \qquad K=1, 2, \cdots,$$ $$v_0=0,$$

where v_K has the interpretation:

(2.22) $$v_K(\pi)=\sup_{\gamma\in\Gamma}\left\{E_{\pi,\gamma}\left[\sum_{k=0}^{K-1} r(k)\right]\right\}=\sup_{\gamma\in\Gamma}\{E_{\pi,\gamma}[K\cdot g(K)]\}.$$

The optimal initial decision (when the ISP is π, and K decisions remain) is the maximal $u\in U$ used to determine $v_K(\pi)$ from v_{K-1}. To obtain the next decision, replace K by $K-1$ and replace π by $\eta(1)$.

It is also known that v_K is convex [2] and piecewise linear with a finite number of faces [27]. This may be casually explained as follows:

Convexity. A decision-maker faced with ISP π might be given additional state information to obtain ISP π' with probability λ, or ISP π'' with probability $(1-\lambda)$, where $\pi=\lambda\pi'+(1-\lambda)\pi''$. He cannot do worse with additional information, so $v_K(\pi)\leqq \lambda v_K(\pi')+(1-\lambda)v_K(\pi'')$.

Piecewise linearity. For any given CS γ, the performance $E_{\pi,\gamma}[g(K)]$ takes the form $\sum_{i\in S}\pi_i w_i(\gamma)$, where $w_i(\gamma)=E_{\cdot,\gamma}[g(K)|s(0)=i]$. But the finite-horizon problem admits a finite number of distinguishable CS's, so $W=\{w(\gamma)\}$ is a finite set. Since $v_K(\pi)=\max_{w\in W}\{\pi w\}$, it is piecewise-linear with at most $\# W$ faces.

Similarly, the discounted (infinite-horizon) FPS control problem is solved by determining the unique fixed point of $\tilde{f}_\beta$,

(2.23) $$\tilde{v}_\beta=\tilde{f}_\beta\tilde{v}_\beta,$$

which satisfies

(2.24) $$\tilde{v}_\beta(\pi)=\sup_{\gamma\in\Gamma}\left\{E_{\pi,\gamma}\left[\sum_{k=0}^{\infty}\beta^k r(k)\right]\right\}.$$

Using the contraction property of discounted dynamic programming operators (see, e.g., Bertsekas [6, § 6.3]), it is readily shown that

(2.25) the sequence $\{\tilde{f}_\beta^k v\}_{k=0}^{\infty}$ converges, uniformly in π, to $\tilde{v}_\beta$, $\forall v\in V$, $0\leqq\beta<1$.

If v is bounded, convex and continuous, then $\tilde{f}_\beta v$ is bounded, convex and continuous; hence, by (2.25),

$$\tilde{v}_\beta \text{ is bounded, convex and continuous.} \tag{2.26}$$

Apparently, $\tilde{v}_\beta$ may be, but need not be, piecewise linear [29].

Finally, we state, for future use, an inequality that follows trivially from the definition of f (by induction on k):

$$\sup_{\pi\in\Pi} \{[f^k v - f^k v'](\pi)\} \leqq \sup_{\pi\in\Pi} \{[v - v'](\pi)\} \qquad k = 0, 1, 2, \cdots. \tag{2.27}$$

This inequality is used to establish the convergence of value-iteration sequences such as (2.21).

3. Consequences of well-posedness in the infinite-horizon undiscounted limit.

3.1. Problem formulation. In essence, our problem is to demonstrate, for any ISP π, the existence of a CS γ_π that maximizes the *infinite-horizon undiscounted* (IHU) performance indices, $\lim_{K\to\infty} \{E_{\pi,\gamma}[g(K)]\}$ and $\lim_{\beta\uparrow 1} \{E_{\pi,\gamma}[\tilde{g}(\beta)]\}$, over all CS's $\gamma \in \Gamma$. Such a problem formulation, however, is unsatisfactory. For instance, the limits defining IHU performance may not exist or may not coincide for certain CS's. Furthermore, such a result would not address the rate of convergence of the performance of γ_π as $K \to \infty$ or $\beta \uparrow 1$, and it is this very consideration that determines whether the IHU formulation will be useful in practice.

Let us instead follow the approach of Bertsekas [6]. We will first state a condition for well-posedness and list some of its attractive consequences. Then, in § 4, we show how it can be verified.

This section is therefore devoted to consequences of

Condition 1. There is a $v_* \in V$ and a constant g such that

$$fv_* = v_* + g. \tag{3.1}$$

When Condition 1 is satisfied, we will frequently make use of

$$C = |v_*|, \tag{3.2}$$

where

$$|v| = \sup_{\pi\in\Pi} \{v(\pi)\} - \inf_{\pi\in\Pi} \{v(\pi)\}, \qquad v \in V. \tag{3.3}$$

When v_* exists, it is not unique, since v_* plus any constant is also a solution of (3.1). Theorem 2 will show that it can be taken, without loss of generality, to be convex. For some v_* satisfying (3.1), define:

$$v_*^+ = v_* - \inf_{\pi\in\Pi} \{v_*(\pi)\}, \tag{3.4}$$

$$v_*^- = v_* - \sup_{\pi\in\Pi} \{v_*(\pi)\}. \tag{3.5}$$

Now v_*^+ is a strictly nonnegative solution of (3.1) and v_*^- is a strictly nonpositive solution of (3.1).

Continuity of v_K and $\tilde{v}_\beta$ may suggest that a continuous solution to (3.1) exists whenever Condition 1 is satisfied. Example 4 of § 5 shows that such is not the case. Theorem 4 of § 4 will establish the continuity of v_* under an additional assumption. In this section, however, we will not require v_* to be either convex or continuous.

3.2. The policy and CS's corresponding to v_*. Let Condition 1 be satisfied and define $\psi^*: \Pi \to U$ to be a "policy" that identifies a maximizing input in (3.1):

$$fv_*(\pi) = \pi q(\psi^*(\pi)) + \sum_{y \in Y} (\pi P(y|\psi^*(\pi))\nu)\, v_*(T(\pi, (\psi^*(\pi), y))). \tag{3.6}$$

We may now construct a CS γ_π for each ISP π so that

$$u(k) = \psi^*(\eta(k)) \qquad \mathcal{P}_{\pi,\gamma_\pi} - \text{a.s.} \tag{3.7}$$

Simply combine (2.16) and (3.7) to obtain

$$\gamma_\pi(z) = \psi^*(T(\pi, z)). \tag{3.8}$$

For ease of notation, also define

$$\mathcal{P}_\pi^* = \mathcal{P}_{\pi,\gamma_\pi}, \quad P_\pi^* = P_{\pi,\gamma_\pi}, \quad E_\pi^* = E_{\pi,\gamma_\pi}. \tag{3.9}$$

Although (3.8) defines γ_π, it is (3.7) that would be used to realize it in practice. Since ψ^* does not depend on k, we say that γ_π is *realized by stationary policy on the state estimate* $\eta(k)$. It is straightforward to show that (2.4)–(2.7) with (3.8) define the same probability spaces $\mathcal{P}_\pi^*$ as do (2.4)–(2.5) with (2.12)–(2.14) and (3.7). We remark that measurability of ψ^* is required when (3.7) is used to define $\mathcal{P}_\pi^*$; it may be established, when v_* is convex, by partitioning Π into subsets

$$\Pi(A) = \{\pi : \pi_i > 0 \Leftrightarrow i \in A\}, \qquad \varnothing \neq A \subseteq S \tag{3.10}$$

over which $T(\cdot, z)$ and v_* are continuous, the latter by [23, Thm. 10.1].

3.3. Convergence of IHU performances. We may now show that the optimal finite-horizon and discounted performances, as well as the performances of γ_π, converge to g as $K \to \infty$ or $\beta \uparrow 1$. By any reasonable definition of IHU performance, this means that γ_π is an IHU-optimal CS for ISP π, and g is the optimal IHU performance or "gain". Since ψ^* realizes γ_π, it is known as the IHU-*optimal policy*.

LEMMA 1. *Let Condition* 1 *be satisfied. Then*

$$\text{(a)} \quad v_*^-(\pi) \leqq v_K(\pi) - Kg \leqq v_*^+(\pi),$$

$$\text{(b)} \quad v_*^-(\pi) \leqq \tilde{v}_\beta(\pi) - (1-\beta)^{-1} g \leqq v_*^+(\pi).$$

Proof. Part (a) clearly holds when $K = 0$, since $v_0 = 0$ by (2.21), and $v_*^+[v_*^-]$ is strictly nonnegative [nonpositive] by (3.4) [(3.5)]. We now prove (a) by induction on K. Let e denote an arbitrary constant and let $v \leqq v'$ signify $v(\pi) \leqq v'(\pi)\ \forall \pi \in \Pi$. From (2.19)–(2.20), we obtain

$$(*) \qquad \begin{aligned} f(v+e) &= (fv) + e, \\ v \leqq v' &\Rightarrow fv \leqq fv'. \end{aligned}$$

These identities enable us to show that

$$v_*^- \leqq v_K - Kg \leqq v_*^+ \Rightarrow fv_*^- \leqq fv_K - Kg \leqq fv_*^+.$$

The proof of (a) is completed by observing that v_*^+ and v_*^- satisfy (3.1), and that $fv_K = v_{K+1}$, by (2.21). Similarly, from (2.19), (2.20) and (3.4), (3.5), we obtain

$$\tilde{f}_\beta v = f(\beta v), \quad \beta v_*^+ \leqq v_*^+, \quad \beta v_*^- \geqq v_*^-,$$

which, along with (*), enables us to prove

$$[\tilde{f}_\beta^k v_*^+](\pi) - \frac{1-\beta^k}{1-\beta} g \leqq v_*^+(\pi),$$

$$[\tilde{f}_\beta^k v_*^-](\pi) - \frac{1-\beta^k}{1-\beta} g \geqq v_*^-(\pi)$$

by induction on k. Taking the limit $k \to \infty$ and applying (2.25) yields (b). □

THEOREM 1. *Let Condition* 1 *be satisfied. Then*

(a) *The optimal expected performance converges, uniformly in* π, *as* $O(K^{-1})$ *or* $O(1-\beta)$, *to* g; *specifically,*

$$\Big|\sup_{\gamma \in \Gamma} \{E_{\pi,\gamma}[g(K)]\} - g\Big| \leqq K^{-1} C, \qquad \pi \in \Pi, \quad K = 1, 2, \cdots,$$

$$\Big|\sup_{\gamma \in \Gamma} \{E_{\pi,\gamma}[\tilde{g}(\beta)]\} - g\Big| \leqq (1-\beta) C, \qquad \pi \in \Pi, \quad 0 \leqq \beta < 1.$$

(b) *The expected performance of* γ_π *converges, uniformly in* π, *as* $O(K^{-1})$ *or* $O(1-\beta)$, *to* g; *specifically,*

$$|E_\pi^*[g(K)] - g| \leqq K^{-1} C, \qquad \pi \in \Pi, \quad K = 1, 2, \cdots,$$

$$|E_\pi^*[\tilde{g}(\beta)] - g| \leqq (1-\beta) C, \qquad \pi \in \Pi, \quad 0 \leqq \beta < 1.$$

(c) *The expected suboptimality of* γ_π *converges, uniformly in* π, *as* $O(K^{-1})$ *or* $O(1-\beta)$, *to* 0; *specifically,*

$$\sup_{\gamma \in \Gamma} \{E_{\pi,\gamma}[g(K)]\} - E_\pi^*[g(K)] \leqq K^{-1} C, \qquad \pi \in \Pi, \quad K = 1, 2, \cdots,$$

$$\sup_{\gamma \in \Gamma} \{E_{\pi,\gamma}[\tilde{g}(\beta)]\} - E_\pi^*[\tilde{g}(\beta)] \leqq (1-\beta) C, \qquad \pi \in \Pi, \quad 0 \leqq \beta < 1.$$

Proof. By (3.1), (3.6)–(3.7) and (3.9),

$$E_\pi^*[r(k) | \eta(k)] = v_*(\eta(k)) + g - E_\pi^*[v_*(\eta(k+1)) | \eta(k)].$$

And so, (2.9)–(2.10) become

$$E_\pi^*[g(K)] - g = K^{-1}(v_*(\pi) - E_\pi^*[v_*(\eta(K))]),$$

$$E_\pi^*[\tilde{g}(\beta)] - g = (1-\beta)\Big(v_*(\pi) - \sum_{k=1}^{\infty} (1-\beta)\beta^{k-1} E_\pi^*[v_*(\eta(k))]\Big).$$

With (3.2), this establishes (b). By (2.9)–(2.10), and (2.22), (2.24),

$$\sup_{\gamma \in \Gamma} \{E_{\pi,\gamma}[g(K)]\} = K^{-1} v_K(\pi),$$

$$\sup_{\gamma \in \Gamma} \{E_{\pi,\gamma}[\tilde{g}(\beta)]\} = (1-\beta)\tilde{v}_\beta(\pi).$$

Lemma 1 is now invoked to establish (a) and (c). □

The sense in which γ_π is an IHU-optimal CS for ISP π may now be made precise.

COROLLARY. *Let Condition* 1 *be satisfied. Then*

$$\begin{array}{lll}
\text{(a)} & \lim\limits_{K\to\infty} \{E^*_\pi[g(K)]\} = g & \forall \pi \in \Pi, \\
\text{(b)} & \lim\limits_{\beta\uparrow 1} \{E^*_\pi[\tilde g(\beta)]\} = g & \forall \pi \in \Pi, \\
\text{(c)} & \limsup\limits_{K\to\infty} \{E_{\pi,\gamma}[g(K)]\} \leqq g & \forall \pi \in \Pi,\ \gamma \in \Gamma, \\
\text{(d)} & \limsup\limits_{\beta\uparrow 1} \{E_{\pi,\gamma}[\tilde g(\beta)]\} \leqq g & \forall \pi \in \Pi,\ \gamma \in \Gamma.
\end{array}$$ □

3.4. Sequences that converge to v_*. Let $\hat\pi$ be an arbitrary element of Π and define:

$$\hat v(\pi) = v(\pi) - v(\hat\pi), \qquad v \in V. \tag{3.11}$$

In an ergodic MDP, it is well-known [10], [14], [15], [16], [19], [25] that

1) $\hat v_*$ is unique.
2) $\hat v_K \to \hat v_*$ as $K \to \infty$.
3) $\hat{\tilde v}_\beta \to \hat v_*$ as $\beta \uparrow 1$.

Even when v_K is asymptotically periodic, a solution of (3.1) may be obtained by the damped value iteration procedure of P. J. Schweitzer [26], [21]:

$$\begin{array}{ll}
\bar v_K = \lambda f \bar v_{K-1} + (1-\lambda)\bar v_{K-1}, & K = 1, 2, \cdots, \\
\bar v_0 = 0, &
\end{array} \qquad 0 < \lambda < 1. \tag{3.12}$$

These ideas are now generalized to the FPS control problem.

It is to be noted that Ross [24] has proved a similar convergence theorem under the assumption that $\{\hat{\tilde v}_\beta\}_{0\leqq\beta<1}$ is equicontinuous. Since we have not excluded the case where v_* is necessarily discontinuous, Ross' theorem is inapplicable. Our analysis is based on the convexity of $\hat{\tilde v}_\beta$ and $\hat{\bar v}_K$, as well as the finite distribution of $\eta(k+1)$ given $u(k)$ and $\eta(k)$. Although we do not require it at this time, a connection between convexity and equicontinuity may be established using Lemma A.1 (in Appendix A).

THEOREM 2. *Let Condition* 1 *be satisfied. Then*

(a) *Any sequence $\beta'_n \uparrow 1$ has a subsequence $\beta_n \uparrow 1$ such that $\hat{\tilde v}_{\beta_n}$ is pointwise convergent. Moreover, the limit function $\hat{\tilde v}_*(\pi) = \lim_{n\to\infty}\{\hat{\tilde v}_{\beta_n}(\pi)\}$ is convex, satisfies* (3.1) *and $|\hat{\tilde v}_*| \leqq 2C$.*

(b) *Any sequence $K'_n \to \infty$ has a subsequence $K_n \to \infty$ such that $\hat{\bar v}_{K_n}$ is pointwise convergent. Moreover, the limit function $\hat{\bar v}_*(\pi) = \lim_{n\to\infty}\{\hat{\bar v}_{K_n}(\pi)\}$ is convex, satisfies* (3.1) *and $|\hat{\bar v}_*| \leqq 2C$.*

To prove Theorem 2, we require two preliminary results.

LEMMA 2. *Any uniformly bounded sequence of convex functions on Π, $\{v'_n\}$, has a pointwise convergent subsequence $\{v_n\}$. Moreover, the limit function $v(\pi) = \lim_{n\to\infty}\{v_n(\pi)\}$ is convex. If $|v_n| \leqq C'$ for all $n = 1, 2, \cdots$, then $|v| \leqq C'$.*

Proof. Using the partition (3.10) of Π, take successive subsequences, for each nonempty subset A of S, to obtain pointwise convergence on $\Pi(A)$. When A contains more than one element, the desired subsequence exists by [23, Thm. 10.9]. The limit v is convex since $\lambda v(\pi) + (1-\lambda)v(\pi') - v(\lambda\pi + (1-\lambda)\pi') \geqq \inf_n\{\lambda v_n(\pi) + (1-\lambda)v_n(\pi') - v_n(\lambda\pi + (1-\lambda)\pi')\} \geqq 0$. Finally $|v| = \sup_{\pi,\pi'\in\Pi}\{v(\pi) - v(\pi')\} \leqq \sup_{\pi,\pi'\in\Pi}\{\sup_n\{v_n(\pi) - v_n(\pi')\}\} \leqq \sup_n |v_n|$, so $|v_n| \leqq C'\ \forall_n \Rightarrow |v| \leqq C'$. □

LEMMA 3. *If $v_n \to v$ (pointwise), then $fv_n \to fv$ (pointwise).*

Proof. For any $\pi \in \Pi$, $\varepsilon > 0$, we must show that there is an M such that $|[fv_n - fv](\pi)| < \varepsilon$, $\forall n \geqq M$. Define the finite set $B = \{T(\pi, (u, y)) \colon u \in U,\ y \in Y,\ \pi P(y|u) \neq 0\}$

and let M be such that $|v_n(\pi')-v(\pi')|<\varepsilon, \forall \pi' \in B, n \geqq M$. The desired result follows immediately from the definition of f. □

Proof of Theorem 2. (a) By (3.11) and Lemma 1, $|\hat{\tilde{v}}_\beta| = |\tilde{v}_\beta| \leqq 2C$. Now $|\hat{\tilde{v}}_\beta(\pi)| = |\tilde{v}_\beta(\pi)-\tilde{v}_\beta(\hat{\pi})| \leqq |\tilde{v}_\beta|$, so $\{\hat{\tilde{v}}_\beta(\pi)\}$ is uniformly bounded. By Lemma 2, there exists a sequence $\beta_n \uparrow 1$ such that $\hat{\tilde{v}}_{\beta_n} \to \hat{\tilde{v}}_*$ (pointwise), $\hat{\tilde{v}}_*$ is convex, and $|\hat{\tilde{v}}_*| \leqq 2C$. By (2.23), $\tilde{f}_\beta \hat{\tilde{v}}_\beta - \hat{\tilde{v}}_\beta = (1-\beta)\tilde{v}_\beta(\hat{\pi})$; Lemma 1 implies $(1-\beta_n)\tilde{v}_{\beta_n}(\hat{\pi}) \to g$; and so $\tilde{f}_{\beta_n}\hat{\tilde{v}}_{\beta_n} - \hat{\tilde{v}}_* \to g$ (pointwise). But Lemma 3 implies $f\hat{\tilde{v}}_{\beta_n} \to f\hat{\tilde{v}}_*$ (pointwise), and uniform boundedness of $\{\hat{\tilde{v}}_\beta\}$ implies $[f_{\beta_n} - f]\hat{\tilde{v}}_{\beta_n} \to 0$. Consequently, $f\hat{\tilde{v}}_* - \hat{\tilde{v}}_* = g$, and so $\hat{\tilde{v}}_*$ satisfies (3.1).

(b) Following Schweitzer [26], we define a modified system that undergoes a transition in the usual manner with probability λ, or does nothing with probability $1-\lambda$. Specifically, for $y_0 \notin Y$, define

$$\begin{aligned}
&\bar{Y} = Y \cup \{y_0\},\\
&\bar{P}(y|u) = \lambda P(y|u) \quad \text{if } y \in Y, \qquad \bar{P}(y_0|u) = (1-\lambda)I,\\
&\bar{R}[i,u,y,j] = R[i,u,y,j] \quad \text{if } y \in Y, \qquad \bar{R}[i,u,y_0,j] = 0,\\
&\bar{g} = \lambda g.
\end{aligned} \tag{3.13}$$

Let $\bar{f}$ be the undiscounted dynamic programming operator (2.20) for the modified system. Clearly,

$$\bar{f}v = \lambda f v + (1-\lambda)v. \tag{3.14}$$

Condition 1 is satisfied for the modified system, since $\bar{f}v_* = v_* + \bar{g}$. Since (3.12) is the modified version of (2.22), Lemma 1 yields

$$v_*^-(\pi) \leqq \bar{v}_K(\pi) - K\bar{g} \leqq v_*^+(\pi), \tag{3.15}$$

and so $|\bar{v}_K| \leqq 2C$. Existence of a pointwise convergent subsequence $\hat{\bar{v}}_{K_n} \to \hat{\bar{v}}_*$, convexity of $\hat{\bar{v}}_*$, and $|\hat{\bar{v}}_*| \leqq 2C$ are obtained from Lemma 2 as in the proof of part (a), above. It remains to show that $\hat{\bar{v}}_*$ satisfies (3.1).

Define

$$\begin{aligned}
x_K &= \bar{v}_{K+1} - \bar{v}_K - \bar{g},\\
L_K &= \sup_{\pi \in \Pi} \{x_K(\pi)\},\\
L &= \limsup_{K \to \infty} \{L_K\}.
\end{aligned} \tag{3.16}$$

From (2.27), we may readily obtain the Schweitzer–Odoni inequalities:

$$x_{K+1}(\pi) \leqq \lambda L_K + (1-\lambda)x_K(\pi), \qquad L_{K+1} \leqq L_K. \tag{3.17}$$

Thus, the Odoni bounds L_K converge monotonically to L. We may now show, by contradiction, that $L=0$. Assume $L>0$ and select δ, M, ε, K, π so that

$$\begin{aligned}
&\delta > 0,\\
&LM > C + \delta\\
&\varepsilon\left[2\sum_{m=0}^{M-1}(1-\lambda)^{-m}\right] \leqq \delta,\\
&L_K \leqq L + \varepsilon\\
&x_{K+M}(\pi) \geqq L - \varepsilon.
\end{aligned} \tag{3.18}$$

For $m=1,\cdots,M-1$, (3.17) yields $L_{K+M-m} \leqq L_K$ and $x_{K+M-m+1}(\pi) \leqq \lambda L_K + (1-\lambda)x_{K+M-m}(\pi)$, or, equivalently, $x_{K+M-m+1}(\pi) - L_K \leqq (1-\lambda)(x_{K+M-m}(\pi) - L_K)$. Thus

$$x_{K+M-m}(\pi) - L_K \geqq (1-\lambda)^{-m}(x_{K+M}(\pi) - L_K), \qquad m=0,1,\cdots,M-1. \tag{3.19}$$

Hence, by (3.16) and (3.18),

$$\begin{aligned} \bar{v}_{K+M}(\pi) - \bar{v}_K(\pi) &= \sum_{m=0}^{M-1} (x_{K+M-m}(\pi) + \bar{g}) \\ &\geqq L_K M - \delta + M\bar{g} \\ &> C + M\bar{g} \end{aligned} \tag{3.20}$$

which contradicts (3.15). Thus $L=0$. By a similar argument, $\liminf_{K\to\infty}\{\inf_{\pi\in\Pi}\{x_K(\pi)\}\} = 0$. Hence $x_K \to 0$, and so, by Lemma 3, $\hat{\bar{v}}_*$ satisfies (3.1). □

4. Sufficient conditions for well-posedness.

4. Sufficient conditions for well-posedness. We now consider the problem of showing, on the basis of simple conditions on **S** and R, that Condition 1 is satisfied.

4.1. Summary of results. Let e^i denote the "unit vector" in Π having entries

$$e^i_j = \begin{cases} 1 & \text{if } i=j, \\ 0 & \text{if } i \neq j, \end{cases} \tag{4.1}$$

and let S^* be a subset of S such that

$$\max_{\pi\in\Pi}\{\tilde{v}_\beta(\pi)\} = \max_{i\in S^*}\{\tilde{v}_\beta(e^i)\}, \qquad 0 \leqq \beta < 1. \tag{4.2}$$

Now each $\tilde{v}_\beta$ is a convex function on the simplex Π, and so it achieves its maximum at a vertex e^i, $i \in S$. Thus S^* may be taken to equal S. In some problems, a smaller set S^* may be identified. In the machine repair problem of Smallwood and Sondik [27], for example, S^* would contain the single state corresponding to "no failures of any kind". Further information on partial ordering of S and its implications in this context may be found in [32].

Also let

$$\begin{aligned} Q^+ &= \max_{i\in S, u\in U} \{q_i(u)\}, \\ Q^- &= \min_{i\in S, u\in U} \{q_i(u)\}, \\ Q &= Q^+ - Q^-. \end{aligned} \tag{4.3}$$

We now define four conditions on **S**, two of which involve S^*.

Condition 2 (*Reachability*). There is a $\rho<1$ and an integer ξ such that

$$\sup_{\gamma\in\Gamma} \max_{0\leqq k\leqq\xi} \{P_{\pi,\gamma}[s(k)=j]\} \geqq 1-\rho \quad \forall \pi\in\Pi, j\in S^*. \tag{4.4}$$

Remark. Condition 2 deals primarily with the extent to which states may be influenced by proper selection of inputs. It assures that for any present value π of $\eta(k)$, there is a CS γ that will bring the system into an optimal state j within ξ time units with positive probability. A decision-maker who exercises this option to "reach" j will not necessarily be able to tell whether the system actually enters j (unless $\rho=0$). For this reason, Condition 2 alone is not sufficient to establish Condition 1. Condition 2 is similar to conditions of Bertsekas [6, p. 358] and Platzman [21], which are used to establish Condition 1 for MDP's. A simple test for Condition 2 is accomplished by computing $\Upsilon \doteq [I+\sum_{u\in U} P(u)]^N$; Condition 2 is satisfied if $\Upsilon_{ij} > 0$ for all $i \in S$, $j \in S^*$.

Condition 2* (*Resetability*). Condition 2 is satisfied with $\rho = 0$.

Remark. Condition 2* is much stronger than Condition 2 because it assures that the desired state will be reached with probability one. Thus the decision-maker will know when state j has been reached. An action that places the system in an optimal state with probability one is called a "reset action", hence the name "resetability". Sondik [28] has shown that

$$\text{Condition } 2^* \Rightarrow \text{Condition 1 with } C \leqq \xi Q. \tag{4.5}$$

Condition 3 (*Detectability*). There is an $a<1$ and an integer ζ such that

$$E_{\pi,\gamma}[a[P(z(\zeta))]] \leqq a \quad \forall \pi \in \Pi, \gamma \in \Gamma, \tag{4.6}$$

where $a[\cdot]$ denotes the "modified ergodic coefficient" of an $N\times N$ substochastic matrix:

$$a[P] = \max\left\{ D\left[\frac{e^iP}{e^iP\nu}, \frac{e^{i'}P}{e^{i'}P\nu}\right] : i, i' \in S, e^iP \neq 0, e^{i'}P \neq 0 \right\} \tag{4.7}$$

and D is a metric on Π defined and discussed in Appendix A.

Remark. Condition 3 deals primarily with the interaction of states and outputs. The function $a[\cdot]$ is closely related to ergodic coefficients for the state estimation process $\{\eta(k)\}$, and Condition 3 implies that the (nondenumerable state) MDP $\{\eta(k), u(k)\}$ is in a certain sense ergodic [22]. This assures continuity of any solution to (3.1), but a controllability assumption is required to establish boundedness. Condition 3 may be verified by showing that $a[P(z)] < 1$ for all z in a suitably large subset of Z^*; this is most easily accomplished by exploiting the notion of *subrectangularity* defined and discussed in Appendix B. Thus, in the MDP, for example, Condition 3 is trivially satisfied. A general procedure for verifying Condition 3 is given in [22].

Condition 3* (*Renewability*). Condition 3 is satisfied and there is an $\eta^* \in \Pi$ such that:

$$\max_{0\leqq k \leqq \zeta} \{P_{\pi,\gamma}[\eta(k) = \eta^*] \geqq 1-a\} \quad \forall \pi \in \Pi, \gamma \in \Gamma. \tag{4.8}$$

Remark. Condition 3* has a "built in" controllability assumption since η^* can be regarded as a state which is entered within a finite time interval with probability one. Ignoring transient states, Condition 3* implies Condition 2. Indeed, under Condition 3*, $\{\eta(k), u(k)\}$ becomes a denumerable-state Markov process on state set $\{T(\eta^*, z): z \in Z^*, \hat{\pi}P(z)\nu > 0\}$, and Condition 1 may be obtained by standard MDP methods. Ross [24] has shown that

$$\text{Condition } 3^* \Rightarrow \text{Condition 1 with } C \leqq \zeta Q(1-a)^{-1}. \tag{4.9}$$

In this section, we will prove

THEOREM 3. *Condition* 2 *and Condition* 3 ⇒ *Condition* 1 *with*

$$C \leqq \frac{(\xi+\zeta)Q}{(1-\rho)(1-a)}. \tag{4.10}$$

THEOREM 4. *Condition* 1 *and Condition* 3 *imply*

(a) *Every convex solution of* (3.1) *is continuous.*

(b) *The sequence* $\{\hat{v}_K\}_{K=0}^{\infty}$ *is uniformly convergent to a convex continuous solution of* (3.1).

Remark. Theorem 3 has a clear analogy in LQG theory, where separate conditions of controllability and observability are required to assure infinite-horizon stability. The analogy to Condition 2* would be a system that could be driven to state 0 by simple output feedback; a more extreme case might be a system whose state never deviates from 0. Condition 3* corresponds to a system whose state estimate never deviates from 0. In the absence of linearity assumptions, Conditions 2* and 3* are, of course, more interesting.

4.2. Methodology for proving Theorem 3. Our aim is to obtain (from Conditions 2 and 3) an inequality of the form

$$|\tilde{v}_\beta| \leqq KQ + \alpha|\tilde{v}_\beta|. \tag{4.11}$$

Then $|\tilde{v}_\beta| \leqq KQ/(1-\alpha)$, so $\{|\hat{\tilde{v}}_\beta|\}$ is uniformly bounded, and v_* may be obtained as in the proof of Theorem 2(a).

We first restate (2.23) as

$$\tilde{v}_\beta(\pi) = \max_{\gamma\in\Gamma}\{E_{\pi,\gamma}[r(0)+\beta\tilde{v}_\beta(\eta(1))]\} \tag{4.12}$$

or, more generally,

$$\tilde{v}_\beta(\pi) = \max_{\gamma\in\Gamma}\left\{E_{\pi,\gamma}\left[\left(\sum_{k=0}^{K-1}\beta^k r(k)\right)+\beta^K\tilde{v}_\beta(\eta(K))\right]\right\}. \tag{4.13}$$

In (4.13), a maximizing γ exists for any $K>0$, $0\leqq\beta<1$, and $\pi\in\Pi$, since the right-hand side of (4.13) involves only the finite horizon K.

The inequality (4.11) will be obtained from (4.13). We may derive (4.5) by setting $K=\xi$, or (4.9) by setting $K=\zeta$. To prove Theorem 3, we proceed in two steps; first bounding $f_\beta^\zeta\tilde{v}_\beta$ according to Condition 3, and then bounding $\tilde{f}_\beta^\xi[\tilde{f}_\beta^\zeta\tilde{v}_\beta]$ according to Condition 2.

4.3. Proof of Theorem 3. Let $j\in S^*$ be such that $\tilde{v}_\beta(e^j)=\max_{\pi\in\Pi}\{\tilde{v}_\beta(\pi)\}$ and let $\tilde{\gamma}\in\Gamma$ maximize (4.13) with $\pi=e^j$ and $K=\zeta$; this is possible by an argument stated immediately following (4.13). For ease of notation let $E_{j,\gamma}$ denote $E_{\pi,\gamma}$ with $\pi=e^j$. Substituting (2.16) into (4.13), we obtain

$$\tilde{v}_\beta(e^j) = E_{j,\tilde{\gamma}}\left[\left(\sum_{k=0}^{\zeta-1}\beta^k r(k)\right)+\beta^\zeta\tilde{v}_\beta(T(e^j,z(k)))\right]. \tag{4.14}$$

But, for any $\pi\in\Pi$,

$$\begin{aligned}\tilde{v}_\beta(\pi) &\geqq E_{\pi,\tilde{\gamma}}\left[\left(\sum_{k=0}^{\zeta-1}\beta^k r(k)\right)+\beta^\zeta\tilde{v}_\beta(T(\pi,z(k)))\right]\\ &= E_{\pi,\tilde{\gamma}}\left[\sum_{k=0}^{\zeta-1}\beta^k r(k)\right]+\pi_j E_{j,\tilde{\gamma}}[\beta^\zeta\tilde{v}_\beta(T(\pi,z(k)))]\\ &\quad+(1-\pi_j)E_{\pi,\tilde{\gamma}}[\beta^\zeta\tilde{v}_\beta(T(\pi,z(k)))|s(0)\neq j].\end{aligned} \tag{4.15}$$

Also, by (2.8), (2.18), and (4.3),

$$Q^- \leqq E_{\pi,\gamma}[r(k)] \leqq Q^+ \quad \forall\pi\in\Pi,\ \gamma\in\Gamma. \tag{4.16}$$

Now Lemma A.2 and (4.14)–(4.16) may be combined to form

$$\begin{aligned}\tilde{v}_\beta(e^j)-\tilde{v}_\beta(\pi) &\leqq \left(\sum_{k=0}^{\zeta-1}\beta^k Q\right)+\beta^\zeta(\pi_j a_{j,\tilde{\gamma}}+1-\pi_j)|\tilde{v}_\beta|\\ &\leqq \zeta Q+(\pi_j a_{j,\tilde{\gamma}}+1-\pi_j)|\tilde{v}_\beta|,\end{aligned} \tag{4.17}$$

where $a_{j,\gamma} = E_{j,\gamma}[a[P(z(\zeta))]]$. If Condition 3 is satisfied, then

$$a_{j,\tilde{\gamma}} \leqq a. \tag{4.18}$$

We will use (4.17) and (4.18), along with some implications of Condition 2, to obtain a bound of the form (4.11). Condition 2 implies that for any ISP $\hat{\pi}$, there is a CS $\hat{\gamma}$ and a $k \leqq \xi$ such that

$$P_{\hat{\pi},\hat{\gamma}}[s(k) = j] \geqq 1 - \rho. \tag{4.19}$$

Since $E_{\hat{\pi},\hat{\gamma}}[P_{\hat{\pi},\hat{\gamma}}[s(k) = j | z(k)]] = P_{\hat{\pi},\hat{\gamma}}[s(k) = j]$, this becomes

$$E_{\hat{\pi},\hat{\gamma}}[\eta_j(k)] \geqq 1 - \rho. \tag{4.20}$$

Again, (4.13) and (4.16) may be used to obtain

$$\begin{aligned} \tilde{v}_\beta(\hat{\pi}') &\leqq \left(\sum_{k'=0}^{k-1} \beta^{k'} Q^+ \right) + \beta^k \tilde{v}_\beta(e^j), \\ \tilde{v}_\beta(\hat{\pi}) &\geqq \left(\sum_{k'=0}^{k-1} \beta^{k'} Q^- \right) + \beta^k E_{\hat{\pi},\hat{\gamma}}[\tilde{v}_\beta(\eta(k))]. \end{aligned} \tag{4.21}$$

Thus

$$\begin{aligned} \tilde{v}_\beta(\hat{\pi}') - \tilde{v}_\beta(\hat{\pi}) &\leqq \left(\sum_{k'=0}^{k-1} \beta^{k'} Q \right) + \beta^k E_{\hat{\pi},\hat{\gamma}}[\tilde{v}_\beta(e^j) - \tilde{v}_\beta(\eta(k))] \\ &\leqq kQ + E_{\hat{\pi},\hat{\gamma}}[\tilde{v}_\beta(e^j) - \tilde{v}_\beta(\eta(k))]. \end{aligned} \tag{4.22}$$

Now substitute (4.17)–(4.18) and (4.20) into (4.22) to obtain

$$\begin{aligned} \tilde{v}_\beta(\hat{\pi}') - \tilde{v}_\beta(\hat{\pi}) &\leqq (k + \zeta)Q + [1 - (1 - \rho)(1 - a)] |\tilde{v}_\beta| \\ &\leqq (\xi + \zeta)Q + [1 - (1 - \rho)(1 - a)] |\tilde{v}_\beta|, \end{aligned} \tag{4.23}$$

which is a bound of the form (4.11). Hence $|\tilde{v}_\beta| \leqq C = (\xi + \zeta)Q/((1 - \rho)(1 - a))$. Proceeding exactly as in Theorem 2(a), there is a sequence $\beta_n \uparrow 1$ such that $\hat{\tilde{v}}_{\beta_n} \to \hat{\tilde{v}}_*$, $\hat{\tilde{v}}_*$ satisfies (3.1), and $|\hat{\tilde{v}}_*| \leqq C$. Thus Condition 1 is satisfied. Q.E.D.

4.4. Proof of Theorem 4. Condition 3 implies that, for every $\pi \in \Pi$ and $\gamma \in \Gamma$, there is a $\hat{z} \in Z^*$ such that $a[P(\hat{z})] < 1$, length$(\hat{z}) = \zeta$, and $P_{\pi,\gamma}[z(\zeta) = \hat{z}] > 0$. Indeed,

$$\delta(\pi) = \min_{\gamma \in \Gamma} \{\max \{P_{\pi,\gamma}[z(\zeta) = \hat{z}]: \hat{z} \in Z^*, \text{length}(\hat{z}) = \zeta, a[P(\hat{z})] < 1\}\} \tag{4.24}$$

is positive and continuous throughout Π. Hence,

$$\delta = \inf_{\pi \in \Pi} \{\delta(\pi)\} > 0. \tag{4.25}$$

To prove part (a) of Theorem 4, let v_* be a convex solution of (3.1), and let $\{\pi^n\}_{n=1}^\infty$ be a sequence in Π such that $\pi^n \to \pi$ and $v_*(\pi) - v_*(\pi^n) \to \varepsilon$. Then $\varepsilon \geqq 0$ since v_* is convex [23, Thm. 10.2]. Let L be the least upper bound on all such discontinuities:

$$L = \sup_{\pi} \{\limsup_{\pi' \to \pi} \{v_*(\pi) - v_*(\pi')\}\}. \tag{4.26}$$

We show that $L = 0$. Select $\hat{z}$ so that $a[P(\hat{z})] < 1$ and $P_\pi^*[z(\zeta) = \hat{z}] = \tilde{\delta} \geqq \delta$. Now Lemma B.3 implies that

$$T(\cdot, \hat{z}) \in \Pi(\hat{A}), \tag{4.27}$$

where $\hat{A} = J(P(\hat{z})) = \{j: P_{ij}(\hat{z}) > 0, \text{ some } i \in S\}$, and $\Pi(\cdot)$ is given by (3.10). Rewrite

(3.1) in the form

$$
\begin{aligned}
v_*(\pi)+g &= \max_{\gamma\in\Gamma}\{E_{\pi,\gamma}[r(0)+v_*(\eta(1))]\} \\
&= E_{\pi,\gamma_\pi}[r(0)+v_*(\eta(1))] \\
&= E_{\pi,\gamma_\pi}[r(0)+v_*(T(\pi, z(1)))]
\end{aligned} \tag{4.28}
$$

or, more generally,

$$
v_*(\pi)+\zeta g = E_{\pi,\gamma_\pi}\left[\left(\sum_{k=0}^{\zeta-1} r(k)\right)+v_*(T(\pi, z(\zeta)))\right]. \tag{4.29}
$$

Likewise,

$$
v_*(\pi^n)+\zeta g \geqq E_{\pi^n,\gamma_\pi}\left[\left(\sum_{k=0}^{\zeta-1} r(k)\right)+v_*(T(\pi^n, z(\zeta)))\right]. \tag{4.30}
$$

But,

$$
\begin{aligned}
&E_{\pi,\gamma_\pi}[v_*(T(\pi, z(\zeta)))]-E_{\pi^n,\gamma_\pi}[v_*(T(\pi^n, z(\zeta)))] \\
&\quad= \sum_{\substack{z\in Z^* \\ i\in S}} P_{i,\gamma_\pi}[z(\zeta)=z](\pi_i v_*(T(\pi, z))-\pi_i^n v_*(T(\pi^n, z))).
\end{aligned} \tag{4.31}
$$

Clearly, $T(\pi^n, z)\to T(\pi, z)$ whenever $\pi P(z)\neq 0$. But $v_*(T(\pi^n, \hat{z}))\to v_*(T(\pi, \hat{z}))$ as well, since $T(\pi^n, \hat{z})$, $T(\pi, \hat{z})\in \Pi(\hat{A})$ by (4.27), and since v_* is continuous throughout $\Pi(\hat{A})$ by [23, Thm. 10.1]. Now (4.29)–(4.31) yield $\varepsilon \leqq (1-\delta)L$. Since L is the least upper bound on ε, this becomes $L\leqq(1-\delta)L$, and so, by (4.25), $L=0$.

To prove part (b), consider the modified system $\bar{\mathbf{S}}=(U, \bar{Y}, S, \bar{P})$ given by (3.13), and used to prove Theorem 2(b). Just as $\bar{f}$ is the modified version of f (i.e., it is generated by $\bar{\mathbf{S}}$ rather than S), and $\bar{v}$ is the modified version of v, let $\bar{Z}^*$, $\bar{\gamma}\in\bar{\Gamma}$, $\bar{z}(k)$, $\bar{P}_{\pi,\bar{\gamma}}$, $\bar{E}_{\pi,\bar{\gamma}}$, $\bar{T}$, $\bar{\delta}$ be modified versions of Z^*, $\gamma\in\Gamma$, $z(k)$, $P_{\pi,\gamma}$, $E_{\pi,\gamma}$, T, δ, respectively. If $\mathbf{S}$ satisfies Condition 3, then so does $\bar{\mathbf{S}}$; so by (4.25),

$$
\bar{\delta}>0. \tag{4.32}
$$

Following Theorem 2(b) and [23, Thm. 10.9], let $\{\hat{\bar{v}}_{K_n}\}$ be a subsequence of $\{\hat{\bar{v}}_K\}$ that converges pointwise to a convex solution $\hat{\bar{v}}_*$ of (3.1) *and* converges uniformly (to $\hat{\bar{v}}_*$) on the closed subsets of $\Pi(A)$, $A\subseteq S$.

Define

$$
L_K=\sup_{\pi\in\Pi}\{|\tilde{\bar{v}}_*(\pi)-\hat{\bar{v}}_K(\pi)|\}, \qquad L=\limsup_{K\to\infty}\{L_K\}. \tag{4.33}
$$

We wish to show that $L=0$, for $\{\hat{\bar{v}}_K\}$ would then be uniformly convergent (on Π) to $\hat{\bar{v}}_*$, and Theorem 4(b) will have been proved.

Further define

$$
\begin{aligned}
\bar{Z}^l &= \{\bar{z}\in\bar{Z}^*\colon \text{Length}\,(\bar{z})=l \text{ and } a[\bar{P}(\bar{z})]<1\}, \\
\bar{\mu}_l &= \inf_{\pi\in\Pi,\bar{\gamma}\in\bar{\Gamma}}\{\bar{P}_{\pi,\bar{\gamma}}[\bar{z}(l)\in\bar{Z}^l]\}, \\
\bar{\Pi}_l &= \{\bar{T}(\pi, \bar{z})\colon \pi\in\Pi,\ \bar{z}\in\bar{Z}^l,\ \pi\bar{P}(\bar{z})\neq 0\}, \\
\bar{L}_{K,l} &= \sup_{\pi\in\bar{\Pi}_l}\{|\hat{\bar{v}}_*(\pi)-\hat{\bar{v}}_K(\pi)|\}.
\end{aligned} \tag{4.34}
$$

By (4.24)–(4.25) and (4.32), $\bar{\mu}_\zeta \geqq \bar{\delta} > 0$. Moreover, by (4.34) and Lemmas B.1 and B.2, $\bar{\mu}_{l+l'} \geqq \bar{\mu}_l + (1-\bar{\mu}_l)\bar{\mu}_{l'}$. Consequently

$$(4.35) \qquad \bar{\mu}_l \uparrow 1 \quad \text{as } l \to \infty.$$

By Lemma B.3, $\bar{\Pi}_l \cap \Pi(A)$ is a closed subset of $\Pi(A)$, $\forall A \subseteq S$. Thus $\{\hat{\bar{v}}_{K_n}\}$ converges uniformly on $\bar{\Pi}_l$. Since $\bar{\Pi}_l$ is nonempty for $l \geqq \zeta$, this may be stated as

$$(4.36) \qquad \lim_{n\to\infty} \{\bar{L}_{K_n,l}\} = 0, \qquad l \geqq \zeta.$$

Once again, let us assume that $L > 0$ and obtain a contradiction.

Select $\hat{l}$, using (4.35), so

$$(4.37) \qquad \bar{\mu}_{\hat{l}} \geqq \tfrac{15}{16}.$$

Also select $\hat{K} = K_{\hat{n}}$, using (4.33) and (4.36), so

$$(4.38) \qquad L_{\hat{K}} \leqq 2L, \qquad \bar{L}_{\hat{K},\hat{l}} \leqq L/8.$$

Now, for all $K' \geqq \hat{K} + \hat{l}$,

$$L_{K'} \leqq 2 \sup_{\pi\in\Pi} \{|[\bar{f}^{K'-\hat{K}}\hat{\bar{v}}_*](\pi) - [\bar{f}^{K'-\hat{K}}\hat{\bar{v}}_K](\pi)|\}.$$

By (2.27),

$$L_{K'} \leqq 2 \sup_{\pi\in\Pi} \{|[\bar{f}^{\hat{l}}\hat{\bar{v}}_*](\pi) - [\bar{f}^{\hat{l}}\hat{\bar{v}}_{\hat{K}}](\pi)|\}.$$

Expanding in the manner of (4.29)–(4.31),

$$(4.39) \qquad L_{K'} \leqq 2 \sup_{\pi\in\Pi,\bar{\gamma}\in\bar{\Gamma}} \{\bar{E}_{\pi,\bar{\gamma}}[[\hat{\bar{v}}_* - \hat{\bar{v}}_{\hat{K}}](\bar{T}(\pi, \bar{z}(\hat{l})))]\}$$

and so, by (4.34),

$$L_{K'} \leqq 2[\bar{\mu}_{\hat{l}}\bar{L}_{\hat{K},\hat{l}} + (1-\bar{\mu}_{\hat{l}})L_{\hat{K}}].$$

Finally, (4.35), (4.37), and (4.38) yield

$$(4.40) \qquad L_{K'} \leqq L/2.$$

But now, by (4.33), $L \leqq L/2$. Moreover, $L_0 = C$ by (3.1) and (3.12); (4.39) with $\hat{K} = 0$ yields $L_{K'} \leqq 2L_0$, and so $L \leqq 2C$. Thus $L = 0$. Q.E.D.

5. Examples. In this section, we give simple illustrative examples of FPS control problems.

Example 1: *The MDP.* In an MDP, the output and state processes coincide. Thus, the entries of matrix $P(y|u)$ equal zero everywhere except in column y, and so $a[P(y|u)] = 0$. Thus, Condition 3 is satisfied with $\zeta = 1$ and $a = 0$. Condition 2 is known as a sufficient condition for infinite-horizon undiscounted well-posedness of MDP's; see [6, p. 358] or [21]. Note that Condition 2 does not imply aperiodicity of the optimal system; a solution to (3.1) may exist even when $\hat{v}_K$ is asymptotically periodic.

Example 2: *The MDP with delayed state observation* [7]. If $y(k) = s(k-1)$, we have an MDP with delayed state observation. The entries of $P(y|u)$ now equal zero everywhere except in row y, so Condition 3 is again trivially satisfied.

Example 3: *A problem having a finite-memory solution.* This example is drawn from the doctoral dissertation of E. Sondik [28].

A wealthy industrialist employs two analysts (u = 1, 2) to manage his holdings. The holdings may be in a loss state (s = 1) or profit state (s = 2). Each analyst's effect on the holdings is modeled as a Markov chain with transition matrix $P(u)$ and profits $q(u)$. At the end of the month, the analyst handling the funds reports on their state. The report is correct with probability $\mu(s, u)$. We combine $P(u)$ and $\mu(s, u)$ to obtain

$$P(1|1)=\begin{bmatrix}.48 & .04\\ .30 & .10\end{bmatrix}, \qquad P(1|2)=\begin{bmatrix}.45 & .20\\ .36 & .24\end{bmatrix},$$

$$P(2|1)=\begin{bmatrix}.32 & .16\\ .20 & .40\end{bmatrix}, \qquad P(2|2)=\begin{bmatrix}.05 & .30\\ .04 & .36\end{bmatrix},$$

$$q(1)=\begin{bmatrix}-4\\ 4\end{bmatrix}, \qquad q(2)=\begin{bmatrix}0\\ 3\end{bmatrix},$$

where

$S=\{1$ (loss state), 2 (profit state)$\}$,

$U=\{1$ (select analyst 1), 2 (select analyst 2)$\}$,

$Y=\{1$ (analyst reports loss state), 2 (analyst reports profit state)$\}$.

Conditions 2 and 3 clearly are satisfied for this problem.

It turns out that the IHU optimal policy is:

$$u(k)=\psi(\eta(k))=\begin{cases}1 & \text{if } \eta_1(k)<.1129,\\ 2 & \text{if } \eta_1(k)>.1129,\\ \text{arbitrary} & \text{if } \eta_1(k)=.1129.\end{cases}$$

Furthermore, the optimal decision may be determined on the basis of the past two input–output pairs alone, according to the rule:

$$u(k)=\begin{cases}1 & \text{if } u(k-2)=y(k-1)=u(k-1)=y(k)=2,\\ 2 & \text{otherwise.}\end{cases}$$

Accordingly, v_* is piecewise linear with three faces. The number of faces of v_* generally equals the number of memory states required to realize ψ [28].

Example 4: *A problem where v_* is necessarily discontinuous.* Let $N=2$, $U=\{1, 2\}$, $Y=\{0\}$, and $P(0|1)=I$, $P(0|2)=\begin{bmatrix}1 & 0\\ 1 & 0\end{bmatrix}$,

$$R[i, u, y, j]=\begin{cases}1 & \text{if } i=u=1,\\ 0 & \text{otherwise.}\end{cases}$$

Since Y contains a single element, the observations convey no useful information, and we are confronted with an open-loop control problem. We may readily show that Condition 2* is satisfied with $S^*=\{1\}$ and $\xi=1$. Condition 3 is not satisfied since the choice $u(k)=1$ for all k causes $P(z(k))=I$, and $a[I]=1$. Now $g=1$ and any solution of (3.1) takes the form

$$v_*(\pi)=\begin{cases}\phi & \text{if } \pi=(1, 0),\\ \phi-1, & \text{otherwise.}\end{cases}$$

Thus every solution of (3.1) is discontinuous.

Example 5: *A problem in which Condition* 1 *is not satisfied.* Consider a system in which learning and rewards are mutually exclusive:

$$U = Y = \{1, 2, 3\}, \qquad N = 2,$$

$$P(3|1) = P(3|2) = I,$$

$$P(1|3) = \begin{bmatrix} p & 0 \\ 0 & 1-p \end{bmatrix}, \qquad P(2|3) = \begin{bmatrix} 1-p & 0 \\ 0 & p \end{bmatrix},$$

$$R[i, u, y, j] = \begin{cases} 1 & \text{if } u = i, \\ 0 & \text{otherwise.} \end{cases}$$

In this system, the state never changes. Input 3 permits state identification (learning). Inputs 1 and 2 are "state guesses" in which the "no information" output, 3, is observed. The optimal performance is $g = 1$, achieved by selecting input 3 infinitely often, but with vanishing frequency. We may show by contradiction that no solution to (3.1) exists: if v_* satisfies (3.1) and π has strictly positive entries, then $\psi(\pi) = 3$ since $T(\pi, (\cdot, 3)) = \pi$ but $\pi q(\cdot) \neq g = 1$; but this implies that the optimal CS selects input 3 at *all times*, in which case $g = 0$. This is, of course, the standard paradox of IHU dual control. Similar difficulties arise in the "two-armed bandit problem" [9] and may be resolved by resorting to IHU formulations other than (3.1) [8].

Appendix A. A metric associated with Condition 3. Define

$$D[\pi, \pi'] = \max\{d[\pi, \pi'], d[\pi', \pi]\}, \tag{A.1}$$

where

$$d[\pi, \pi'] = 1 - \min\{\pi_i/\pi'_i \colon \pi'_i > 0\}. \tag{A.2}$$

The function D has a number of remarkable properties. It is a metric on Π, closely related to the metric used in [22] to establish ergodic properties of $\{\eta(k)\}$. In the metric topology induced by D, Π is separated into the disconnected subsets $\Pi(A)$ given by (3.10). Thus D is discontinuous with respect to conventional metrics on Π exactly where a convex function may be discontinuous in the conventional sense. This enables us to prove

LEMMA A.1. *Any convex function v on Π satisfies*

$$|v(\pi) - v(\pi')| \leqq D[\pi, \pi']\,|v|.$$

Proof. Assume without loss of generality that $v(\pi) > v(\pi')$, and let $\pi'' = \pi' + (D[\pi, \pi'])^{-1}(\pi - \pi')$. Now $\pi = (1 - D[\pi, \pi'])\pi' + D[\pi, \pi']\pi''$ and the desired result follows from the convexity of v, provided only that $\pi'' \in \Pi$. Clearly, $\Sigma_{i \in S}\, \pi''_i = 1$. So it remains to show that $\pi''_i \geqq 0$, $\forall i \in S$. If $\pi_i \geqq \pi'_i$, then $\pi_i - \pi'_i \geqq 0$ and so $\pi''_i = \pi'_i + (D[\pi, \pi'])^{-1}(\pi_i - \pi'_i) \geqq 0$. If $\pi_i < \pi'_i$, then $D[\pi, \pi'] \geqq 1 - (\pi_i/\pi'_i)$ and so $\pi''_i \geqq \pi'_i + [1 - (\pi_i/\pi'_i)]^{-1}(\pi_i - \pi'_i) = 0$. □

We may use Lemma A.1 to obtain an expression involving $a[\cdot]$. It shows that if z is a string of most recent input–output pairs, then $a[P(z)]$ may be used to bound the value of knowing what occurred before z.

LEMMA A.2. *Let v be a convex function on Π. Then*

$$|v(T(\pi, z)) - v(T(\pi', z))| \leqq a[P(z)]\,|v|, \qquad \pi P(z) \neq 0, \quad \pi' P(z) \neq 0.$$

Proof. From the definition (A.1)–(A.2) of D, it is clear that

$$D[\pi, \lambda\pi' + (1-\lambda)\pi''] \leqq \max\{D[\pi, \pi'], D[\pi, \pi'']\}, \tag{A.3}$$

$$\pi, \pi', \pi'' \in \Pi, \quad 0 \leqq \lambda \leqq 1.$$

From (2.15), we obtain

$$\text{(A.4)} \qquad T(\pi, z) = \sum_{i \in \{i : e^i P(z) \neq 0\}} [\pi_i (e^i P(z)\nu)(\pi P(z)\nu)^{-1}] T(e^i, z), \qquad \pi P(z) \neq 0.$$

Combining (A.3) and (A.4) with the definition (4.7) of $a[\cdot]$ yields

$$\text{(A.5)} \qquad D[T(\pi, z), T(\pi', z)] \leqq a[P(z)], \qquad \pi P(z) \neq 0, \quad \pi' P(z) \neq 0.$$

The desired result follows immediately from Lemma A.1. □

Appendix B. Subrectangular matrices. A substochastic matrix $P = [P_{ij}]$ is said to be *subrectangular* if

$$\text{(B.1)} \qquad P_{ij} > 0 \text{ and } P_{i'j'} > 0 \Rightarrow P_{ij'} > 0 \text{ and } P_{i'j} > 0.$$

Let P be a subrectangular matrix and define:

$$\text{(B.2)} \qquad \begin{aligned} I(P) &= \{i : P_{ij} > 0, \text{ some } j\}, \\ J(P) &= \{j : P_{ij} > 0, \text{ some } i\}. \end{aligned}$$

Then we may easily verify that

$$\text{(B.3)} \qquad \begin{aligned} \{j : P_{ij} > 0\} &= \begin{cases} J(P) & \text{if } i \in I(P), \\ \varnothing & \text{if } i \notin I(P), \end{cases} \\ \{i : P_{ij} > 0\} &= \begin{cases} I(P) & \text{if } j \in J(P), \\ \varnothing & \text{if } j \notin J(P). \end{cases} \end{aligned}$$

Using (B.3), we obtain three fundamental properties of subrectangular matrices required in § 4.

LEMMA B.1. *$a[P] < 1 \Leftrightarrow P$ is subrectangular.*

LEMMA B.2. *The product of any matrix with a subrectangular matrix is subrectangular.*

LEMMA B.3. *If P is subrectangular, then $\{T(\pi, z) : \pi \in \Pi, \pi P(z) \neq 0\}$ is a closed subset of $\Pi(J(P(z)))$, where $\Pi(\cdot)$ is defined by* (3.10).

For a more detailed discussion of subrectangular matrices and their applications, see [22].

Acknowledgments. I am grateful to Alvin W. Drake and to Sanjoy K. Mitter, my dissertation supervisors at MIT, for their continuing encouragement, and to Wolfgang Runggaldier of the University of Padua, and Chelsea C. White of the University of Virginia, who carefully reviewed earlier drafts of this paper and provided valuable technical suggestions.

REFERENCES

[1] K. J. ASTROM, *Optimal control of Markov processes with incomplete state information*, J. Math. Anal. Appl., 10 (1965), pp. 174–205.

[2] ———, *Optimal control of Markov processes with incomplete state information II: The convexity of the loss function*, Ibid., 26 (1969), pp. 403–406.

[3] M. ATHANS, *The Role and Use of the Stochastic Linear-Quadratic-Gaussian Problem in Control System Design* (Special issue on linear-quadratic-Gaussian problem), IEEE Trans. Automatic Control, AC-16 (1971), pp. 529–552.

[4] Y. BAR-SHALOM AND E. TSE, *Dual effect, certainty equivalence and separation in stochastic control*, IEEE Trans. Automatic Control, AC-19 (1974), pp. 494–500.

[5] R. BELLMAN, *A Markovian decision process*, J. Math. and Mech., 6 (1957), pp. 679–684.

[6] D. BERTSEKAS, *Dynamic Programming and Stochastic Control*, Academic Press, New York, 1976.

[7] D. M. BROOKS AND C. T. LEONDES, *Markovian decision processes with state-information lag*, Operations Res., 21 (1973), pp. 904–907.

[8] T. M. COVER AND M. E. HELLMAN, *The two-armed-bandit problem with time-invariant finite memory*, IEEE Trans. Information Theory, IT-16 (1970), pp. 185–195.

[9] M. H. DEGROOT, *Optimal Statistical Decisions*, McGraw-Hill, New York, 1970.

[10] C. DERMAN, *Finite State Markovian Decision Processes*, Academic Press, New York, 1970.

[11] A. W. DRAKE, *Observation of a Markov process through a noisy channel*, SC.D. thesis, Dept. of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, MA, 1962.

[12] A. A. FELDBAUM, *Optimal Control Systems*, Academic Press, New York, 1965.

[13] J. FLYNN, *Conditions for the equivalence of optimality criteria in dynamic programming*, Ann. Statist., 4 (1976), pp. 936–953.

[14] N. A. J. HASTINGS, *Dynamic Programming with Management Application*, Butterworth, London and Crane–Russak, New York, 1973.

[15] R. A. HOWARD, *Dynamic programming and Markov processes*, MIT Press, Cambridge, MA, 1960.

[16] ———, *Dynamic Probabilistic Systems, Vols. I and II*, Wiley, New York, 1971.

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

[18] H. J. KUSHNER, *Introduction to Stochastic Control*, Holt, Rinehart and Winston, New York, 1971.

[19] H. MINE AND S. OSAKI, *Markovian Decision Processes*, Academic Press, New York, 1970.

[20] A. PAZ, *Introduction to Probabilistic Automata*, Academic Press, New York, 1970.

[21] L. K. PLATZMAN, *Improved conditions for convergence in undiscounted Markov renewal programming*, Operations Res., 25 (1977), pp. 529–533.

[22] ———, *Stability of recursive state estimators—The finite-state case*, to appear.

[23] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[24] S. M. ROSS, *Arbitrary state Markovian decision processes*, Ann. Math. Stat., 6 (1968), pp. 2188–2122.

[25] ———, *Applied probability models with optimization applications*, Holden-Day, San Francisco, CA, 1970.

[26] P. J. SCHWEITZER, *Iterative solution of the functional equations of undiscounted Markov renewal programming*, J. Math. Anal. Appl., 34 (1971), pp. 495–501.

[27] R. D. SMALLWOOD AND E. J. SONDIK, *The optimal control of partially-observable Markov processes over a finite horizon*, Operations Res., 21 (1973), pp. 1071–1081.

[28] E. J. SONDIK, *The optimal control of partially-observable Markov processes*, Ph.D. thesis, Stanford Information Systems Laboratory Tech. Rep. 6252–4, Stanford Univ., Stanford, CA, 1971.

[29] ———, *The optimal control of partially-observable Markov processes over the infinite horizon: Discounted costs*, Operations Res., 26 (1978), pp. 282–304.

[30] C. T. STRIEBEL, *Sufficient statistics in the optimum control of stochastic systems*, J. Math. Anal. Appl., 12 (1965), pp. 576–592.

[31] ———, *Optimal Control of Discrete-Time Stochastic Systems*, Springer-Verlag, New York, 1975.

[32] C. C. WHITE, *Monotone control laws for noisy countable-state Markov chains*, to appear.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 4, July 1980

0363–0129/80/1804–0005 $01.00/0

SCHAUDER DECOMPOSITIONS, APPROXIMATIONS AND CONTROL PROBLEMS*

HIDEAKI KANEKO† AND WILLIAM H. RUCKLE†

Abstract. A *Schauder decomposition* for a Banach space X is a sequence $\{P_n\}$ of finite rank continuous projections such that (a) $P_nP_m = P_mP_n = P_{\min\{m,n\}}$ and (b) $\lim_n P_nx = x$ for each x in X. Schauder decompositions can be used to approximate the solution to optimal control problems defined on X. For example, let S and T denote continuous linear operators from X into itself; let u be a point in the range of S and let p be a continuous seminorm on X. The problem:

(I) find x (and c) in X such that (a) $S(x) = u$, (b) $c = x - Tx$, (c) $p(c)$ is a minimum,

can be discretized to the problem:

(II) find x_n (and c_n) in the range of P_n such that (a) $Sx_n = P_nu$, (b) $c_n = x_n - P_nTx_n$, (c) $p(c_n)$ is a minimum.

We discuss conditions under which the minima found in solving (II) converge to the minimum in (I) as $n \to \infty$. Then we illustrate our theory by computing approximate solutions to the problem:

(III) find functions x (and c) such that (a) $x(t)$ is given for t in $[0, \frac{1}{3}) \cup [\frac{2}{3}, 1]$, (b) $c(t) = x(t) - \int_0^t x(s)\, ds$, (c) $\int_0^1 |c(t)|^2\, dt$ is a minimum.

Introduction. Schauder bases and decompositions are classic means of approximations in the abstract theory of Banach spaces. A reasonable question is whether these theoretic devices can be of service in solving practical problems. In studying the literature the authors have not found explicit reference to the theory of Schauder bases in the treatment of concrete examples. For example, even though the author of the huge compendium [4] has a well-known interest in the theory of best approximations, this work contains no applications of Schauder bases to practical approximation. On the other hand, many authors use Schauder decompositions as approximation schemes without recording this fact. For instance, in the paper [1] the arguments essentially rest upon properties of Schauder bases and decompositions. The trapazoidal rule for integration can be described in terms of Schauder's original basis of $C[0, 1]$. In this paper, we offer further evidence for a positive answer to the above question in a case where the approximation problem is more complicated. We preface our main discussion (§ 2) with a basic theory of approximation in a Banach space with a Schauder decomposition (§ 1). In § 2, we describe an optimal control problem which is essentially to drive a function from an initial state to a terminal state in such a way as to minimize a given functional. We then show how this problem can be embedded in a Banach space having a tailor made Schauder basis. We also derive formulas for the matrix which occurs in the discretized problem. In the final section (§ 3), we present the results of numerical calculations for two typical problems. Our main intention for including these calculations is to demonstrate the relevancy of the preceeding theory. The numerical approximation which we use is indeed crude, but it is clean cut and seems appropriate to the problems. For the first of these problems we compare the approximate solutions with an exact solution obtained by means of the calculus of variations.

1. Approximation in a Banach space having a Schauder decomposition. The purpose of this section is to develop a basic theory of approximation in a Banach space

* Received by the editors March 28, 1979, and in revised form October 15, 1979.

† Department of Mathematical Sciences, Clemson University, Clemson, South Carolina 39631.

having a Schauder decomposition. The theory which we develop is somewhat rudimentary but most useable as we shall illustrate in § 2 and § 3.

DEFINITION 1.1. A *Schauder decomposition* for a normed linear space X is a sequence $\{P_n\}$ of continuous finite rank projections from X into X such that

$$\text{(a)} \qquad P_nP_m = P_mP_n = P_{\min\{m,n\}},$$

$$\text{(b)} \qquad \lim_n P_nx = x \quad \text{for } x \text{ in } X.$$

The usual definition of Schauder decomposition does not require that each P_n have finite rank [3]. Condition (a) implies that $P_n(X) \subset P_m(X)$ if $n \leqq m$. The concept of Schauder decomposition is a generalization of the concept of a Schauder basis; see [3] for a discussion. For the purposes of approximation, the actual basis expansion may be less convenient than some other vector basis of the range space $P_n(X)$. This is illustrated in § 2.

THEOREM 1.2. *Let X be a Banach space with a Schauder decomposition $\{P_n\}$, let C be a closed and bounded subset of X such that $P_n(C) \subset C$ for each n, and let z be a point in $X \backslash C$.*

(a) *For each n there exists an element x_n of best approximation of z in $P_n(C)$.*

(b) *If x_n is as in* (a) *and x is the weak limit of any subsequence of $\{x_n\}$, then x is an element of best approximation of z by means of elements of C.*

Proof. (a) The set $P_n(C)$ is bounded and closed in the finite dimensional space $P_n(X)$. To verify that $P_n(C)$ is closed, suppose each y_k is in $P_n(C)$ and $y_k \to y$ in X. Since C is closed and $P_n(C) \subset C$, y is in C. Moreover, $P_ny = \lim_k P_ny_k = \lim_k y_k = y$ so that y is in $P_n(C)$. This implies $P_n(C)$ is compact. This means there exists a point x_n in $P_n(X)$ on which the continuous function $q(x) = \|z - x\|$ attains its minimum.

(b) Let $\{x_{n_k}\}$ be a subsequence of $\{x_n\}$ which converges weakly to the point x. The sequence $\{x_{n_k} - z\}$ converges weakly to $\{x - z\}$ so that by Lemma II. 3.27 of [1]

$$\begin{aligned} \|z - x\| &\leqq \liminf_k \|z - x_{n_k}\| \\ &= \inf_n \|z - x_n\|. \end{aligned} \tag{1.1}$$

The last equality holds since $\{\|z - x_n\|\}$ is a decreasing sequence. From (1.1) we conclude that

$$\|z - x\| \leqq \|z - y\| \tag{1.2}$$

for y in $\bigcup_n P_n(C)$. But $\bigcup_n P_n(C)$ is dense in C by Definition 1.1(b). This means that (1.2) holds for each y in C.

The condition $P_n(C) \subset C$ is essential to guarantee that x_n is not closer to z than $\inf\{\|x - z\|: x \text{ in } C\}$. It may seem at first that this condition is contrived and diminishes the applicability of Theorem 1.2. In the sequel we shall illustrate that we can tailor the Schauder decomposition to suit the problem in such a way that this condition is satisfied.

COROLLARY 1.3. *Suppose X is a reflexive Banach space with a Schauder decomposition $\{P_n\}$. Let C be a weakly closed and bounded subset of X such that $P_n(C) \subset C$ for each n. Then for each z in $X \backslash C$, there exists an element of best approximation for z in C.*

Proof. Since X is reflexive, and C is a weakly closed and bounded subset of X, C must be weakly sequentially compact [2, p. 430]. Hence, the sequence $\{x_n\}$ obtained in (a) of Theorem 1.2 has a subsequence which weakly converges to a point in C. By Theorem 1.2 (b) this point is an element of best approximation of z in C.

THEOREM 1.4. *Suppose X be a Banach space with a Schauder decomposition $\{P_n\}$, and p is a continuous seminorm on X. Let D be a bounded closed convex subset of X such that $\overline{\bigcup_n P_n(D)} = D$. If x_n in $P_n(D)$ is such that $p(x_n) = \inf\{p(y)\colon y \in P_n(D)\}$ then $\{p(x_n)\}$ decreases to* $\inf\{p(x)\colon x \in D\}$.

Proof. Since $P_n(D) \subset D$ for each n it follows from Definition 1.1(a) that $P_mP_n(D) = P_m(D) \subset P_n(D)$ if $m \leqq n$. This means that $\{p(x_n)\}$ is a decreasing sequence. If $d = \lim_n p(x_n)$ then $d \leqq p(y)$ for y in $\bigcap_n P_n(D)$ so that $d \leqq p(y)$ for y in $\overline{\bigcup_n P_n(D)} = D$.

COROLLARY 1.5. *Suppose X is a Banach space with a Schauder decomposition $\{P_n\}$, and D is a* compact *subset of X such that $\bigcup_n P_n(D) = D$. If $\{x_n\}$ is a sequence determined by Theorem* 1.4, *there is a subsequence of $\{x_n\}$ which converges to a point x_0 of D for which $p(x_0) = \inf\{p(x)\colon x \in D\}$.*

2. A control problem and its finite dimensional reduction. In this section, we formulate a control problem and suggest a means to solve it numerically by means of Schauder decompositions. This control problem arises in connection with a typical feedback system illustrated in Fig. 2.1. In the system illustrated, the input $x(t)$ at time t is the sum of the output $Tx(t)$ and a control function $c(t)$. We can represent this algebraically by the equation

$$x = Tx + c.$$

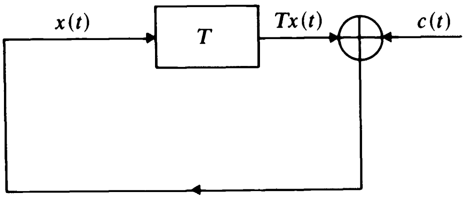


FIG. 2.1

The output of the system $Tx(t)$ is usually considered to be determined by the state of x from time 0 to time t. For example, one might describe T in terms of an integral operator

$$Tx(t) = \int_0^t A(t, s)x(s)\,ds,$$

where $A(t, s)$ is continuous in t and s. The purpose of the control c could be to drive x from an initial state

$$x(t) = u_1(t), \qquad 0 \leqq t \leqq \tfrac{1}{3}$$

to a finite state

$$x(t) = u_2(t), \qquad \tfrac{1}{3} \leqq t \leqq 1$$

at the least possible cost $p(c)$.

We first consider the control problem from an entirely functional analytic viewpoint. Let X denote a Banach space with a Schauder decomposition $\{P_n\}$, and let S and T denote continuous linear operations from X into itself. Let u be an arbitrary but fixed point in the range of S, and let p be a continuous seminorm on X. We then pose the problem

PROBLEM 2.1. Find x and c in X such that

$$\begin{aligned} &\text{(a)} \quad Sx = u, \\ &\text{(b)} \quad c = x - Tx, \\ &\text{(c)} \quad p(x) \text{ is a minimum.} \end{aligned}$$

In a sequel, we shall show that the motivating control problem is indeed a special case of Problem 2.1. The formal resemblance should be clear. Our present task will be to discretize Problem 2.1 by means of the Schauder decomposition of X and to determine the value of solutions of the discretized problem as approximations of the solution of Problem 2.1. The discretized problem is the following

PROBLEM 2.2. Find x_n and c_n in $R(P_n)$, the range of P_n such that

$$\begin{aligned} &\text{(a)} \quad Sx_n = P_n u \\ &\text{(b)} \quad c_n = x_n - P_n T x_n \\ &\text{(c)} \quad p(c_n) \text{ is a minimum.} \end{aligned}$$

Consider this question: if $\{x_n, c_n\}$ is a solution to Problem 2.2 in $R(P_n)$ how good is it as an approximation to the solution of Problem 2.1? Of course, since $\{P_n u\}$ converges to u, we can approximate u as closely as we like. But how does $p(c_n)$ compare with m given by

$$m = \inf\{p(c) \colon c = x - Tx,\ Sx = u\}?$$

For each $n = 1, 2, \cdots$, let

$$r_n = \inf\{p(c) \colon c = x - P_n Tx,\ x \in R(P_n),\ Sx = P_n u\}.$$

THEOREM 2.3. *If each P_n commutes with S, then* $\limsup_n r_n \leqq m$.

Proof. Denote $I - T$ by Q. For x in $R(P_n)$

$$P_n Q x = P_n(I - T)x = P_n x - P_n T x = x - P_n T x.$$

Since $\lim_n P_n Q y = Q y$ for each y in X, the seminorm q defined by $q(y) = \sup_k p(P_k Q y)$ is continuous on X by the uniform boundedness principle. Given $\varepsilon > 0$, let y in X be such that $Sy = u$ and $p(Qy) m + \varepsilon/2$. Let N be such that for $n \geqq N$

$$q(P_n y - y) + p(P_n Q y - Q y) < \varepsilon/2.$$

For $n \geqq N$ we have $SP_n y = P_n S y = P_n u$ so that

$$\begin{aligned} r_n &\leqq p(P_n Q P_n y) \\ &\leqq p(Qy) + p(-Qy + P_n Q y) + p(-P_n Q y + P_n Q P_n y) \\ &\leqq p(Qy) + p(P_n Q y - Q y) + q(P_n y - y) \\ &< m + \varepsilon. \end{aligned}$$

Therefore, $\limsup_n r_n \leqq m$.

COROLLARY 2.4. *Suppose for $\varepsilon > 0$, x_n in $R(P_n)$ is such that $Sx_n = P_n u$ and $p(x_n - P_n T x_n) < r_n + \varepsilon$. If x is the limit of any subsequence of $\{x_n\}$ then $Sx = u$ and $p(x - Tx) \leqq m + \varepsilon$.*

Proof. We may assume that $\lim_n x_n = x$. It follows that $Sx = \lim_n S x_n = \lim_n P_n u =$

u. Since

$$\begin{aligned}\|Tx-P_nTx_n\| &\leqq \|Tx-P_nTx\|+\|P_nTx-P_nTx_n\|\\ &\leqq \|Tx-P_nTx\|+\left(\sup_n \|P_n\|\right)\|Tx-Tx_n\|,\end{aligned}$$

we conclude that $\lim_n P_nTx_n = Tx$ and

$$\begin{aligned}p(x-Tx) &= \lim_n p(x_n-P_nTx_n)\\ &\leqq \limsup_n r_n+\varepsilon\\ &\leqq m+\varepsilon.\end{aligned}$$

COROLLARY 2.5. *Suppose for each $n=1,2,\cdots, x_n$ in $R(P_n)$ is a solution to Problem* 2.2. *If x is the limit of any subsequence of $\{x_n\}$, then x is a solution of Problem* 2.1.

Our next theorem gives conditions on the operators S and T under which the approximations converge.

THEOREM 2.6. *Suppose* (i) *T is a compact linear operator,* (ii) *S commutes with each P_n,* (iii) *there is $M>0$ such that for each $\varepsilon>0$ and $n=1,2,\cdots$, one can find x_n in $R(P_n)$ with $Sx_n=P_nu$, $\|x_n\|\leqq M$ and $p(x_n-P_nTx_n)\leqq r_n+\varepsilon$. Then* $\lim_n r_n=m$.

Proof. Since T is compact, $\lim_n (I-P_n)T=0$ in the uniform operator topology. Let v be any vector such that $Sv=u$. Given $\varepsilon>0$, let N be such that

$$\begin{aligned}&\sup\{p((I-P_n)Ty): \|y\|\leqq M\}\leqq \varepsilon/4,\\ &p((I-P_n)v)<\varepsilon/4,\\ &p(T(I-P_n)v)<\varepsilon/4\end{aligned}$$

for all $n\geqq N$. The first inequality holds since T is compact, the second and third since $\lim_n (I-P_n)v=0$. Given $n\geqq N$, let x_n in $R(P_n)$ be such that $Sx_n=P_nu$, $\|x_n\|\leqq M$ and $p(x_n-P_nTx_n)\leqq r_n+\varepsilon/4$. Then $S(x_n+(I-P_n)v)=u$ so that

$$\begin{aligned}m &\leqq p(x_n+(I-P_n)v-T(x_n+(I-P_n)v))\\ &\leqq p(x_n-P_nTx_n)+p((I-P_n)Tx)+p((I-P_n)v)+p(T(I-P_n)v)\\ &\leqq r_n+\varepsilon.\end{aligned}$$

Therefore, we conclude $m\leqq \liminf_n r_n$. This, together with Theorem 2.3 shows $\lim_n r_n=m$.

THEOREM 2.7. *If T is compact, and S commutes with each P_n, then for each $\varepsilon>0$ we can find N such that $\|c-(x-Tx)\|<\varepsilon\|x\|$ whenever c and x are solutions of Problem* 2.2 *with $n\geqq N$.*

Proof. Let N be such that $\|T-P_nT\|<\varepsilon$ for $n\geqq N$. If x and c are solutions to Problem 2.2 in $R(P_n)$, we have

$$\begin{aligned}\|c-(x-Tx)\| &= \|c-(x-P_nTx)+Tx-P_nTx\|\\ &= \|Tx-P_nTx\|\\ &\leqq \varepsilon\|x\|.\end{aligned}$$

We now return to a concrete version of the control problem described at the beginning of the section, and analyze it by constructing an appropriate Banach space with a Schauder decomposition. Let $A(t, s)$ be a continuous function from $[0, 1]\times[0, 1]$

into $[0, 1]$. Let $u_1(t)$ be a continuous function defined on $[0, \frac{1}{3})$ and $u_2(t)$, a continuous function on $[\frac{2}{3}, 1]$.

PROBLEM 2.8. Find functions x and c such that

$$\text{(a)} \qquad x(t) = \begin{cases} u_1(t), & 0 \leqq t < \frac{1}{3}, \\ u_2(t), & \frac{2}{3} \leqq t \leqq 1, \end{cases}$$

$$\text{(b)} \qquad c(t) = x(t) - \int_0^t A(t, s)x(s)\, ds,$$

$$\text{(c)} \qquad \left(\int_0^1 |c(t)|^p\, dt\right)^{1/p} \text{ is a minimum } (p \geqq 1).$$

In order to treat Problem 2.8 after the manner of Problem 2.1, we define a Banach space Z with an underlying Schauder decomposition with the properties needed for applying the appropriate theorems. For $n = 1, 2, \cdots$ let H_n denote the set of 3^n functions $H_{n1}, H_{n2}, \cdots, H_{n3^n}$ defined on $[0, 1]$ by

$$H_{ni}(t) = \begin{cases} 1 & \text{for } t \text{ in } \left[\frac{i-1}{3^n}, \frac{i}{3^n}\right), \\ 0 & \text{otherwise} \end{cases}$$

$(i = 1, 2, \cdots, 3^n)$ except that $H_{n3^n}(1)$ is 1 instead of 0. Let Z denote the closed linear span of $\bigcup_n H_n$ in the space $B[0, 1]$ of functions bounded on $[0, 1]$. Here $B[0, 1]$ has the uniform norm

$$\|x\| = \sup \{|x(t)|: 0 \leqq t \leqq 1\}.$$

By a uniform continuity argument, it follows that Z contains $C[0, 1]$ as well as the function $u(t)$ defined to be $u_1(t)$ for $0 \leqq t \leqq \frac{1}{3}$, $u_2(t)$ for $\frac{2}{3} \leqq t \leqq 1$, and 0 otherwise, where u_1 and u_2 are the functions in Problem 2.8. For x in Z we define

$$P_n x(t) = \sum_{i=1}^{3^n} x\left(\frac{i-1}{3^n}\right) H_{ni}(t).$$

Then P_n is a projection from Z onto the linear span of H_n in Z. We shall verify that $\{P_n\}$ is a Schauder decomposition of Z. If $n \geqq m$, $P_n H_{mi} = H_{mi}$ so that $\lim_n P_n H = H$ for each H in $\bigcup_n H_n$. Since $\|P_n x\| \leqq \|x\|$ for each x in Z, it follows from the Banach–Steinhaus theorem [2, p. 60] that $\lim_n P_n x = x$ for all x in Z. This verifies Definition 1.1(b). To verify Definition 1.1(a), note that if $m < n$

$$\begin{aligned} (P_m P_n x) &= P_m\left(\sum_{i=1}^{3^n} x\left(\frac{i-1}{3^n}\right) H_{ni}\right) \\ &= \sum_{i=1}^{3^n} x\left(\frac{i-1}{3^n}\right) P_m(H_{ni}). \end{aligned}$$

Now $P_m(H_{ni})$ is 0 unless $(i-1)/3^n$ is $(j-1)/3^m$ for some j in which case $P_m(H_{ni})$ is H_{mj}. Therefore, we conclude that

$$P_m P_n x = \sum_{j=1}^{3^m} x\left(\frac{j-1}{3^m}\right) H_{mj} = P_m x.$$

On the other hand,

$$P_nP_mx = \sum_{i=1}^{3^m} x\left(\frac{i-1}{3^m}\right)P_n(H_{mi})$$
$$= \sum_{i=1}^{3^m} x\left(\frac{i-1}{3^m}\right)H_{mi} = P_mx.$$

We can now define Sx, Tx and $p(x)$ for x in Z by

$$Tx(t) = \int_0^t A(t, s)x(s)\,ds,$$
$$Sx(t) = \begin{cases} x(t), & t \text{ in } [0, \frac{1}{3}) \cup [\frac{2}{3}, 1], \\ 0, & t \text{ in } [\frac{1}{3}, \frac{2}{3}), \end{cases}$$
$$p(x) = \left(\int_0^1 |c(t)|^p\,dt\right)^{1/p}.$$

Obviously $SP_n = P_nS$ for each n, so if we add the condition "Find x and c in Z" to Problem 2.8, we can apply Theorem 2.3. Problem 2.8 can then be discretized to the following problem in $R(P_n) = [H_n]$.

PROBLEM 2.9. Find x and c in $[H_n]$ such that

(a) $Sx = P_nu$,

(b) $c = x - P_nTx$,

(c) $\left(\int_0^1 |c(t)|^p\,dt\right)^{1/p}$ is a minimum.

If we set $c_i = c((i-1)/3^n)$, $x_i = x((i-1)/3^n)$ and $u_i = u((i-1)/3^n)$ for $i = 1, 2, \cdots, 3^n$, then Problem 2.9 assumes the following form.

PROBLEM 2.10. Find vectors $\mathbf{x}$ and $\mathbf{c}$ in R^{3^n} such that

(a) $x_i = a_i, \qquad i = 1, 2, \cdots, 3^{n-1}, 2\cdot 3^{n-1}+1, \cdots, 3^n$,

(b) $c = \mathbf{x} - A\mathbf{x}$,

(c) $\left(\sum_{i=1}^{3^n} |c_i|^p\right)^{1/p}$ is a minimum.

Here A is a $3^n \times 3^n$ matrix whose entries are computed as follows:

$$c_i = c\left(\frac{i-1}{3^n}\right)$$
$$= x\left(\frac{i-1}{3^n}\right) - Tx\left(\frac{i-1}{3^n}\right)$$
$$= x\left(\frac{i-1}{3^n}\right) - \int_0^{(i-1)/3^n} A\left(\frac{i-1}{3^n}, s\right)x(s)\,dx$$
$$= x\left(\frac{i-1}{3^n}\right) - \sum_{j=1}^{i-1} \int_{(j-1)/3^n}^{(i-1)/3^n} A\left(\frac{i-1}{3^n}, s\right)x\left(\frac{j-1}{3^n}\right)ds$$

since x being in $R(P_n)$ is actually a step function. Thus we have

$$c_i = x_i - \sum_{j=1}^{i-1} \left(\int_{(j-1)/3^n}^{j/3^n} A\left(\frac{i-1}{3^n}, s\right) ds \right) x_j.$$

This means that $A = (a_{ij})$, where

$$a_{ij} = \begin{cases} \displaystyle\int_{(j-1)/3^n}^{j/3^n} A\left(\frac{i-1}{3^n}, s\right) ds, & j \leqq i-1, \\ 0, & j > i-1. \end{cases}$$

We shall obtain some specific numerical solutions to Problem 2.10 in the next section.

The theory developed here also applies to Problem 2.9 with a delay namely the following.

PROBLEM 2.11. Find x and c such that

$$\text{(a)} \qquad x(t) = \begin{cases} u_1(t), & t \text{ in } [0, \frac{1}{3}), \\ u_2(t), & t \text{ in } [\frac{2}{3}, 1], \end{cases}$$

$$\text{(b)} \qquad c(t) = x(t) - \int_0^t A(t, s) x(s - \tfrac{1}{3}) \, ds,$$

$$\text{(c)} \qquad \left(\int_0^1 |c(t)|^p \, dt \right)^{1/p} \text{ is a minimum.}$$

In Problem 2.11, $x(t) = 0$ for $t < 0$. The only change needed to reduce Problem 2.11 to the form of Problem 2.9 is to redefine T by

$$Tx(t) = \int_0^t A(t, s) x(s - \tfrac{1}{3}) \, ds$$
$$= \begin{cases} 0 & \text{if } t \leqq \frac{1}{3}, \\ \displaystyle\int_{1/3}^{t-1/3} A(t, s + \tfrac{1}{3}) x(s) \, ds, & t > \frac{1}{3}. \end{cases}$$

In this case, the entries of the matrix A in Problem 2.10 are given by

$$a_{ij} = \begin{cases} 0 & \text{if } i \leqq 3^{n-1} \text{ for all } j, \\ \displaystyle\int_{(j-1)/3^n - 1/3}^{j/3^n - 1/3} A\left(\frac{i-1}{3^n}, s + \tfrac{1}{3}\right) ds & \text{if } i > 3^{n-1} \text{ and } 3^{n-1} < j \leqq i, \\ 0 & \text{if } j > i. \end{cases}$$

3. Two numerical examples. We shall now use the ideas developed in § 2 to approximate solutions to Problem 2.8 in two particular cases:

$$\text{(A)} \qquad x(t) = \begin{cases} 0, & 0 \leqq t < \frac{1}{3}, \\ 1, & \frac{2}{3} \leqq t \leqq 1, \end{cases}$$

$$p = 2, \quad A(s, t) = 1, \quad (s, t) \in I \times I,$$

$$\text{(B)} \qquad x(t) = \begin{cases} 1, & 0 \leqq t < \frac{1}{3}, \\ e^t, & \frac{2}{3} \leqq t \leqq 1, \end{cases}$$

$$p = 2, \quad A(s, t) = 1, \quad (s, t) \in I \times I.$$

In case (A), we were able to find an exact solution to Problem 2.8 by means of an arduous but unenlightening application of the calculus of variations. For the purpose of comparison, the functional $(\int_0^1 |c(t)|^2\,dt)^{1/2}$ is a minimum when $x(t) = Pe^{2/3-t} + Qe^t$, where P is approximatedly .142573 and Q is .110974. The minimum value of $(\int_0^1 |c(t)|^2\,dt)^{1/2}$ is .20189. We expect that in case (B) an exact solution will be much harder to find.

After reducing Problem 2.8 to Problem 2.10, we can further modify Problem 2.10 to put it in a form more suitable for computation. We want to minimize

$$\begin{aligned} p(\mathbf{c}) &= 3^{-n} \sum_{i=1}^{3^n} c_i^2 \\ &= 3^{-n}\mathbf{c}^T\mathbf{c} = 3^{-n}((I - \bar{A})\mathbf{x})^T((I - A)\mathbf{x}) \\ &= \mathbf{x}^T(I - A)^T(I - A)\mathbf{x} \\ &= \mathbf{x}^T Bx, \end{aligned}$$

where $B = (I - A)^T(I - A)$ is a symmetric $3^n \times 3^n$ matrix. In Problem 2.8, the first and last 3^{n-1} elements of the vector $\mathbf{x}$ are given. Hence the minimization problem is simply solving 3^{n-1} equations for 3^{n-1} elements which occupy the middle third of the vector $\mathbf{x}$.

We partition $\mathbf{x}$ and B thus

$$\mathbf{x} = \begin{bmatrix} \mathbf{a} \\ \mathbf{z} \\ \mathbf{b} \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} B_{11} & B_{12} & B_{13} \\ B_{21} & B_{22} & B_{23} \\ B_{31} & B_{32} & B_{33} \end{bmatrix},$$

where $\mathbf{a}$, $\mathbf{z}$, $\mathbf{b}$ are 3^{n-1} dimensional vectors and the submatrices B_{ij} are $3^{n-1} \times 3^{n-1}$ matrices, $i, j = 1, 2, 3$. We then have

$$\begin{aligned} \mathbf{c}^T\mathbf{c} &= \mathbf{x}^T B\mathbf{x} \\ &= \mathbf{z}^T(B_{21}\mathbf{a} + B_{23}\mathbf{b}) + (\mathbf{a}^T B_{12} + \mathbf{b}^T B_{32})z + \mathbf{z}^T B_{22}\mathbf{z} + R, \end{aligned}$$

where R does not depend on $\mathbf{z}$, i.e., is constant. Since B is symmetric we obtain

$$\mathbf{c}^T\mathbf{c} = 2(\mathbf{a}^T B_{12} + b^T B_{32})\mathbf{z} + \mathbf{z}^T B_{22}\mathbf{z} + R. \tag{3.1}$$

We then take the partial derivatives of $\mathbf{c}^T\mathbf{c}$ with respect to the components of $\mathbf{z}$ and set them equal to zero to obtain

$$B_{22}\mathbf{z} = -B_{12}\mathbf{a} - B_{32}\mathbf{b},$$

which we can solve for $\mathbf{z}$ using standard FORTRAN routines.

Numerical results for problems A and B are tabulated in Tables 1 and 2.

TABLE 1
Problem A

	$n=2$	$n=3$	$n=4$	Exact
$x(\frac{1}{3})$	.32110	.32993	.33315	.35385
$x(\frac{10}{27})$		.33037	.33345	.35246
$x(\frac{11}{27})$		.33124	.33420	.35155
$x(\frac{4}{9})$	.32467	.33255	.33541	.35113
$x(\frac{13}{27})$		.33430	.33707	.35119
$x(\frac{14}{27})$		.33650	.33918	.35172
$x(\frac{5}{9})$	.33184	.33914	.34176	.35275
$x(\frac{16}{27})$		.34222	.34480	.35425
$x(\frac{17}{27})$		.34576	.34830	.35624
$\min\,(\int_0^1 \|c(t)\|^2\,dt)^{1/2}$	.23393	.21315	.20636	.20189

TABLE 2
Problem B

	$n=3$	$n=4$
$x(\frac{1}{3})$	.82110	.80606
$x(\frac{10}{27})$	.83410	.81898
$x(\frac{11}{27})$	.84819	.83302
$x(\frac{4}{9})$	.86341	.84818
$x(\frac{13}{27})$	.87976	.86449
$x(\frac{14}{27})$	.89728	.88198
$x(\frac{5}{9})$	.91599	.90066
$x(\frac{16}{27})$	.93591	.92056
$x(\frac{17}{27})$	.95707	.94170
$\min(\int_0^1 \lvert c(t)\rvert^2\,dt)^{1/2}$	.71659	.65922

4. Remarks. As we stated in the Introduction, there are in the literature many examples of Schauder decomposition which are not identified as such. Since these examples include the most obvious and useful Schauder decompositions, we cannot claim that our present work contributes new and useful numerical techniques. For example, the Schauder decomposition introduced in [2] is simply a piecewise constant approximation scheme—the first type of approximation that one would consider.

On a more positive note, we point out that the theory developed in § 1 quickly shows that the given approximation scheme or those like it lead to a useable approximation of the solution to the control problems of § 2. This means that in order to attack such problems one need only verify that the scheme results from a Schauder decomposition or, even better, tailor a Schauder decomposition to suit the problem as we have done. Thus our work has the potential of (a) unifying diverse results on convergence under a single theory; (b) suggesting a way to attack problems like those in § 2 without working out a special theory of convergence.

Toward point (a) above, we shall note in a future paper that the nth degree polynomial approximation of a function form a Schauder decomposition of the Sobolev spaces W_n^p; $n = 1, 2, \cdots$; $1<p<\infty$, and exploit this fact to verify some known results on nth degree polynomial approximations.

Finally we note that the matrix problem solved in § 3 is a special case of a least squares problem. The method we used to solve it, is essentially Algorithm 2.10 given in Chapter 5 of [5]. Because of the special nature of the problem we were able to reduce the size of the matrix by two-thirds.

REFERENCES

[1] E. M. CLIFF AND J. A. BURNS, *A piecewise linear approximation scheme for hereditary optimal control problems*, Virginia Polytechnic Institute and State Univ. Blacksburg, VA, November 1976.

[2] N. DUNFORD AND J. T. SCJWARTZ, *Linear Operators. Part I: General Theory*, Interscience, New York, 1958.

[3] W. H. RUCKLE, *The infinite sum of closed subspaces of an F-space*, Duke Math. J., 31 (1964), pp. 543–554.

[4] I. SINGER, *Bases in Banach Spaces I*, Springer, New York, 1970.

[5] G. W. STEWART, *Introduction to Matrix Computations*, Academic Press, New York, 1973.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 4, July 1980

0363-0129/80/1804-0006 $01.00/0

WELL-POSEDNESS OF SOME EVOLUTION PROBLEMS IN THE THEORY OF AUTOMATIC FEED-BACK CONTROL FOR SYSTEMS WITH DISTRIBUTED PARAMETERS*

AART VAN HARTEN† AND HANS SCHUMACHER‡

Abstract. The subject of this paper is the evolution, in time, of systems of diffusion type controlled by an automatic feed-back mechanism using a finite number of observators and control-inputs. For the initial-boundary value problem for a functional partial differential equation of parabolic type, which the state variable has to satisfy, the following types of questions are considered: existence, uniqueness, regularity and continuous dependence on data of a solution. Answers to these questions are given in function spaces of Hölder and Sobolev type.

1. Introduction.

1.1. Purpose of this paper. In this paper we shall consider a system for which, in the uncontrolled situation, the behavior is described by an evolution equation of diffusion type ((1.1.1) with $\Pi \equiv 0$) together with initial and boundary conditions (1.1.2–3). In the controlled situation an automatic feed-back control mechanism is applied to this system. The effect of this control is accounted for by the term Πu in (1.1.1). Π is called the feed-back control operator.

$$\frac{\partial u}{\partial t} = (L+\Pi)u + f, \qquad \text{FPDE}, \tag{1.1.1}$$

$$Bu = \phi, \qquad \text{BC}, \tag{1.1.2}$$

$$u(\cdot, 0) = \psi, \qquad \text{IC}. \tag{1.1.3}$$

In the uncontrolled case a rather well-developed theory for existence, uniqueness, regularity and continuous dependence on (f, ϕ, ψ) of the state u is available.

In Ladyzenskaja, Solonnikov, Uraltseva (1967) such a theory is given using function spaces of Hölder type. We shall refer to this type of theory as "C-theory".

A theory using function spaces of Sobolev type can be found in Lions, Magenes (1972). To the latter type of theory we shall refer as "H-theory".

In the controlled case, the structure of the operator Π will be such that (1.1.1) is no longer a PDE but a FPDE, where "F" abbreviates "functional".

The purpose of this paper is to give a "C-theory" as well as a "H-theory" for the existence, uniqueness, regularity and continuous dependence on (f, ϕ, ψ) of a solution of (1.1.1-2-3) in the controlled case.

As for the structure of the operator Π we shall take an approach, in which, right from the beginning, we shall deal with control in feed-back form using only a finite number of observators and control inputs. This in contrast with the work of Lions (1971), where much more of an open loop approach for the control term is taken.

Now we shall first demonstrate, with an example, what type of automatic feed-back control mechanisms we have in mind in § 1.2. In § 1.3 we shall explain, in full generality, the setting in which we shall consider the problem (1.1.1-2-3). In § 1.4 we shall discuss the strategy to attack (1.1.1-2-3). Next we shall look at the concept of compatibility of the data in § 2. In § 3 we shall deal with a theorem that appears to be very useful for the derivation of our main results. Then these main results will be discussed in §§ 4 and 5. In

* Received by the editors November 28, 1978, and in revised form September 24, 1979.

† Mathematical Institute, University of Utrecht, Budapestlaan 6, Utrecht, The Netherlands.

‡ Dept. of Mathematics, Free University of Amsterdam, De Boelelaan 1081, Amsterdam, The Netherlands.

§ 6 the developed theory is applied to some examples. There we shall also come back to the example of § 1.2 for explicit calculation of some solutions.

1.2. An example taken from physics. Consider the distribution of temperature u in a plate D with an isolating boundary ∂D. Let g be the autonomous production/absorption of heat and let ψ be the initial distribution of temperature. Suppose, that in order to control the system, temperature is permanently observed in the points: $y_1 \in \bar{D}, \cdots, y_p \in \bar{D}$, $\bar{D} = D \cup \partial D$. This information is fed-back to a heating/cooling apparatus characterised by control input functions on $D: c_1, \cdots, c_q$ (see Fig. 1).

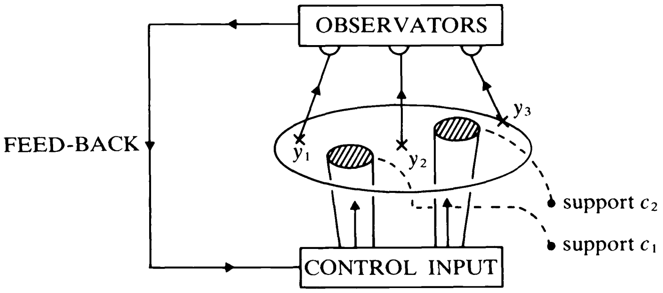


FIG. 1

Suppose that ideal values for the temperature in $y_1, \cdots, y_p$ at time $t \geqq 0$ are known to be $I_1(t), \cdots, I_p(t)$.

As for the feed-back we consider the following examples.

(i) Instantaneous feed-back: the control action producing/absorbing heat in a volume element dx at $x \in D$ during a time interval $(t, t+dt)$ is given by

$$\left[\sum_{i=1}^{q} \sum_{j=1}^{p} c_i(x, t) h_{ij}(t)(u(y_j, t) - I_j(t))\right] dx\, dt.$$

(ii) Feed-back with (simple) memory: the control action at x in dx during $(t, t+dt)$ is now given by

$$\left[\sum_{i=1}^{q} \sum_{j=1}^{p} c_i(x, t)\bar{h}_{ij}(t) \int_0^t \mu\, e^{-\mu(t-\tau)}(u(y_j, \tau) - I_j(\tau))\, d\tau\right] dx\, dt.$$

Here $\mu^{-1} > 0$ is the characteristic time for loss of memory.

A model for this controlled system is given by

(1.2.1) $$\frac{\partial u}{\partial t} = (\Delta + \Pi)u + f, \qquad \text{FPDE},$$

(1.2.2) $$\frac{\partial u}{\partial \vec{n}} = 0, \qquad \text{BC},$$

(1.2.3) $$u(\cdot, 0) = \psi, \qquad \text{IC}$$

with Δ the Laplacian, $\partial/\partial\vec{n}$ the normal derivative.

For the feed-back control operator we get

(1.2.4) $$\Pi = \sum_{i=1}^{q} \sum_{j=1}^{p} c_i F_{ij} \delta_{y_j}$$

with observators δ_{y_j} defined by

$$(1.2.5) \qquad (\delta_{y_j}u)(t) = u(y_j, t),$$

and with feed-back coupling operators F_{ij} defined by

$$(1.2.6) \qquad \begin{aligned} (F_{ij}\chi)(t) &= h_{ij}(t)\chi(t) && \text{in case (i)} \\ &= \bar{h}_{ij}(t)\int_0^t \mu\, e^{-\mu(t-\tau)}\chi(\tau)\, d\tau && \text{in case (ii).} \end{aligned}$$

The inhomogeneous term f in (1.2.1) is given by

$$(1.2.7) \qquad f = g - \sum_{i=1}^{q}\sum_{j=1}^{p} c_i F_{ij} I_j.$$

Note, that the operator Π of (1.2.4) has a nonlocal character in space. In case (ii), Π also has a delayed character in time. These facts cause (1.2.1) to be an equation of more complex type than a PDE.

1.3. Description of the general problem. Let us now discuss the general setting in which we consider the problem (1.1.1-2-3).

$u(x, t)$ will be the state of the system; x, the vector of space variables; $x \in \bar{D}$ and t, the time variable, $t \in [0, T]$. D will be a bounded, open domain $\subset \mathbb{R}^n$; $\bar{D}$ denotes the closure of D, $\partial D = \bar{D}\backslash D$. For simplicity we suppose, that ∂D is smooth. Let us introduce the notation: $Q = D \times (0, T)$, $\Gamma = \partial D \times (0, T)$. L will be a linear, second order, uniformly elliptic PDO:

$$(1.3.1) \qquad L = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}D_{ij} + \sum_{i=1}^{n} a_i D_i + a_0$$

with $D_i = \partial/\partial x_i$, $D_{ij} = \partial^2/\partial x_i \partial x_j$ and $\exists E > 0, \forall (x, t) \in \bar{Q}, \forall \xi \in \mathbb{R}^n : \sum_{i=1}^n \sum_{j=1}^n a_{ij}(x, t)\xi_i\xi_j \geqq E\sum_{i=1}^n \xi_i^2$. All coefficients of L are allowed to depend on x and t. For simplicity we suppose: $a_{ij}, a_i, a_0 \in C^\infty(\bar{Q})$, $1 \leqq i, j \leqq n$.

The operator B will be linear and of order ν with $\nu = 0$ or $\nu = 1$.

$$(1.3.2) \qquad \begin{aligned} B &= 1, && \text{Dirichlet BC,} \quad \nu = 0, \\ B &= \sum_{i=1}^{n} b_i D_i + b_0, && \text{Neumann BC,} \quad \nu = 1. \end{aligned}$$

The coefficients b_i, b_0, $1 \leqq i \leqq n$ are allowed to depend on x and t. For simplicity we suppose b_i, $b_0 \in C^\infty(\bar{\Gamma})$, $1 \leqq i \leqq n$. Furthermore we assume, that $\forall (x, t) \in \bar{\Gamma}$, $\sum_{i=1}^n b_i(x, t)\vec{n}_i(x) > 0$ with $\vec{n}(x)$ the outward directed normal on ∂D at x.

As for the regularity of the data f, ϕ, ψ we suppose

$$(1.3.3) \qquad \begin{array}{l|l} f \in C^{\beta,\beta/2}(\bar{Q}), & f \in H^{\beta,\beta/2}(Q), \\ \phi \in C^{\hat{\beta},\hat{\beta}/2}(\bar{\Gamma}), & \phi \in H^{\hat{\beta},\hat{\beta}/2}(\Gamma), \\ \psi \in C^{\beta_0}(\bar{D}), & \psi \in H^{\beta_0}(D), \\ \beta, \hat{\beta}, \beta_0 \geqq 0 & \beta, \hat{\beta}, \beta_0 \geqq 0. \end{array}$$

Here we used for the first time a form of specification which will often be used afterwards in the same way: at the left of the vertical bar the specification refers to

"C-theory", at the right of the vertical bar the specification refers to "H-theory", i.e.,

$$\text{“C-theory”}\ \Big|\ \text{“H-theory”}.$$

$C^{\beta,\beta/2}(\bar{Q})$ denotes the Hölder space of order β in the space directions and of order $\beta/2$ in the time direction. For the definition of $C^{\beta,\beta/2}(\bar{Q})$, $C^{\hat{\beta},\hat{\beta}/2}(\bar{\Gamma})$, $C^{\beta_0}(\bar{D})$ and their usual norms $|\cdot|_{\beta,\beta/2}$, $|\cdot|_{\hat{\beta},\beta/2}$, $|\cdot|_{\beta_0}$ we refer to Ladyzenskaja, Solonnikov, Uraltseva (1967) with the important remark that these spaces are indicated there with C replaced by H.

For the definition of the Sobolev spaces $H^{\beta,\beta/2}(Q)$, $H^{\hat{\beta},\hat{\beta}/2}(\Gamma)$ and $H^{\beta_0}(D)$, and their usual norms $\|\cdot\|_{\beta,\beta/2}$, $\|\cdot\|_{\hat{\beta},\beta/2}$, $\|\cdot\|_{\beta_0}$, we refer to Lions, Magenes (1972, vol. II, p. 6) and Adams (1975, p. 208).

From now on the notation β, $\hat{\beta}$, β_0 will be reserved for and will always refer to the indices indicating the regularity of the data as introduced in (1.3.3).

In addition to the regularity of the data it is important to consider the compatibility of f, ϕ, ψ. This is done in § 2.

Let us now describe the general form of the feed-back control operator Π. We shall take

$$\Pi = \sum_{i=1}^{q} \sum_{j=1}^{p} c_i F_{ij} P_j \tag{1.3.4}$$

with

control input functions c_i, $\quad 1 \leqq i \leqq q$;

observators P_j, $\quad 1 \leqq j \leqq p$;

feed-back coupling operators F_{ij}, $\quad 1 \leqq i \leqq q$, $1 \leqq j \leqq p$.

Further we suppose

(1.3.5)	$\Pi \in L(C^{\check{\gamma},\check{\gamma}/2}(\bar{Q}) \to C^{\gamma,\gamma/2}(\bar{Q}))$, $\check{\gamma} \geqq \gamma \geqq 0$	$\Pi \in L(H^{\check{\gamma},\check{\gamma}/2}(Q) \to H^{\gamma,\gamma/2}(Q))$, $\check{\gamma} \geqq \gamma \geqq 0$.

For Banach spaces X, Y we denote by $L(X \to Y)$ the space of bounded, linear operators from X into Y.

An important role will be played by

$$\alpha = \check{\gamma} - \gamma. \tag{1.3.6}$$

α is called the order of the feed-back control operator.

It will not be surprising that for our theory of well-posedness of the problem (1.1.1-2-3), we have to distinguish between the cases where the order of the feed-back control operator Π is smaller than, or larger than/equal to, the order of the diffusion operator L. In § 4 we shall deal with the case $0 \leqq \alpha < 2$ and in § 5 we consider the case $\alpha \geqq 2$.

Our requirements for P_j, F_{ij}, c_i, $1 \leqq i \leqq q$, $1 \leqq j \leqq p$ are specified in (1.3.7-8-9).

	$P_j \in L(C^{\check{\gamma},\check{\gamma}/2}(\bar{Q}) \to C^{\gamma/2}[0, T])$	$P_j \in L(H^{\check{\gamma},\check{\gamma}/2}(Q) \to H^{\gamma/2}(0, T))$
	and:	and:
(1.3.7)	$\exists \tilde{P}_j \in C^{\gamma/2}([0, T] \to C^{\alpha}(\bar{D})')$ such that	$\exists \tilde{P}_j \in H^{\gamma/2}((0, T) \to H^{\alpha}(D)')$ such that
	$\forall u \in C^{\check{\gamma},\check{\gamma}/2}(\bar{Q})$, $\forall t \in [0, T]$,	$\forall u \in H^{\check{\gamma},\check{\gamma}/2}(Q)$ a.e. in $t \in (0, T)$,
	$(P_j u)(t) = \tilde{P}_j(t) u(\cdot, t)$.	$(P_j u)(t) = \tilde{P}_j(t) u(\cdot, t)$.

By $C^\alpha(\bar{D})'$, $H^\alpha(D)'$ we denote the topological dual spaces of $C^\alpha(\bar{D})$, $H^\alpha(D)$ equipped with their usual norms.

Note that the contents of (1.3.7) are that P_j acts instantaneously with its value on u at time t given by $\tilde{P}_j(t)u(\cdot, t)$, where $\tilde{P}_j(t)$ possesses a prescribed regularity in t.

If $P_j : C^\infty(\bar{Q}) \to C[0, T]$ (this is not necessarily true in the case of H-theory if $0 \leqq \gamma \leqq 1$, but is implied by (1.3.7) in the other cases) then a consequence of the fact that P_j acts instantaneously as defined in (1.3.7) is: $\forall u, v \in C^\infty(\bar{Q})$, $\forall t \in [0, T]$, $u(\cdot, t) = v(\cdot, t) \Rightarrow (P_j u)(t) = (P_j v)(t)$.

$$(1.3.8) \qquad \begin{array}{l|l} F_{ij} \in L(C^{\gamma/2}[0, T]), & F_{ij} \in L(H^{\gamma/2}(0, T)), \\ F_{ij} \text{ nonanticipative} & F_{ij} \text{ nonanticipative.} \end{array}$$

For a Banach space X we abbreviate $L(X \to X)$ to $L(X)$.

Nonanticipativity of F_{ij} means, that for all elements ξ, η of $C^{\gamma/2}[0, T]$, $H^{\gamma/2}(0, T)$ and $\forall t \in (0, T)$:

$$\xi|_{(0,t)} = \eta|_{(0,t)} \Rightarrow F_{ij}\xi|_{(0,t)} = F_{ij}\eta|_{(0,t)}.$$

$$(1.3.9) \qquad c_i \in C^{\gamma,\gamma/2}(\bar{Q}) \;\Big|\; c_i \in H^{\gamma,\gamma/2}(Q).$$

The operator Π will be called of type (γ, α), if (1.3.5-6-7-8-9) are satisfied.

From now on the notation γ, α will be reserved for, and will always refer to, the type of the operator Π as introduced above. A natural question is of course whether the requirements (1.3.6-7-8-9) already imply (1.3.5).

In the case of "C-theory" this is indeed true. In the case of "H-theory" it is true if $\gamma > 1$ but not necessarily if $0 \leqq \gamma \leqq 1$! One of the reasons is, that in general, for $c_i \in H^{\gamma,\gamma/2}(Q)$ and $\xi \in H^{\gamma/2}(0, T)$ with $0 \leqq \gamma \leqq 1$, the function defined by $c_i(x, t)\xi(t)$ is not an element of $H^{\gamma,\gamma/2}(Q)$.

A sufficient supplement to (1.3.6-7-8-9) in the case of "H-theory" with $0 \leqq \gamma \leqq 1$, in order to imply (1.3.5), is

$$(1.3.10) \qquad \begin{array}{l} \underset{0<t<T}{\operatorname{ess\,sup}} \{\|c_i(\cdot, t)\|_\gamma + \|\tilde{P}(t)\|_{H^\alpha(D)'}\} < \infty \quad \text{if } \gamma \in [0, 1], \\ \text{and furthermore, if } \gamma > 0,\ \exists \bar{\delta} > \gamma/2 \text{ such that} \\ \underset{0<t<\tau<T}{\operatorname{ess\,sup}} |t-\tau|^{-\bar{\delta}}\{\|c(\cdot, t) - c(\cdot, \tau)\|_0 + \|\tilde{P}(t) - \tilde{P}(\tau)\|_{H^\alpha(D)'}\} < \infty. \end{array}$$

Note, that if in the example of § 1.2 we take $c_i \in C^\infty(\bar{Q})$, h_{ij}, $\bar{h}_{ij} \in C^\infty[0, T]$, then (1.3.5-6-7-8-9) are satisfied for any choice of $\gamma \in [0, \infty)$ with

$$(1.3.11) \qquad \alpha = 0 \;\Big|\; \alpha = \frac{n}{2} + \varepsilon, \quad \varepsilon > 0 \text{ arbitrarily small}$$

The choice of α in the case of "H-theory" is a consequence of the fact that for $s > n/2$, $H^s(D)$ is continuously imbedded in $C(\bar{D})$, see Adams (1975, p. 97).

It will be clear that the setting (1.3.5-6-7-8-9) for the control term is rather rich: it admits many examples quite different from the ones of § 1.2. The following additional examples may serve to demonstrate this.

Firstly it is allowed to take observators which use derivatives in space-directions and/or integration, averaging in space.

In the model of § 1.2 one can, for example, think of an observator $\tilde{P}$ that observes the flow of heat through the smooth boundary S of a sub-domain Ω of D. This leads to

$(\bar{P}u)(t)=\int_S(\partial u/\partial\vec{n})(x,t)\,dx$, where $\partial/\partial\vec{n}$ denotes the normal derivative in outward direction on S. Such an observator would fit in (1.3.7) with

$$\gamma\geqq 0,\quad \alpha=1 \;\Big|\; \gamma>0,\quad \alpha=\tfrac{3}{2}. \tag{1.3.12}$$

In the case of "H-theory" the choice of α is here a consequence of the theorem on traces on the boundary, see Lions, Magenes (1972, vol. II, p. 9).

Secondly it is allowed that the feed-back coupling uses time-derivatives in an essential way.

In the following example the feed-back coupling operator F maps the function on which it operates to a smoothed piecewise linear approximation: $(F\xi)(t)=\sum_{0\leqq n<t}[\xi(n)+\dot{\xi}(n)(t-n)]H(t-n)$.

Here $\dot{\xi}$ denotes the time derivative of ξ. The function H is supposed to be $\in C^\infty[0,\infty)$ and further $H\geqq 0$, $H(0)=0$, $H(t)=1$ for $\varepsilon\leqq t\leqq 1$, $H(t)=0$ for $t\geqq 1+\varepsilon$, ε a positive constant $\ll 1$.

Such a feed-back coupling would fit in (1.3.8) if

$$\gamma\geqq 1 \;\Big|\; \gamma>\tfrac{3}{2}. \tag{1.3.13}$$

The choice of γ in (1.3.13), in the case of "H-theory", is a consequence of the imbedding theorem (see Adams (1975, p. 97)).

Finally it is useful to remark that certain feed-back operators Π which are not a priori in the form specified here above can be rewritten to that form.

For example, $\Pi=cFP$ with $c\in C^\infty(\bar{Q})$, F the identity operator and $(Pu)\,(t)=\delta_y(u+s(\partial u/\partial t))$. Here s denotes a constant >0 and δ_y is as in (1.2.5).

Note that this P predicts in a simple way the value of $\delta_y u$ at the time $t+s$ from the data at time t.

Certainly P doesn't satisfy the conditions given in (1.3.7), because of the $\delta_y(\partial/\partial t)$ operation. But this operation can be eliminated!

From (1.1.1) we derive

$$\delta_y\frac{\partial u}{\partial t}=\delta_y(Lu+f+c\delta_y u)+s\delta_y c\delta_y\frac{\partial u}{\partial t}.$$

In other words: $\delta_y(\partial u/\partial t)=z\delta_y(Lu+f+c\delta_y u)$ if we suppose $z=(1-s\delta_y c)^{-1}$ to be nonsingular.

Now we can rewrite $\Pi u=\hat{\Pi}u+\hat{f}$ with $\hat{\Pi}=c_1\delta_y(1+sL)$.

Here we denote $c_1=zc$ and $\hat{f}=scz\,\delta_y f$.

Equation (1.1.1) can be rewritten as $\partial u/\partial t=(L+\hat{\Pi})u+g$ with $g=f+\hat{f}$. This equation is of the same type as (1.1.1) with the difference that Π has been replaced by $\hat{\Pi}$.

Now $\hat{\Pi}$ satisfies the conditions in (1.3.5-6-7-8-9) for any $\gamma\in[0,\infty)$ with

$$\alpha=2 \;\Big|\; \alpha=2+\frac{n}{2}+\varepsilon,\quad \varepsilon>0 \text{ arbitrarily small.} \tag{1.3.14}$$

The verification that the choice for α in the case of "H-theory" suffices is analogous to (1.3.11).

1.4. On the strategy to attack the problem. Using the compatibility of the data the problem (1.1.1-2-3) will first be rewritten as

$$\begin{aligned} &\frac{\partial u}{\partial t} = (L+\Pi)w + f_0, \\ (1.4.1) \qquad &Bw = \phi_0, \\ &w(\cdot, 0) = 0, \end{aligned}$$

where f_0 and ϕ_0 vanish up to a certain order for $t \downarrow 0$.

The advantage of (1.4.1) is that the solution W of the corresponding uncontrolled problem is a "nice" function, where originally this is not necessarily true. Now (1.4.1) is equivalent to

$$(1.4.2) \qquad w = S\Pi w + W$$

with S the solution operator of the uncontrolled problem with homogeneous initial— and boundary conditions.

In the case that the order of Π is less than 2, it is possible to show directly from the compactness and nonanticipativity of $S\Pi$ the invertibility of $I - S\Pi$. This is done by the important lemma given in § 3. The solution of (1.4.2) is given by $(I - S\Pi)^{-1}W$ and can be analyzed.

This method fails, however, if the order of Π is larger than 2, since $S\Pi$ is not compact then. In this case we deduce from (1.4.2) a Volterra integral equation for the functions P_iw, $i = 1, \cdots, p$. This is done by operating on (1.4.2) with $\tilde{P}_j(t)$. This Volterra integral equation is then solved by using the lemma of § 3. The properties of w are next found from (1.4.2), since we have $w = S\sum_{i=1}^{q}\sum_{j=1}^{p} c_iF_{ij}P_jw + W$, where the P_jw's are now known.

2. On the compatibility of the data. In order to have a solution of (1.1.1-2-3) which is sufficiently regular at $t = 0$ near the boundary of D, it is necessary that the data f, ϕ, ψ are compatible there in a certain sense. This is well-known in the uncontrolled case. Let us now investigate compatibility in the controlled situation.

In order to do so it is very useful to introduce the following function spaces:

$$(2.1) \qquad \begin{array}{l|l} \underset{0}{C}^{\beta,\beta/2}(\bar{Q}), & \underset{0}{H}^{\beta,\beta/2}(Q), \\ \\ \underset{0}{C}^{\hat{\beta},\hat{\beta}/2}(\bar{\Gamma}), & \underset{0}{H}^{\hat{\beta},\beta/2}(\Gamma), \\ \\ \underset{0}{C}^{\delta}[0, T] & \underset{0}{H}^{\delta}(0, T). \end{array}$$

They are defined as the closure in the spaces without the subscript $_0$ of the ∞-differentiable functions with supports that have no point in common with $\{t = 0\}$. It can be shown (see Ladyzenskaja, Solonnikov, Uraltseva (1967) and Lions, Magenes (1972)), that usually the function spaces introduced in (2.1) can also be described as the subspaces of the function spaces without the subscript $_0$ consisting of functions for which the time derivatives at $t = 0$ vanish up to some number. This number is given below together with an indication when the characterization is valid.

$$\begin{array}{ll|lll} [\beta/2] & \text{if } \beta \geqq 0, & [\frac{1}{2}(\beta-1)] & \text{if } \beta \geqq 0, & \frac{1}{2}(\beta+1) \notin \mathbb{N}, \\ [\hat{\beta}/2] & \text{if } \hat{\beta} \geqq 0, & [\frac{1}{2}(\hat{\beta}-1)] & \text{if } \hat{\beta} \geqq 0, & \frac{1}{2}(\hat{\beta}+1) \notin \mathbb{N}, \\ [\delta] & \text{if } \delta \geqq 0 & [\delta-\frac{1}{2}] & \text{if } \delta \geqq 0, & \delta+\frac{1}{2} \notin \mathbb{N}. \end{array} \tag{2.2}$$

As usual $[\delta]$ denotes the largest integer $\leqq \delta$.

If the number given in (2.2) is less than 0 the function space with subscript $_0$ coincides with the space without subscript $_0$.

Now we use the function spaces introduced in (2.1) to define the concept of compatibility of the data f, ϕ, ψ.

The data will be called compatible of type β_2, β_1, $\hat{\beta}_1$, if $\beta_2 \geqq \max(2, \gamma)$, $\beta_1 \geqq 0$ and

$$\begin{array}{l|l} \exists v \in C^{\beta_2,\beta_2/2}(\bar{Q}) \text{ such that} & \exists v \in H^{\beta_2,\beta_2/2}(Q) \text{ such that} \\ v(\cdot\,, 0) = \psi, & v(\cdot\,, 0) = \psi, \\ \phi - Bv \in \underset{0}{C}^{\hat{\beta}_1,\hat{\beta}_1/2}(\bar{\Gamma}), & \phi - Bv \in \underset{0}{H}^{\hat{\beta}_1,\hat{\beta}_1/2}(\Gamma), \\ f - \dfrac{\partial v}{\partial t} + (L+\Pi)v \in \underset{0}{C}^{\beta_1,\beta_1/2}(\bar{Q}) & f - \dfrac{\partial v}{\partial t} + (L+\Pi)v \in \underset{0}{H}^{\beta_1,\beta_1/2}(Q). \end{array} \tag{2.3}$$

From now on the notation β_2, β_1, $\hat{\beta}_1$ will be reserved for, and will always refer to, the indices indicating the type of compatibility as introduced in (2.3). Let us introduce

$$\begin{array}{l|l} |(f, \phi, \psi)|^{cp}_{\beta_2,\beta_1,\hat{\beta}_1} = & \|(f, \phi, \psi)\|^{cp}_{\beta_2,\beta_1,\hat{\beta}_1} = \\ \quad \inf\{|v|_{\beta_2,\beta_2/2} | v \text{ satisfies (2.3)}\} & \quad \inf\{\|v\|_{\beta_2,\beta_2/2} | v \text{ satisfies (2.3)}\}. \end{array} \tag{2.4}$$

Here we use the convention $\inf \varnothing = \infty$. An equivalent formulation of (2.3) is

$$\begin{array}{l|l} |(f, \phi, \psi)|^{cp}_{\beta_2,\beta_1,\hat{\beta}_1} < \infty & \|(f, \phi, \psi)\|^{cp}_{\beta_2,\beta_1,\hat{\beta}_1} < \infty. \end{array} \tag{2.5}$$

However, it would be preferable if under certain conditions we would be able to derive a less abstract, better verifiable criterion for the compatibility of the data than (2.3).

In order to do so let us first give some easy consequences of the nonanticipativity of the feed-back coupling.

$$\begin{array}{l|l} F_{ij}\underset{0}{C}^{\gamma/2}[0, T] \subset \underset{0}{C}^{\gamma/2}[0, T] & F_{ij}\underset{0}{H}^{\gamma/2}(0, T) \subset \underset{0}{H}^{\gamma/2}(0, T) \\ \text{for } 1 \leqq i \leqq q, \quad 1 \leqq i \leqq p, & \text{for } 1 \leqq i \leqq q, \quad 1 \leqq j \leqq p, \\ \Pi\, \underset{0}{C}^{\check{\gamma},\check{\gamma}/2}(\bar{Q}) \subset \underset{0}{C}^{\gamma,\gamma/2}(\bar{Q}) & \Pi\, \underset{0}{H}^{\check{\gamma},\check{\gamma}/2}(Q) \subset \underset{0}{H}^{\gamma,\gamma/2}(Q). \end{array} \tag{2.6}$$

It appears to be possible to calculate the time-derivatives at $t=0$ of $F_{ij}\xi$, in terms of the time-derivatives at $t=0$ of ξ itself, by a simple matrix multiplication. More precisely suppose:

$$\begin{array}{l|l} \xi \in C^{\gamma/2}[0, T] & \xi \in H^{\gamma/2}(0, T), \quad \gamma > 1, \quad \frac{1}{2}(\gamma+1) \notin \mathbb{N}. \end{array} \tag{2.7}$$

Define

$$d(\gamma)=[\gamma/2] \quad \Big| \quad d(\gamma)=[\tfrac{1}{2}(\gamma-1)].$$

If we define Der ξ as the vector in $\mathbb{R}^{d(\gamma)+1}$ with

$$(\text{Der }\xi)_l=\frac{1}{l!}\frac{d^l\xi}{dt^l}(0), \qquad 0\leqq l\leqq d(\gamma) \tag{2.8}$$

then we have the relation

$$\text{Der } F_{ij}\xi=A^{ij}\text{ Der }\xi \tag{2.9}$$

with A^{ij} the $(d(\gamma)+1)\times(d(\gamma)+1)$ matrix with the matrix elements

$$A^{ij}_{kl}=\left\{\frac{1}{k!}\left(\frac{d}{dt}\right)^k F_{ij}(t^l)\right\}\bigg|_{t=0}, \qquad 0\leqq k,\, l\leqq d(\gamma) \tag{2.10}$$

(where with (t^l) we mean the function $t\to t^l$).

In order to prove (2.10) we observe that $\xi(t)-\sum_{l=0}^{d(\gamma)}(\text{Der }\xi)_l t^l$ defines an element of $\underset{0}{C}^{\gamma/2}[0,T]$, $\underset{0}{H}^{\gamma/2}(0,T)$. Using (2.6) we find that $F_{ij}\xi-\sum_{l=0}^{d(\gamma)}(\text{Der }\xi)_l F_{ij}(t^l)$ is in $\underset{0}{C}^{\gamma/2}[0,T]$, $\underset{0}{H}^{\gamma/2}(0,T)$, and from this fact (2.9–10) are straightforwardly found.

A special situation arises if the matrices A^{ij} have a lower triangle form:

$$k<l\Rightarrow A^{ij}_{kl}=0. \tag{2.11}$$

This will be the case for example, if in addition to (1.3.8), we suppose that

$$F_{ij}\in L(C[0,T]) \quad \Big| \quad F_{ij}\in L(L_2(0,T)). \tag{2.12}$$

This statement is proven in the following way.

Suppose $A^{ij}_{kl}\neq 0$ with $k<l$. Then $F_{ij}(t^{l-1})$ behaves as $A^{ij}_{kl}t^k$ for $t\downarrow 0$. Consider $\xi_\mu(t)=t^l\bar{H}(\mu t)$ with $\bar{H}\in C^\infty[0,\infty)$, $\bar{H}(\tau)=1$ for $\tau\in[0,1]$, $\bar{H}(\tau)=0$ for $\tau\geqq 2$. A simple calculation shows that F_{ij} cannot be bounded in the sense of (2.12) on the functions ξ_μ for $\mu\to\infty$. From this contradiction we see that (2.11) follows from (2.12).

In the example of § 1.2 with h_{ij}, $\bar{h}_{ij}\in C^\infty[0,T]$, (2.12) is certainly fulfilled and consequently we are in the situation of (2.11) there.

The following notation will be useful:

$$c_{i,l}=\left\{\frac{1}{l!}\left(\frac{\partial}{\partial t}\right)^l c_i\right\}\bigg|_{t=0}, \qquad \tilde{P}_{j,l}=\left\{\frac{1}{l!}\left(\frac{\partial}{\partial t}\right)^l \tilde{P}_j\right\}\bigg|_{t=0},$$

(2.13) $a_{ij,l}$, $a_{i,l}$ and $a_{0,l}$ are defined analogously to $c_{i,l}$

$$L_l=\sum_{i=1}^n\sum_{j=1}^n a_{ij,l}D_{ij}+\sum_{i=1}^n a_{i,l}D_i+a_{0,l}.$$

Note that

$$\begin{array}{ll|l} & c_{i,l}\in C^{\gamma-2l}(\bar{D}), & c_{i,l}\in H^{\gamma-2l+1}(D), \\ (2.14) & P_{j,l}\in C^\alpha(\bar{D})' & P_{j,l}\in H^\alpha(D)' \\ & \quad\text{for } 0\leqq l\leqq d(\gamma) & \quad\text{if } \gamma>1, \tfrac{1}{2}(\gamma-1)\notin\mathbb{N} \\ & & \quad\text{and } 0\leqq l\leqq d(\gamma). \end{array}$$

Our first result for a more direct characterization of compatibility concerns the case $0 \leqq \alpha < 2$.

LEMMA 2.1. *Suppose that*:

ASSUMPTION 2.1.1. *The indices indicating regularity of the data* (*see* (1.3.3)), *compatibility of the data* (*see* (2.3)), *and the type of the operator* (*see* (1.3-5-6-7-8-9)) *satisfy*:

$$\beta_2 \geqq 2, \quad \beta_1 = \beta = \gamma = \beta_2 - 2, \quad 0 \leqq \alpha < 2,$$
$$\text{(2.15)} \qquad \begin{array}{l|l} \beta_0 = \beta_2, & \beta_0 = \beta_2 - 1, \\ \hat{\beta} = \hat{\beta}_1 = \beta_2 - \nu, & \hat{\beta} = \hat{\beta}_1 = \beta_2 - \frac{1}{2} - \nu, \\ \beta_2 \notin \mathbb{N} & \frac{1}{2}(\beta_2 - 1) \notin \mathbb{N}, \beta_2 - \frac{1}{2} \notin \mathbb{N}. \end{array}$$

ASSUMPTION 2.1.2. *If* $d(\gamma) \geqq 0$ *with* $d(\gamma)$ *defined as in* (2.7), *then the following linear algebraic system of equations in the variables* $z_{s,s',m}$, $1 \leqq s \leqq p$, $0 \leqq m \leqq d(\gamma)$, $0 \leqq s' \leqq d(\gamma) - m$ *possesses only the trivial solution*

$$z_{s,s',0} = 0, \qquad 1 \leqq s \leqq p, \quad 0 \leqq s' \leqq d(\gamma),$$
$$\text{(2.16)} \qquad \textit{if } d(\gamma) \geqq 1 \textit{ then for } 1 \leqq m+1 \leqq d(\gamma), 1 \leqq s \leqq p, 0 \leqq s' \leqq d(\gamma) - m - 1:$$
$$(m+1) z_{s,s',m+1} = \sum_{j=1}^{p} \sum_{l=0}^{d(\gamma)} \sum_{j'=0}^{l} J_{j,l}^{s,s',m} z_{j,j',l-j'}.$$

Here we define:

$$J_{j,l}^{s,s',m} = \sum_{i=1}^{q} \sum_{m'=0}^{m} \sum_{k=0}^{m'} \tilde{P}_{s,s'} b_{i,k}^{m,m'} A_{kl}^{ij}$$

$$b_{i,k}^{m,m'} = \begin{cases} c_{i,m-k} & \textit{if } m' = m \\ \displaystyle\sum_{\substack{l_1 \geqq 0, \cdots, l_n \geqq 0 \\ n \geqq 1 \\ m' + l_1 + \cdots + l_n = m - n}} \left(\overrightarrow{\prod_{r=1}^{n}} \left(m - \sum_{j=1}^{r} l_j \right)^{-1} L_{l_r} \right) c_{i,m'-k} & \textit{if } m' < m, \end{cases}$$

$$\overrightarrow{\prod_{r=1}^{n}} \operatorname{term}(r) = \operatorname{term}(1) \cdot \cdots \cdot \operatorname{term}(n).$$

Under these conditions there are explicitly determinable operators $M_k \in$

$$\text{(2.17)} \qquad \begin{array}{c|c} L(C^{\beta,\beta/2}(\bar{Q}) + C^{\beta_0}(\bar{D}) & L(H^{\beta,\beta/2}(Q) \times H^{\beta_0}(D) \\ \to C^{\beta_0 - 2k - \nu}(\partial D)), & \to H^{\beta_0 - 2k - \nu - 1/2}(\partial D)), \\ 0 \leqq k \leqq \left[\dfrac{\beta_0 - \nu}{2}\right] = k_0, & 0 \leqq k \leqq \left[\dfrac{\beta_0 - \nu - 1/2}{2}\right] = k_0 \end{array}$$

only dependent on the operators L, Π, *B, such that*:

$$\begin{array}{l|l} |(f,\phi,\psi)|^{cp}_{\beta_2,\beta_1\hat{\beta}_1}<\infty & \textit{If } \beta_2>\tfrac{3}{2}+\nu \textit{ then:} \\ \Leftrightarrow M_k(f,\psi)=\left\{\left(\frac{\partial}{\partial t}\right)^k\phi\right\}\Big|_{t=0}, & \|(f,\phi,\psi)\|^{cp}_{\beta_2.\beta_1\hat{\beta}_1}<\infty \\ \qquad 0\leqq k\leqq k_0. & \Leftrightarrow M_k(f,\psi)=\left\{\left(\frac{\partial}{\partial t}\right)^k\phi\right\}\Big|_{t=0}, \\ & 0\leqq k\leqq k_0. \\ & \textit{If } \beta_2<\tfrac{3}{2}+\nu \textit{ then always} \\ & \|(f,\phi,\psi)\|^{cp}_{\beta_2,\beta_1\hat{\beta}_1}<\infty. \end{array} \tag{2.18}$$

Moreover, there is a constant $K_{cp}>0$ *such that for all compatible triples* (f,ϕ,ψ):

$$\begin{array}{l|l} |(f,\phi,\psi)|^{cp}_{\beta_2,\beta_1,\hat{\beta}_1} & \|(f,\phi,\psi)\|^{cp}_{\beta_2,\beta_1,\hat{\beta}_1} \\ \quad \leqq K_{cp}(|f|_{\beta,\beta/2}+|\psi|_{\beta_0}) & \quad \leqq K_{cp}(\|f\|_{\beta,\beta/2}+\|\psi\|_{\beta_0}). \end{array} \tag{2.19}$$

Proof of Lemma 2.1. "⇒" Suppose that v satisfies (2.3). Define

$$v_1=\left\{\frac{1}{l!}\left(\frac{\partial}{\partial t}\right)^l v\right\}\Big|_{t=0}, \qquad f_l=\left\{\frac{1}{l!}\left(\frac{\partial}{\partial t}\right)^l f\right\}\Big|_{t=0}. \tag{2.20}$$

Note, that:

$$\begin{array}{r|l} v_l\in C^{\beta_2-2l}(\bar{D}), & v_l\in H^{\beta_2-2l-1}(D), \\ f_l\in C^{\beta-2l}(\bar{D}) & f_l\in H^{\beta-2l-1}(D), \end{array} \tag{2.21}$$

and

$$v_0=\psi. \tag{2.22}$$

If $d(\gamma)\geqq 0$ then we find for $0\leqq m\leqq d(\gamma)$ by expansion in powers of t of $\partial v/\partial t-(L+\Pi)v-f$:

$$\begin{aligned}(m+1)v_{m+1}&=\sum_{l=0}^{m}L_l v_{m-l}+f_m\\ &\quad+\sum_{i=1}^{q}\sum_{j=1}^{p}\sum_{k=0}^{m}\sum_{l=0}^{d(\gamma)}\sum_{j'=0}^{l}c_{i,m-k}A^{ij}_{kl}\tilde{P}_{j,j'}v_{l-j'}.\end{aligned} \tag{2.23}$$

We eliminate $\sum_{l=0}^{m}L_l v_{m-l}$ from (2.23). This leads to

$$\begin{aligned}(m+1)v_{m+1}&=\bar{N}_{m+1}(f,\psi)\\ &\quad+\sum_{i=1}^{q}\sum_{j=1}^{p}\sum_{m'=0}^{m}\sum_{k=0}^{m'}\sum_{l=0}^{d(\gamma)}\sum_{j'=0}^{l}b^{m,m'}_{i,k}A^{ij}_{kl}\tilde{P}_{j,j'}v_{l-j'}\end{aligned} \tag{2.24}$$

with $b_{i,k}^{m,m'}$ as introduced in Assumption 2.1.2 and:

$$\bar{N}_{m+1}(f,\psi)=\Big(L_m+\sum_{\substack{l_1\geqq 0,\cdots,l_n\geqq 0\\ n\geqq 2\\ l_1+\cdots+l_n=m+1-n}}\overrightarrow{\prod_{i=1}^{n}}\Big(m-\sum_{j=1}^{i}l_j\Big)^{-1}L_{l_i}\Big)\psi$$

$$+f_m+\Big(\sum_{\substack{l_0\geqq 0,\cdots,l_n\geqq 0\\ n\geqq 1\\ l_0+\cdots+l_n=m-n}}\overrightarrow{\prod_{i=1}^{n}}\Big(m-\sum_{j=1}^{i}l_j\Big)^{-1}L_{l_i}\Big)f_{l_0}.$$

Operating at (2.24) with $\tilde{P}_{ss'}$, we find the following set of linear algebraic equations for $\tilde{P}_{ss'}v_m$, $1\leqq s\leqq q$, $0\leqq m\leqq d(\gamma)$, $0\leqq s'\leqq d(\gamma)-m$, if $d(\gamma)\geqq 0$:

(2.25) $$P_{s,s'}v_0=P_{s,s'}\psi,\qquad 1\leqq s\leqq p,\qquad 0\leqq s'\leqq d(\gamma);$$

if $d(\gamma)\geqq 1$, then for $1\leqq m+1\leqq d(\gamma)$, $1\leqq s\leqq p$, $0\leqq s'\leqq d(\gamma)$,

$$(m+1)\tilde{P}_{s,s'}v_{m+1}=\tilde{P}_{s,s'}\bar{N}_{m+1}(f,\psi)+\sum_{j=1}^{p}\sum_{l=0}^{d(\gamma)}\sum_{j'=0}^{l}J_{j,l}^{s,s',m}\tilde{P}_{j,j'}v_{l-j'}.$$

Because of Assumption 2.1.2, we know that (2.25) possesses a unique solution. This solution is denoted by

(2.26) $$\tilde{P}_{s,s'}v_m=E_{s,s',m}(f,\psi),\qquad 1\leqq s\leqq p,\quad 0\leqq m\leqq d(\gamma),\quad 0\leqq s'\leqq d(\gamma)-m.$$

Substitution of the result of (2.26) in (2.24) provides us with explicit expressions of the following type:

(2.27) $$v_m=N_m(f,\psi),\qquad 0\leqq m\leqq d(\gamma)+1,$$

where, by definition, $N_0(f,\psi)=\psi$. It is not difficult to verify that N_m is an operator $\in$

(2.28) $$\begin{array}{r|r} L(C^{\beta,\beta/2}(\bar{Q})\times C^{\beta_0}(\bar{D}) & L(H^{\beta,\beta/2}(Q)\times H^{\beta_0}(D) \\ \to C^{\beta_0-2m}(\bar{D})) & \to H^{\beta_0-2m}(D)). \end{array}$$

It will now be clear how the operators M_k are defined: in the case of Dirichlet BC:

(2.29) $$M_k(f,\psi)=\text{restriction to } \partial D \text{ of } N_k(f,\psi);$$

in the case of Neumann BC:

(2.30) $$M_k=k!\sum_{l=0}^{k}B_lN_{k-l}\quad\text{with}$$

$$B_l=\sum_{i=1}^{n}b_{il}D_i+b_{0,l},\qquad b_{i,l}=\left\{\frac{1}{l!}\Big(\frac{\partial}{\partial t}\Big)^l b_i\right\}\Bigg|_{t=0}.$$

An expansion of $\phi-Bv$ in powers of t yields exactly the contents of (2.18) "$\Rightarrow$".

"$\Leftarrow$" Given that if $k_0\geqq 0$, the relations $M_k(f,\psi)=\{(\partial/\partial t)^k\phi\}|_{t=0}$ are satisfied for $0\leqq k\leqq k_0$ with M_k's as introduced in (2.29–30) we shall show the existence of a function v as required in (2.3). Define:

(2.31) $$v_k=N_k(f,\psi),\qquad 0\leqq k\leqq d(\gamma)+1$$

with N_k as introduced in (2.27).

Using the theory as developed in the Theorems 4.1–2–3–4 of Ladyzenskaja, Solonnikov, Uraltseva (1967, pp. 298–301), we find the existence of a function v such that:

$$\left\{\frac{1}{k!}\left(\frac{\partial}{\partial t}\right)^k v\right\}\bigg|_{t=0} = v_k \quad \text{for } 0 \leqq k \leqq d(\gamma)+1 \tag{2.32}$$

and moreover:

$$\begin{array}{r|r} |v|_{\beta_2,\beta_2/2} \leqq C \displaystyle\sum_{k=0}^{d(\gamma)+1} |v_k|_{\beta_0-2k} & \|v\|_{\beta_2,\beta_2/2} \leqq C \displaystyle\sum_{k=0}^{d(\gamma)+1} \|v_k\|_{\beta_0-2k} \\ \leqq K_{cp}(|f|_{\beta,\beta/2}+|\psi|_{\beta_0}) & \leqq K_{cp}(\|f\|_{\beta,\beta/2}+\|\psi\|_{\beta_0}) \end{array} \tag{2.33}$$

with certain constants C, $K_{cp} > 0$ only dependent of L, Π.

By construction we now have $v(\cdot, 0) = \psi$, and it is easy to show that $\partial v/\partial t - (L+\Pi)v - f \in$

$$\underset{0}{C}^{\beta_1,\beta_1/2}(\bar{Q}) \;\Big|\; \underset{0}{H}^{\beta_1,\beta_1/2}(Q). \tag{2.34}$$

Namely, if $d(\gamma) \geqq 0$, then expansion in powers of t yields for $1 \leqq m+1 \leqq d(\gamma)+1$:

$$\left\{\frac{1}{m!}\left(\frac{\partial}{\partial t}\right)^m\left(\frac{\partial v}{\partial t} - (L+\Pi)v - f\right)\right\}\bigg|_{t=0} = 2.23^* = 2.24^* = 2.25^* = 0$$

because of the definition of the operators N_m. The superscript * means: put the right-hand side of the indicated equation in parentheses, change the equality sign of the indicated equation into a minus sign and take v_k's as introduced in (2.31–32) in this expression.

It is also completely straightforward to show that $\phi - Bv \in$

$$\underset{0}{C}^{\hat{\beta}_1,\hat{\beta}_1/2}(\bar{\Gamma}) \;\Big|\; \underset{0}{H}^{\beta_1,\hat{\beta}_1/2}(\Gamma). \tag{2.35}$$

Let us give the calculation in the case of Neumann BC:

$$\begin{aligned}\left\{\left(\frac{\partial}{\partial t}\right)^k(\phi - Bv)\right\}\bigg|_{t=0} &= \left\{\left(\frac{\partial}{\partial t}\right)^k \phi\right\}\bigg|_{t=0} - k!\sum_{l=0}^{k} B_l v_{k-l} \\ &= \left\{\left(\frac{\partial}{\partial t}\right)^k \phi\right\}\bigg|_{t=0} - k!\sum_{l=0}^{k} B_l N_{k-l}(f, \psi) = 0\end{aligned}$$

by the definition of M_k and the given relations.

So (2.18) "$\Leftarrow$" and (2.19) have also been demonstrated. □

It is interesting to notice the following

COROLLARY (to Lemma 2.1). *Assumption* 2.1.2 *is certainly satisfied if the matrices* A^{ij} *have a lower triangle form* (*see* (2.11–12)).

This is true since in this case $l > m \Rightarrow J_{j,l}^{s,s',m} = 0$. As a consequence the system of (2.16) (and also (2.23)) possesses a recursive structure with respect to m, i.e., in order to calculate $z_{s,s',m+1}(v_{m+1})$ we only use $z_{s,s',k}(v_k)$ with $0 \leqq k \leqq m$. So indeed the only solution of (2.16) is then the trivial one.

In the following lemma we shall give an analogue of the previous results in the case $\alpha \geqq 2$.

LEMMA 2.2. *Suppose that*:

ASSUMPTION 2.2.1. *The indices indicating regularity of the data* (*see* (1.3.3)), *compatibility of the data* (*see* (2.3)) *and the type of the operator* Π (*see* (1.3.5-6-7-8-9)) *satisfy*:

$$\beta_2 \geqq \alpha \geqq 2, \quad \beta_1 = \gamma = \beta_2 - \alpha, \quad \beta = \beta_2 - 2,$$

$$\begin{array}{l|l} \beta_0 = \beta_2, & \beta_0 = \beta_2 - 1, \\ \hat{\beta} = \hat{\beta}_1 = \beta_1 + 2 - \nu, & \hat{\beta} = \hat{\beta}_1 = \beta_1 + \frac{3}{2} - \nu, \\ c_i \in C^{\beta,\beta/2}(\bar{Q}),\ 1 \leqq i \leqq q, & c_i \in H^{\beta,\beta/2}(Q),\ 1 \leqq i \leqq q, \\ \beta_2 \notin \mathbb{N} & \frac{1}{2}(\beta_2 - 1) \notin \mathbb{N},\ \beta_2 - \frac{1}{2} \notin \mathbb{N}. \end{array} \tag{2.36}$$

ASSUMPTION 2.2.2. *It is assumed, that the contents of Assumption* 2.1.2 *are valid with* γ as defined in Assumption 2.2.1.

Under these conditions there are explicitly determinable operators $M_k \in$

$$\begin{array}{r|r} L(C^{\beta,\beta/2}(\bar{Q}) \times C^{\beta_0}(\bar{D}) \rightarrow & L(H^{\beta,\beta/2}(Q) \times H^{\beta_0}(D) \rightarrow \\ C^{\beta_0 - 2k - \nu}(\partial D)) & H^{\beta_0 - 2k - \nu - (1/2)}(\partial D)) \\ \textit{for } 0 \leqq k \leqq \bar{k}_0, & \textit{for } 0 \leqq k \leqq \bar{k}_0, \\ \bar{k}_0 = \min\left(\left[\frac{\beta_0 - \nu}{2}\right], \left[\frac{\beta_1 + 2}{2}\right]\right) & \bar{k}_0 = \min\left(\left[\frac{\beta_0 - (1/2) - \nu}{2}\right], \left[\frac{\beta_1 + 1}{2}\right]\right) \end{array} \tag{2.37}$$

only dependent on the operators L, Π, B *such that*:

$$\begin{array}{r|l} & \textit{If } \beta_2 > \frac{3}{2} + \nu \textit{ then}: \\ |(f, \phi, \psi)|^{cp}_{\beta_2,\beta_1,\hat{\beta}_1} < \infty & \|(f, \phi, \psi)\|^{cp}_{\beta_2,\beta_1,\hat{\beta}_1} < \infty \\ \Leftrightarrow M_k(f, \psi) = \left\{\left(\frac{\partial}{\partial t}\right)^k \phi\right\}\Big|_{t=0}, & \Leftrightarrow M_k(f, \psi) = \left\{\left(\frac{\partial}{\partial t}\right)^k \phi\right\}\Big|_{t=0}, \\ 0 \leqq k \leqq \bar{k}_0. & \qquad\qquad 0 \leqq k \leqq \bar{k}_0. \\ & \textit{If } \beta_2 < \frac{3}{2} + \nu \textit{ then always} \\ & \|(f, \phi, \psi)\|^{cp}_{\beta_2,\beta_1,\hat{\beta}_1} < \infty. \end{array} \tag{2.38}$$

Moreover, there is a constant $K_{cp} > 0$ *such that for all compatible triples* (f, ϕ, ψ):

$$\begin{array}{r|l} |(f, \phi, \psi)|^{cp}_{\beta_2,\beta_1,\hat{\beta}_1} & \|(f, \phi, \psi)\|^{cp}_{\beta_2,\beta_1,\hat{\beta}_1} \\ \leqq K_{cp}(|f|_{\beta,\beta/2} + |\psi|_{\beta_0}) & \quad \leqq K_{cp}(\|f\|_{\beta,\beta/2} + \|\psi\|_{\beta_0}). \end{array} \tag{2.39}$$

Proof of Lemma 2.2. "$\Rightarrow$" This part of the proof is completely analogous to the corresponding part of the proof of Lemma 2.1. One can even use the same text as before with, in the final conclusion, (2.18) replaced by (2.38). However, a few remarks should be made.

It is now used, that $P_{s,s'}v_m$ is well-defined for $1 \leqq s \leqq p$, $0 \leqq m \leqq d(\gamma)$, $0 \leqq s' \leqq d(\gamma) - m$ if $d(\gamma) \geqq 0$. This is the case since $\beta_0 - 2d(\gamma) \geqq \alpha$.

Further we note that v_m is given by the formula in (2.20) for $0 \leqq m \leqq d(\beta_2)$, but that the formula of (2.27) is only valid for $0 \leqq m \leqq d(\gamma) + 1$ and it is certainly possible that $d(\beta_2) > d(\gamma) + 1$!

"$\Leftarrow$" Given that if $\bar{k}_0 \geqq 0$, the relations $M_k(f, \psi) = \{(\partial/\partial t)^k \phi\}|_{t=0}$ are satisfied for $0 \leqq k \leqq \bar{k}_0$ we now construct a function v as required in (2.3) as follows. Define:

$$\begin{array}{ll} v_k = N_k(f, \psi), & 0 \leqq k \leqq d(\gamma) + 1 \\ v_k = 0, & d(\gamma) + 2 \leqq k \leqq d(\beta_2) \quad \text{if } d(\beta_2) > d(\gamma) + 1. \end{array} \tag{2.40}$$

As before the theory of Ladyzenskaja, Solonnikov, Uraltseva (1967) provides us with a function v such that

$$\left\{\frac{1}{k!}\left(\frac{\partial}{\partial t}\right)^k v\right\}\bigg|_{t=0} = v_k, \qquad 0 \leqq k \leqq d(\beta_2) \tag{2.41}$$

for which (2.33) holds.

The proof is concluded in the same way as the proof of Lemma 2.1. □

COROLLARY (to Lemma 2.2). *Assumption* 2.2.2 *is certainly satisfied if the matrices* A^{ij} *have a lower triangle form* (*see* (2.11–12)).

This is in fact the same remark as the Corollary to Lemma 2.1.

3. Nilpotency of nonanticipative compact linear operators. Here we give and prove a lemma with contents as indicated in the title of this section, which will be very useful in the next sections.

Let B be a nontrivial Banach space with norm $\|\cdot\|_B$. We suppose, that on B we have a strongly continuous semi-group of bounded linear operators $\{U(\tau);\ \tau \geqq 0\}$ with the following properties:

(a) *existence of a finite time-horizon* $T > 0$, *i.e.*,

$$\begin{aligned} &\ker U(\tau) \neq B \quad \textit{for}\ \tau \in [0, T), \\ &\ker U(\tau) = B \quad \textit{for}\ \tau \geqq T. \end{aligned} \tag{3.1}$$

(b)

$$\forall \tau \in [0, T] \quad \ker U(\tau) = \operatorname{ran} U(T-\tau). \tag{3.2}$$

The $U(\tau)$'s are called time shifts.

Examples of Banach spaces with time shifts and a finite time horizon $T > 0$ are $\underset{0}{C}^{\mu}[0, T]$, $\underset{0}{H}^{\mu}(0, T)$, $\underset{0}{C}^{2\mu,\mu}(\bar{Q})$, $\underset{0}{H}^{2\mu,\mu}(Q)$ with $Q = D \times (0, T)$.

On these spaces we define time shifts in the obvious way:

$$(U(\tau)f)(\cdot, t) = \begin{cases} 0 & \text{for } t \in [0, \min(\tau, T)] \\ f(\cdot, t-\tau) & \text{for } t \in (\tau, T] \text{ if } \tau < T \end{cases} \tag{3.3}$$

(the $\cdot$ in $(\cdot, t)$ indicates possible other variables).

The properties a, b are easily verified in these examples. An operator $A \in L(B)$ will be called nonanticipative if:

$$\forall \tau \in [0, \infty) \quad A \ker U(\tau) \subset \ker U(\tau). \tag{3.4}$$

Note that the nonanticipativity of the feed-back coupling as defined in (1.3.8) induces nonanticipativity of the F_{ij}'s in the sense of (3.4) on $\underset{0}{C}^{\gamma/2}[0, T]$ in the case of "C-theory" and on $\underset{0}{H}^{\gamma/2}(0, T)$ in the case of H-theory.

Now we shall prove the following result:

LEMMA 3.1. *Let B be a nontrivial Banach space with time shift with a finite time horizon $T > 0$ for which* (a) *and* (b) *are satisfied.*

If $A \in L(B)$ is nonanticipative and compact then A is nilpotent, i.e., $\sigma(A) = \{0\}$.

Proof of Lemma 3.1. This lemma is a straightforward application of Corollary 4.3.11 of Ringrose (1971, p. 177).

The continuous chain of closed subspaces of B mentioned in that corollary is here given by

$$\tilde{\mathscr{F}} = \{\ker U(\tau) | \tau \in [0, T]\}. \tag{3.5}$$

Of course $\ker U(\tau)$ is closed, for $U(\tau)$ is a bounded operator on B.

The system $\tilde{\mathscr{F}}$ is totally ordered by inclusion:

$$\ker U(\tau_1) \supset \ker U(\tau_2) \quad \text{for } \tau_1 \geqq \tau_2. \tag{3.6}$$

The chain $\tilde{\mathscr{F}}$ also satisfies

$$0 = \ker U(0) \in \tilde{\mathscr{F}}, \qquad B = \ker U(T) \in \tilde{\mathscr{F}}. \tag{3.7}$$

Further we have the property that for any subset $I \subset [0, T]$,

$$\bigcap_{\tau \in I} \ker U(\tau) = \ker U(\inf I), \tag{3.8}$$

$$\overline{\bigcup_{\tau \in I} \ker U(\tau)} = \ker U(\sup I). \tag{3.9}$$

As for (3.8) it is clear, that $\bigcap_{\tau \in \mathrm{I}} \ker U(\tau) \supset \ker U$ $(\inf I)$.

Let τ_n now be a sequence in I such that $\tau_n \downarrow \inf I$ for $n \uparrow \infty$ and let f satisfy $U(\tau_n)f = 0$ for $n \in N$.

By the strong continuity of the semi-group we have $U(\lim_{n\uparrow\infty} \tau_n)f = \lim_{n\uparrow\infty} U(\tau_n)f = 0$. Consequently $\bigcap_{\tau \in I} \ker U(\tau) \subset \ker U(\inf I)$.

As for (3.9) it is clear, that $\overline{\bigcap_{\tau \in I} \ker U(\tau)} \subset \ker U(\sup I)$. If $f \in \ker U(\tau_0)$, $\tau_0 = \sup I$ then $f = U(T - \tau_n)\, g$ because of (b).

Let τ_n now be a sequence in I such that $\tau_n \uparrow \tau_0$ for $n \uparrow \infty$ and define $f_n = U(T - \tau_n)g$.

By the strong continuity of the semi-group we have $U(T - \tau_0)g = \lim_{n\uparrow\infty} U(T - \tau_n)g$; so $f = \lim_{n\uparrow\infty} f_n$. Consequently $\overline{\bigcup_{\tau \in I} \ker U(\tau)} \supset \ker U(\sup I)$.

The properties of $\tilde{\mathscr{F}}$ given in (3.7–8–9) imply that conditions (i) and (ii) of Ringrose (1971, pp. 166–167) are satisfied.

Now it remains to show that in Ringrose's notation $M_- = M$ for each $M \in \tilde{\mathscr{F}}$.

If $M = \ker U(\tau_0)$ then we have in this situation: $M_- = \mathrm{cl}\, \{\bigcup L | L \in \tilde{\mathscr{F}},\ L \subsetneqq M\} = \overline{\bigcup_{\tau < \tau_0} \ker U(\tau)}$.

Indeed $\ker U(\tau) \neq \ker U(\tau_0)$ for $\tau < \tau_0$ since if we suppose $\ker U(\tau) = \ker U(\tau_0)$, then we would have $\operatorname{ran} U(T - \tau) = \operatorname{ran} U(T - \tau_0)$,

$$\Rightarrow U(\tau) \operatorname{ran} U(T - \tau) = U(\tau) \operatorname{ran} U(T - \tau_0)$$

$$\Rightarrow \{0\} = \operatorname{ran} U(T - \tau_0 + \tau_1) = \ker U(\tau_0 - \tau_1)$$

$$\Rightarrow \ker U(\varepsilon) = \{0\} \text{ for some } \varepsilon > 0$$

$$\Rightarrow \ker U(t) = \{0\}\ \forall t \geqq 0$$

which would contradict (a). Now $M_- = M$ follows from (3.9). Herewith all conditions of the corollary have been verified. □

A direct consequence of Lemma 3.1 is that for such an operator A the equation for $w \in B$,

$$w = Aw + g \quad \text{with } g \in B \tag{3.10}$$

possesses a unique solution $\in B$:

$$w=(I-A)^{-1}g \tag{3.11}$$

with $(I-A)^{-1}\in L(B)$, $\|w\|_B \leqq \|(I-A)^{-1}\|_{L(B)}\|g\|_B$.

This is the way Lemma 3.1 will be used further on.

4. Well-posedness if the order of the feed-back control operator is less than the order of the diffusion operator. Here we consider the case $0\leqq\alpha<2$. We are now ready to prove the following result on existence, uniqueness, regularity, and continuous dependence on the data of a solution of (1.1.1–2–3).

THEOREM 4.1. *Let us suppose*:

ASSUMPTION 4.1. *The indices indicating the regularity of the data* (*see* (1.3.3)), *and the type of the operator* Π (*see* (1.3.5–6–7–8–9) *satisfy*

$$\begin{array}{l|l} \beta>0, \quad \beta\notin\mathbb{N}, & \beta\geqq 0, \quad \tfrac{1}{2}(\beta+1)\notin\mathbb{N}, \quad \beta+\tfrac{1}{2}\notin\mathbb{N}, \\ \beta_0=\beta+2, & \beta_0=\beta+1, \\ \hat\beta=\beta+2-\nu, & \hat\beta=\beta+\tfrac{3}{2}-\nu, \\ \gamma=\beta, \quad 0\leqq\alpha<2 & \gamma=\beta, \quad 0\leqq\alpha<2. \end{array} \tag{4.1}$$

ASSUMPTION 4.2. *The data are compatible in the following sense*:

$$|(f,\phi,\psi)|^{cp}_{\beta_2,\beta_1,\hat\beta_1}<\infty \;\Big|\; \|(f,\phi,\psi)\|^{cp}_{\beta_2,\beta_1,\hat\beta_1}<\infty \tag{4.2}$$

with $\beta_2=\beta+2$, $\beta_1=\beta$, $\hat\beta_1=\hat\beta$.

Then there exists a unique solution u of (1.1.1–2–3) *in the space*:

$$C^{2s,s}(\bar Q) \;\Big|\; H^{2s,s}(Q) \tag{4.3}$$

with $2s=\beta+2$. *This solution depends continuously on the data in the following sense*:

$$\begin{array}{l|l} |u|_{2s,s}\leqq K(|f|_{\beta,\beta/2} & \|u\|_{2s,s}\leqq K(\|f\|_{\beta,\beta/2} \\ \qquad +|\phi|_{\hat\beta,\hat\beta/2} & \qquad +\|\phi\|_{\hat\beta,\hat\beta/2} \\ \qquad\quad +|(f,\phi,\psi)|^{cp}_{\beta_2,\beta_1,\hat\beta_1}) & \qquad\quad +\|(f,\phi,\psi)\|^{cp.}_{\beta_2,\beta_1,\hat\beta_1}) \end{array} \tag{4.4}$$

with a constant $K>0$ *independent of* f, ϕ, ψ.

Note that in Assumption 4.1 the exceptional cases for β are exactly as in the uncontrolled problem (see Ladyzenskaja, et al. (1967, Thms. 5.2–3, p. 320), and Lions, Magenes (1972, Thm. 6.2, p. 37)). Further it will be clear that the compatibility condition of Assumption 4.2 is nececessary in order to have (4.3).

Note, that theorem 4.1 is applicable to the example of § 1.2 with h_{ij}, $\bar h_{ij}\in C^\infty[0,T]$ in the case of "C-theory". In the case of "H-theory", this is also true if $D\subset\mathbb{R}^n$ with $n\leqq 3$, see (1.3.11).

Proof of Theorem 4.1. Let v be a function for which (2.3) is satisfied with

$$|v|_{\beta_2,\beta_2/2}\leqq|(f,\phi,\psi)|^{cp}_{\beta_2,\beta_1,\hat\beta_1}+\varepsilon \;\Big|\; \|v\|_{\beta_2,\beta_2/2}\leqq\|(f,\phi,\psi)\|^{cp}_{\beta_2,\beta_1,\hat\beta_1}+\varepsilon \tag{4.5}$$

with $\varepsilon>0$ arbitrarily small.

Put $w=u-v$; then

$$\begin{aligned} \frac{\partial w}{\partial t} &= (L+\Pi)w+f_0, && \text{FPDE},\\ Bw &= \phi_0, && \text{BC},\\ w(\cdot,0) &= 0, && \text{IC} \end{aligned} \tag{4.6}$$

with

$$\begin{array}{l|l} f_0\in \underset{0}{C}^{\beta,\beta/2}(\bar{Q}), & f_0\in \underset{0}{H}^{\beta,\beta/2}(Q),\\ \phi_0\in \underset{0}{C}^{\hat{\beta},\hat{\beta}/2}(\bar{\Gamma}), & \phi_0\in \underset{0}{H}^{\hat{\beta},\hat{\beta}/2}(\Gamma),\\ |f_0|_{\beta,\beta/2}\leqq |f|_{\beta,\beta/2}+C|v|_{\beta_2,\beta_2/2}, & \|f_0\|_{\beta,\beta/2}\leqq \|f\|_{\beta,\beta/2}+C\|v\|_{\beta_2,\beta_2/2},\\ |\phi_0|_{\hat{\beta},\hat{\beta}/2}\leqq |\phi|_{\hat{\beta},\hat{\beta}/2}+C|v|_{\beta_2,\beta_2/2} & \|\phi_0\|_{\hat{\beta},\hat{\beta}/2}\leqq \|\phi\|_{\hat{\beta},\hat{\beta}/2}+C\|v\|_{\beta_2,\beta_2/2}. \end{array} \tag{4.7}$$

In this proof $C>0$ will always denote some constant only dependent of L, B, Π, β, $\hat{\beta}$, β_0, β_2, β_1, $\hat{\beta}_1$, γ, α, D.

We shall now solve (4.6). In order to do so we shall rewrite (4.6) in the form (4.10) below.

Using the solution theory for the parabolic problem as given in Ladyzenskaja, Solonnikov, Uraltseva (1967) and Lions, Magenes (1972, Thm. 6.2, p. 37), we see that:

(i) the uncontrolled problem corresponding to (4.6) possesses a unique solution W with

$$\begin{array}{r|r} W\in \underset{0}{C}^{\beta+2,(\beta+2)/2}(\bar{Q}), & W\in \underset{0}{H}^{\beta+2,(\beta+2)/2}(Q),\\ |W|_{\beta+2,(\beta+2)/2}\leqq C(|f_0|_{\beta,\beta/2} & \|W\|_{\beta+2,(\beta+2)/2}\leqq C(\|f_0\|_{\beta,\beta/2}\\ +|\phi_0|_{\hat{\beta},\hat{\beta}/2}) & +\|\phi_0\|_{\hat{\beta},\hat{\beta}/2}); \end{array} \tag{4.8}$$

(ii) the solution operator S for the uncontrolled problem with homogeneous IC and BC is an element of:

$$\begin{array}{r|r} L(\underset{0}{C}^{\beta,\beta/2}(\bar{Q}) & L(\underset{0}{H}^{\beta,\beta/2}(Q)\\ \to \underset{0}{C}^{\beta+2,(\beta+2)/2}(\bar{Q})) & \to \underset{0}{H}^{\beta+2,(\beta+2)/2}(Q)). \end{array} \tag{4.9}$$

Now we can rewrite (4.6) in the following way:

$$w=Aw+W \quad \text{with } A=S\Pi. \tag{4.10}$$

Equations (4.6) and (4.10) are equivalent for

$$w\in \underset{0}{C}^{2s,s}(\bar{Q}) \;\Big|\; w\in \underset{0}{H}^{2s,s}(Q).$$

We shall show that

$$(4.11)\qquad A\in L(\underset{0}{C}^{2s,s}(\bar{Q}))\;\Big|\; A\in L(\underset{0}{H}^{2s,s}(Q)),$$

and A is compact and nonanticipative.

Then Lemma 3.1 yields (see (3.15–16)) that (4.10) possesses a unique solution

$$(4.11)\qquad A\in L(\underset{0}{C}^{2s,s}(\bar{Q}))\;\Big|\; A\in L(\underset{0}{H}^{2s,s}(Q)),$$

and

$$(4.13)\qquad |w|_{2s,s}\leqq C|W|_{2s,s}\;\Big|\;\|w\|_{2s,s}\leqq C\|W\|_{2s,s}.$$

Herewith the first part of theorem 4.1 has then been proven. The estimate (4.4) is a trivial consequence of $u=w+v$ and (4.5–7–8–13) if we take the limit $\varepsilon\downarrow 0$.

So it remains to prove (4.11).

(a) Using (2.6) and (4.9) we find $\Pi\in$

$$(4.14)\qquad L(\underset{0}{C}^{\check{\gamma},\check{\gamma}/2}(\bar{Q})\to\underset{0}{C}^{2\gamma,\gamma}(\bar{Q}))\;\Big|\; L(\underset{0}{H}^{\check{\gamma},\check{\gamma}/2}(Q)\to\underset{0}{H}^{2\gamma,\gamma}(Q))$$

with $\check{\gamma}=\gamma+\alpha=\beta+\alpha<\beta+2=2s$.

Because of the compact imbedding of

$$(4.15)\qquad \underset{0}{C}^{2s,s}(\bar{Q})\quad\text{in}\;\underset{0}{C}^{\check{\gamma},\check{\gamma}/2}(\bar{Q})\;\Big|\;\underset{0}{H}^{2s,s}(Q)\quad\text{in}\;\underset{0}{H}^{\check{\gamma},\check{\gamma}/2}(Q),$$

we obtain that A is an element of the space indicated in (4.11) and that A is compact. Note that we used the assumption $0\leqq\alpha<2$ here!

(b) The operator S is nonanticipative:

$$(4.16)\qquad \forall\tau\in(0,T],\;u(\cdot,t)=0\text{ for }0<t<\tau\quad\Rightarrow\quad(Su)(\cdot,t)=0\text{ for }0<t<\tau.$$

For $u\in C^{\infty}(\bar{Q})$ this is a consequence of the causal type of maximum principle valid for equations of parabolic type (see Protter, Weinberger (1967) or Friedman (1975)). By continuity of S, the property is then valid on the whole of the domain of S.

The operator Π is nonanticipative because of the nonanticipativity of the feed-back coupling, see (1.3.8). This immediately implies the nonanticipativity of A. □

COROLLARY (to Theorem 4.1). *If Assumption* 2.1.2 *is satisfied then the charac/terization of compatibility of the data given in Lemma* 2.1, (2.18) *is valid and* (4.4) *can be replaced by the estimate*

$$(4.17)\qquad \begin{aligned}|u|_{2s,s}\leqq K(&|f|_{\beta,\beta/2}\\ &+|\phi|_{\hat{\beta},\beta/2}+|\psi|_{\beta_0})\end{aligned}\;\Bigg|\;\begin{aligned}\|u\|_{2s,s}\leqq K(&\|f\|_{\beta,\beta/2}\\ &+\|\phi\|_{\hat{\beta},\beta/2}+\|\psi\|_{\beta_0}\end{aligned}$$

with a constant $K>0$ independent of f, ϕ, ψ. This implies continuous dependence of the data in the usual sense.

This corollary follows from the fact, that Assumptions 4.1 and 4.2 imply Assumption 2.1.1. As a consequence we can apply Lemma 2.1. Then (4.17) is found by inserting (2.19) into (4.4).

5. The case where the order of the feed-back control operator equals or exceeds the order of the diffusion operator

5.1. A result on well-posedness. Now we shall consider the case $\alpha\geqq 2$. In order to derive a well-posedness result we have to make a number of assumptions.

ASSUMPTION 5.1.1. *The indices indicating the regularity of the data* (*see* (1.3.3)) *and the type of the operator* Π (*see* (1.3.5–6–7–8–9)) *satisfy*:

(5.1.1)	$\beta \geqq 2\alpha - 4 \geqq 0$, $\beta \notin \mathbb{Z}, \beta - \alpha \notin \mathbb{Z}$, $\beta_0 - \beta + \alpha$, $\hat{\beta} = \beta + 2 - \nu$, $\gamma = \beta$ *and also the choice* $\gamma = 2\delta$ *with* $\delta = \frac{1}{2}(\beta + 2 - \alpha)$ *is allowed.*	$\beta \geqq 2\alpha - 4 \geqq 0, \frac{1}{2}(\beta + 1) \notin \mathbb{Z}, \beta + \frac{1}{2} \notin \mathbb{Z}$, $\frac{1}{2}(\beta - \alpha + 1) \notin \mathbb{Z}, \beta + \frac{1}{2} - \frac{\alpha}{2} \notin \mathbb{Z}$, $\beta_0 = \beta + \alpha - 1$, $\hat{\beta} = \beta + \frac{3}{2} - \nu$, $\gamma = \beta$ *and also the choice* $\gamma = 2\delta$ *with* $\delta = \frac{1}{2}(\beta + 2 - \alpha)$ *is allowed.*
(5.1.2)		$\operatorname*{ess\,sup}_{0<t<T} \lVert c_i(\cdot, t) \rVert_\gamma < \infty$ *if* $0 \leqq \delta < \frac{1}{2}$; $\exists \bar{\delta} > \delta$ *such that*: $\operatorname*{ess\,sup}_{0<t<\tau<T} \lvert t - \tau \rvert^{-\delta} \lVert c_i(\cdot, t) - c_i(\cdot, \tau) \rVert_0 < \infty$ *if* $0 < \delta < \frac{1}{2}$.

ASSUMPTION 5.1.2. *The data are compatible in the following sense*:

$$\lvert (f, \phi, \psi) \rvert^{cp}_{\beta_2, \beta_1, \hat{\beta}_1} < \infty \quad \Big| \quad \lVert (f, \phi, \psi) \rVert^{cp}_{\beta_2, \beta_1, \hat{\beta}_1} < \infty \tag{5.1.3}$$

with $\beta_2 = \beta + \alpha$, $\beta_1 = \beta$, $\hat{\beta}_1 = \hat{\beta}$.

Assumptions 5.1.1 and 5.1.2 are rather analogous to Assumptions 4.1 and 4.2. However, in this case we have to assume more.

Introduce $\mathscr{C}_i(x, t; \tau)$, $1 \leqq i \leqq q$ as the solution of the following problem:

$$\frac{\partial \mathscr{C}_i}{\partial t} = L\mathscr{C}_i, \qquad \text{PDE}, \tag{5.1.4}$$

$$B\mathscr{C}_i = 0, \qquad \text{BC}, \tag{5.1.5}$$

$$\mathscr{C}_i(\cdot, \tau; \tau) = c_i(\cdot, \tau), \qquad \text{IC}. \tag{5.1.6}$$

Note that the $\mathscr{C}_i$'s are well-defined.

Using a perturbation argument and Theorem XIV of Dunford, Schwartz (1963, § 8.1), we see that $\mathscr{C}_i(\cdot, t; \tau)$ is a continuous function of t for $t \in [\tau, T]$ in $L_2(D)$ and moreover that $\mathscr{C}_i(\cdot, \cdot; \tau) \in C^\infty(\bar{D}x(\tau, T])$ for all $\tau \in [0, T)$.

Next we introduce, for $1 \leqq i \leqq q$, $1 \leqq k \leqq p$:

$$Z_{ki}(t, \tau) = \tilde{P}_k(t)\mathscr{C}_i(\cdot, t; \tau).$$

Note, that Z_{ki} is well-defined for $t > \tau$.

Our following assumption requires that the Z_{ki}'s are sufficiently regular for $t \downarrow \tau$.

ASSUMPTION 5.1.3. *The functions* Z_{ki}, $1 \leqq k \leqq p$, $1 \leqq i \leqq q$ *have the following property*:

$$Z_{ki} \in C^{\delta}\{0 \leqq \tau \leqq t \leqq T\} \quad \Big| \quad Z_{ki} \in L^{\delta}_{\infty}\{0 < \tau < t < T\}. \tag{5.1.7}$$

Here for a domain Ω we define $L^{\delta}_{\infty}(\Omega)$ as the space $\tilde{W}^{\delta,\infty}(\Omega)$ defined in Adams (1975, 7.49).

Now the following result will be proven.

THEOREM 5.1. *Under the Assumptions* 5.1.1, 5.1.2, 5.1.3 *there exists a unique solution u of* (1.1.1–2–3) *with the following properties*:

$$\begin{array}{ll|ll} u \in C^{2s,s}(\bar{Q}), & & u \in H^{2s,s}(Q), & \\ P_j u \in C^{\delta}[0, T], & 1 \leqq j \leqq p & P_j u \in H^{\delta}(0, T), & 1 \leqq j \leqq p \end{array} \tag{5.1.8}$$

with $2s = \beta + 4 - \alpha$. *Moreover this solution depends continuously on the data in the following sense*:

$$\begin{array}{l|l} |u|_{2s,s} + \sum_{j=1}^{p} |P_j u|_{\delta} & \|u\|_{2s,s} + \sum_{j=1}^{p} \|P_j u\|_{\delta} \\ \quad \leqq K(|f|_{\beta,\beta/2} + |\phi|_{\hat{\beta},\hat{\beta}/2} & \quad \leqq K(\|f\|_{\beta,\beta/2} + \|\phi\|_{\hat{\beta},\beta/2} \\ \quad + |(f, \phi, \psi)|^{cp}_{\beta_2,\beta_1,\hat{\beta}_1}) & \quad + \|(f, \phi, \psi)\|^{cp}_{\beta_2,\beta_1,\hat{\beta}_1}) \end{array} \tag{5.1.9}$$

with a constant $K > 0$ *independent of* f, ϕ, ψ.

Note, that for $\alpha > 2$ in this theorem, a loss of regularity in the space directions takes place from $t = 0$ to $t > 0$, namely:

$$\begin{array}{r|r} \psi \in C^{\beta_0}(\bar{D}),\ u(\cdot, t) \in C^{\beta_2}(\bar{D}) & \psi \in H^{\beta_0}(D),\ u(\cdot, t) \in H^{\beta_2 - 1}(D) \\ \text{for } t > 0 & \text{for } t > 0 \\ \text{with } \beta_2 < \beta_0 & \text{with } \beta_2 - 1 < \beta_0. \end{array}$$

The Assumption 5.1.3 is difficult to understand at a first glance. However, it is used in a very essential way in our proof of Theorem 5.1. In the case of time-independent operators L, B and time independent control input functions c_i and observators $\tilde{P}_j$, this assumption can be relaxed substantially, see Corollary 5.1.c. In Corollary 5.1.b. we shall give some concrete conditions under which Assumption 5.1.3 is certainly fulfilled.

Proof of Theorem 5.1. In exactly the same way as in the proof of Theorem 4.1, (1.1.1–2–3) can be rewritten in the equivalent form:

$$w = S\Pi w + W \tag{5.1.10}$$

with

$$W \in \underset{0}{C}^{\beta+2,(\beta+2)/2}(\bar{Q}) \quad \Big| \quad W \in \underset{0}{H}^{\beta+2,(\beta+2)/2}(Q).$$

However, now we cannot prove that the operator $S\Pi$ is compact on a suitable space and from here on we have to proceed differently. In this case we shall exploit, in an essential way, that the control uses a finite number of observators and control inputs. This is done by deriving a system of equations for the observation functions $P_k w$ and analyzing existence, uniqueness, and regularity of solutions of this system of equations.

In order to do so we observe that (5.1.10) can be written as

$$w(\cdot, t)=W(\cdot, t)+\sum_{i=1}^{q}\sum_{j=1}^{p}\int_0^t \mathscr{C}_i(\cdot, t; \tau)\{F_{ij}P_jw\}(\tau)\,d\tau. \tag{5.1.11}$$

Suppose that w satisfies (5.1.11) and that

$$\begin{array}{l|l} w\in C^{2s,s}(\bar{Q}), & w\in H^{2s,s}(Q), \\ P_jw\in C^{\delta}[0, T], \quad 1\leqq j\leqq p & P_jw\in H^{\delta}(0, T), \quad 1\leqq j\leqq p. \end{array} \tag{5.1.12}$$

We shall show that the functions P_kw, $1\leqq k\leqq p$ satisfy the following equation:

$$(P_kw)(t)=(P_kW)(t)+\sum_{i=1}^{q}\sum_{j=1}^{p}\int_0^t Z_{ki}(t, \tau)\{F_{ij}P_jw\}(\tau)\,d\tau. \tag{5.1.13}$$

In order to show (5.1.13) we operate on both sides of (5.1.11) with $\tilde{P}_k(t)$. It will be clear that we find (5.1.13) if for $1\leqq k, j\leqq p$, $1\leqq i\leqq q$:

$$\tilde{P}_k(t)\int_0^t \mathscr{C}_i(\cdot, t; \tau)\chi_{ij}(\tau)\,d\tau=\int_0^t \tilde{P}_k(t)\mathscr{C}_i(\cdot, t; \tau)\chi_{ij}(\tau)\,d\tau$$

with $\chi_{ij}=F_{ij}w$.

Let us show that the interchangement of the integral $\int_0^t$ and the operation of $\tilde{P}_k(t)$ is allowed here.

Using the remark following (5.1.6) and the Assumption 5.1.3 it is found that $\mathscr{C}_i(\cdot, t; \tau)$ is a continuous function of $\tau\in[0, t]$ in the sense of the (somewhat unusual) norm $\|\cdot\|_{k,t}=\|\cdot\|_0+|\tilde{P}_k(t)|$. Consequently $\|\mathscr{C}_i(\cdot, t; \tau)\chi_{ij}(\tau)\|_{k,t}$ is integrable with respect to τ on $(0, t)$.

This means that the integral $\int_0^t \mathscr{C}_i(\cdot, t, \tau)\chi_{ij}(\tau)\,d\tau$ is well-defined in Bochner's sense with respect to $\|\cdot\|_{k,t}$.

An application of Corollary 2 of Yosida (1965, p. 134) shows that the interchangement of $\int_0^t$ and $\tilde{P}_k(t)$ is indeed allowed. □

On the other hand suppose, that ξ_k, $1\leqq k\leqq p$ satisfies

$$\begin{array}{l|l} \xi_k\in \underset{0}{C}^{\delta}[0, T] & \xi_k\in \underset{0}{H}^{\delta}(0, T), \end{array} \tag{5.1.14}$$

$$\xi_k(t)=(P_kW)(t)+\sum_{i=1}^{q}\sum_{j=1}^{p}\int_0^t Z_{k,i}(t, \tau)\{F_{ij}\xi_j\}(\tau)\,d\tau. \tag{5.1.15}$$

Then w defined by

$$w(\cdot, t)=W(\cdot, t)+\sum_{i=1}^{q}\sum_{j=1}^{p}\int_0^t \mathscr{C}_i(\cdot, t; \tau)\{F_{ij}\xi_j\}(\tau)\,d\tau \tag{5.1.16}$$

is an element of

$$\begin{array}{l|l} \underset{0}{C}^{2s,s}(\bar{Q}) & \underset{0}{H}^{2s,s}(Q), \end{array} \tag{5.1.17}$$

and w satisfies (5.1.11). Let us demonstrate this.

Operating with $\tilde{P}_k(t)$ on both sides of (5.1.16) we find:

$$\xi_k=P_kw. \tag{5.1.18}$$

Substitution of (5.1.18) in (5.1.16) gives that w satisfies (5.1.11). In order to show the regularity given in (5.1.17) we note that

$$\sum_{i=1}^{q}\sum_{j=1}^{p}\int_0^t \mathscr{C}_i(\cdot\,, t;\tau)\{F_{ij}\xi_j\}(\tau)\,d\tau=\Big(\sum_{i=1}^{q} S\tilde{c}_i\Big)(\cdot\,, t)$$

with $\tilde{c}_i(\cdot\,, t)=c_i(\cdot\,, t)\sum_{j=1}^{p}(F_{ij}\xi_j)(t)$. Now we have

$$\begin{array}{r|l} \tilde{c}_i\in \underset{0}{C}^{2\delta,\delta}(\bar{Q}), & \tilde{c}_i\in \underset{0}{H}^{2\delta,\delta}(\bar{Q}), \\ S\tilde{c}_i\in \underset{0}{C}^{2s,s}(\bar{Q}) & S\tilde{c}_i\in \underset{0}{H}^{2s,s}(\bar{Q}). \end{array}$$

In the case of "H-theory" and $0\leqq\delta<\frac{1}{2}$ we have used (5.1.2) here. □

We shall now show that (5.1.15) possesses a unique solution ξ in the following function space:

$$\text{(5.1.19)}\qquad \begin{array}{r|l} \{L_2(0, T)\}^p & \{L_2(0, T)\}^p. \end{array}$$

As for the regularity of this solution ξ we shall show that (5.1.14) holds.

Let us introduce the following shorthand vector notation of (5.1.17).

$$\text{(5.1.20)}\qquad \xi=\eta+A\xi$$

with

$$\eta_k=P_kW,$$

$$(A\xi)_k=\sum_{i=1}^{q}\sum_{j=1}^{p}\int_0^t Z_{k,i}(t,\tau)\{F_{ij}\xi_j\}(\tau)\,d\tau.$$

Now A defines a nonanticipative, compact, linear operator on

$$\text{(5.1.21)}\qquad \begin{array}{r|l} \{L_2(0, T)\}^p, & \{L_2(0, T)\}^p, \\ \{\underset{0}{C}^{\delta}[0, T]\}^p & \{\underset{0}{H}^{\delta}(0, T)\}^p. \end{array}$$

The only nontrivial part of this statement is the compactness of A.

It is clear, that $A\in L(\{L_2(0, T)\}^p\rightarrow\{L_\infty(0, T)\}^p)$ and this implies the compactness of $A\in L(\{L_2(0, T)\}^p)$.

It is also easy to verify that $A\in L(\{\underset{0}{H}^{\delta}(0, T)\}^p\rightarrow\{\dot{L}^{\delta}_{\infty}(0, T)\}^p)$ which proves the compactness of $A\in L(\{\underset{0}{H}^{\delta}(0, T)\}^p)$.

Now consider A as an element of $L(\{\underset{0}{C}^{\delta}[0, T]\}^p)$. Define a sequence $A^{(n)}$ by replacing $Z_{k,i}$ in the definition of A by $Z^{(n)}_{k,i}$ with $Z^{(n)}_{k,i}\in C^\infty(\{0\leqq\tau\leqq t\leqq T\})$ and $Z^{(n)}_{k,i}\rightarrow Z_{k,i}$ in $C^{\delta}(\{0\leqq\tau\leqq t\leqq T\})$ for $n\rightarrow\infty$.

It is not difficult to show that $A^{(n)}\in L(\{\underset{0}{C}^{\delta}[0, T]\}^p\rightarrow\{\underset{0}{C}^{\delta+1}[0, T]\}^p)$, i.e., $A^{(n)}$ is a compact element of $L(\{\underset{0}{C}^{\delta}[0, T]\}^p)$ and that $A^{(n)}\rightarrow A$ in $L(\{\underset{0}{C}^{\delta}[0, T]\}^p)$ for $n\rightarrow\infty$. A well-known theorem (see Kato (1966, p. 158)) gives the desired compactness result for A in this case.

So we are in the situation that we can apply Lemma 3.1. We note that

$$(5.1.22) \qquad \eta \in \{\underset{0}{C}^{\delta}[0, T]\}^p \;\Big|\; \eta \in \{\underset{0}{H}^{\delta}(0, T)\}^p$$

and the above mentioned result is found. □

We have now proved the existence of a unique solution u of (1.1.1–2–3) with the properties of (5.1.8).

It is not difficult to deduce the estimate (5.1.9), since all operations to find the solution u appear to be continuous in an obvious sense.

COROLLARY a (to Theorem 5.1). *If Assumption* 2.2.2 *is satisfied and if, moreover,*

$$(5.1.23) \qquad f, c_i \in C^{\bar{\beta},\bar{\beta}/2}(\bar{Q}),\ 1 \leqq i \leqq q \;\Big|\; f, c_i \in H^{\bar{\beta},\bar{\beta}/2}(Q),\ 1 \leqq i \leqq q,$$

with $\bar{\beta} = \beta + \alpha - 2$, *then the characterization of compatibility given in Lemma* 2.2, (2.38) *is valid and* (5.1.9) *can be replaced by*

$$(5.1.24) \qquad \begin{array}{l|l} |u|_{2s,s} + \sum_{j=1}^{p} |P_j u|_\delta & \|u\|_{2s,s} + \sum_{j=1}^{p} \|P_j u\|_\delta \\ \quad \leqq K(|f|_{\bar{\beta},\bar{\beta}/2} + |\phi|_{\hat{\bar{\beta}},\hat{\beta}/2} + |\psi|_{\beta_0}) & \quad \leqq K(\|f\|_{\bar{\beta},\bar{\beta}/2} + \|\phi\|_{\hat{\bar{\beta}},\hat{\beta}/2} + \|\psi\|_{\beta_0}) \end{array}$$

with a constant K independent of f, ϕ, ψ.

This follows from the fact that Assumptions 5.1.1 and 5.1.2, supplemented with (5.1.23), imply Assumption 2.2.1 with β replaced by $\bar{\beta}$. As a consequence we can apply Lemma 2.2. Then (5.1.24) is found by inserting (2.39) into (5.1.9).

COROLLARY b (to Theorem 5.1). *The Assumption* 5.1.3 *is certainly fulfilled if*:

(i) *the control input functions satisfy*:

$$(5.1.25) \qquad \begin{array}{l} c_i \in C^\infty(\bar{Q}), \\ \textit{distance (support } c_i, \bar{\Gamma}) > 0, \qquad 1 \leqq i \leqq q; \end{array}$$

or

(ii) *the coefficients of the operators L, B are time-independent and the control input functions have the following form*:

$$(5.1.26) \qquad c_i = \sum_{l=1}^{d} \hat{c}_{il} e_l, \qquad d < \infty, \quad 1 \leqq i \leqq q$$

with $\{e_l;\, l \in \mathbb{N}\}$ *the eigenfunctions of the stationary uncontrolled problem*:

$$(5.1.27) \qquad \begin{array}{lll} Le = \lambda e & \textit{PDE}, & \lambda \textit{ spectral parameter}, \\ Be = 0 & \textit{BC}, & \end{array}$$

and the $\hat{c}_{il}$*'s functions of time only*:

$$(5.1.28) \qquad \hat{c}_{il} \in C^\infty[0, T], \qquad 1 \leqq l \leqq d, \quad 1 \leqq i \leqq q.$$

In the first case, we get that $\mathscr{C}_i \in C^\infty(\bar{D} \times \{0 \leqq \tau \leqq t \leqq T\})$, $1 \leqq i \leqq q$, since each $c_i(\cdot, \tau)$ satisfies the compatibility conditions of the problem (5.1.4) up to any order.

In the second case, we can even specify the form of each $\mathscr{C}_i$, $1 \leqq i \leqq q$:

$$(5.1.29) \qquad \mathscr{C}_i(\cdot, t; \tau) = \sum_{l=1}^{d} \hat{c}_{i,l}(\tau) e_l(\cdot) \operatorname{Pol}_l(t - \tau)\, e^{\lambda_l(t-\tau)}$$

with $e_l \in C^\infty(\bar{D})$, λ_l the eigenvalue corresponding to e_1 and Pol_l some polynomial.

It will be clear that, indeed, in both cases Assumption 5.1.3 is fulfilled.

COROLLARY c (to Theorem 5.1). *If the operators L, B have time-independent coefficients and if the observators $\tilde{P}_j$, and the control input functions c_i are time-independent, i.e.,*

$$\tilde{P}_j(t)=\tilde{P}_j^s, \qquad c_i(x,t)=c_i^s(x) \tag{5.1.30}$$

with

$$\begin{array}{l|l} \tilde{P}_j^s\in C^\alpha(\bar{D})', & \tilde{P}_j^s\in H^\alpha(D)', \\ c_i^s\in C^\gamma(\bar{D}), & c_i^s\in H^\gamma(D), \end{array}$$

then the Assumption 5.1.3 *can be relaxed. In this situation the $Z_{k,i}$'s are only dependent of the variable $t-\tau$, i.e.,*

$$Z_{k,i}(t,\tau)=Z_{k,i}^s(t-\tau), \qquad 1\leqq k\leqq p, 1\leqq i\leqq q. \tag{5.1.31}$$

It is sufficient to require:

$$\begin{array}{l|l} Z_{k,i}^s\in L_1(0,T), & Z_{k,i}^s\in L_1(0,T) \ \mathit{if}\ \delta>\tfrac{1}{2} \\ 1\leqq k\leqq p, \quad 1\leqq i\leqq q & Z_{k,i}^s\in L_r(0,T) \ \mathit{if}\ 0\leqq\delta<\tfrac{1}{2} \\ & \qquad \mathit{with}\ r=2(1+2\delta)^{-1}, \\ & \qquad\qquad 1\leqq k\leqq p, 1\leqq i\leqq q. \end{array} \tag{5.1.32}$$

Proof of Corollary 5.1.c. The interchangement of $\tilde{P}_k(t)$ and $\int_0^t$ necessary to derive (5.1.13) is also allowed under the conditions of (5.1.32). This is easily derived by the same reasoning as before.

In the case of "H-theory" and $0\leqq\delta<\frac{1}{2}$ we use that $H^\delta(0,T)$ is continuously inbedded in $L_{\bar{r}}(0,T)$ with $\bar{r}=2(1-2\delta)^{-1}$ (see Adams (1975)) and that $r^{-1}+\bar{r}^{-1}=1$.

Now we have for the operator A of (5.1.20),

$$\begin{aligned} (A\xi)_k(t) &= \int_0^t Z_{ki}^s(t-\tau)(F_{ij}\xi_j)(\tau)\,d\tau \\ &= \int_0^t Z_{ki}^s(\tau)(F_{ij}\xi_j)(t-\tau)\,d\tau \\ &= \left\{\int_0^T Z_{ki}^s(\tau)U(\tau)(F_{ij}\xi)\,d\tau\right\}(t) \end{aligned} \tag{5.1.33}$$

with $U(\tau)$ the shift operator introduced in (3.4).

So we have the following formula:

$$A=\left(\int_0^T Z^s(\tau)U(\tau)\,d\tau\right)F, \tag{5.1.34}$$

where Z^s, F are matrices with matrix elements Z_{ki}^s, F_{ij}. We also used Assumption 5.1.3 to show the compactness of the operator A. But this can now be shown only with the use of (5.1.32–34).

In order to prove the compactness of A in this case, we approximate A by a sequence $A^{(n)}$ obtained by replacing, in (5.1.34), Z^s by $Z^{s,n}$ with $Z^{s,n}\in\{C^\infty[0,T]\}^{p\times q}$, $n\in\mathbb{N}$ and $Z^{s,n}\to Z^s$ in the sense of $\{L_2(0,T)\}^{p\times q}$ for $n\to\infty$.

It is not difficult to verify, that:

$$A^{(n)}\in L(\{L_2(0,T)\}^p\Rightarrow\{\underset{0}{H}^1(0,T)\}^p) \quad \text{that is} \quad A^{(n)}\in L(\{L_2(0,T)\}^p)$$

is compact. $A^{(n)} \to A$ in the sense of $L(\{L_2(0, T)\}^p)$ for $n \to \infty$, so (see Kato (1966)) $A \in L(\{L_2(0, T)\}^p)$ is compact.

In the same way, we can show that $A \in L(\{\underset{0}{H}^\delta(0, T)\}^p)$ is compact and that $A \in L(\{\underset{0}{C}^\delta[0, T]\}^p)$ is compact.

Now it is clear that the proof of Theorem 5.1 keeps holding in this situation just as before. □

5.2. An example which shows how bad things can be. Consider the following problem:

$$(5.2.1)\qquad \begin{aligned} &\frac{\partial u}{\partial t}(x, t) = \frac{\partial^2 u}{\partial x^2}(x, t) - \frac{\partial^2 u}{\partial x^2}(0, t), &&\text{FPDE,}\\ &u(0, t) = u(1, t) = 0, &&\text{BC,}\\ &u(\cdot, 0) = \psi, &&\text{IC.} \end{aligned}$$

Note that $D = (0, 1) \subset \mathbb{R}$, $L = \partial^2/\partial x^2$, $\Pi = cF\delta_0''$ with $c \equiv -1$, F the identity, $(\delta_0'' u)(t) = (\partial^2 u/\partial x^2)(0, t)$, i.e., $p = q = 1$.

So the operators L, B have time-independent coefficients and the control function and observator are time-independent. The feed-back control is instantaneous.

It is not difficult to give a family of solutions of (5.2.1) for special initial data ψ_λ, $\lambda \in [1, \infty)$.

We take ψ_λ the solution of

$$(5.2.2)\qquad (L - \lambda)\psi_\lambda = -1, \qquad \psi_\lambda(o) = \psi_\lambda(1) = 0.$$

Note that ψ_λ is well-defined and $\psi_\lambda \in C^\infty[0, 1]$.

Now consider u_λ defined by

$$(5.2.3)\qquad u_\lambda(x, t) = e^{\lambda t}\psi_\lambda(x).$$

It is easy to verify that u_λ satisfies (5.2.1) with the initial data ψ_λ.

Take $T > 0$. Then we find:

$$(5.2.4)\qquad |u_\lambda|_{0,0} \geqq K\lambda^{-1} \exp(\lambda T), \quad\Big|\quad \|u_\lambda\|_{0,0} \geqq K\lambda^{-1} \exp(\lambda T)$$

with some constant $K > 0$ independent of $\lambda \in [1, \infty)$.

But it is also clear that

$$(5.2.5)\qquad \begin{array}{l|l} |(0, 0, \psi_\lambda)|^{cp}_{\beta_2,\beta_1,\hat\beta_1} & \|(0, 0, \psi_\lambda)\|^{cp}_{\beta_2,\beta_1,\hat\beta_1} \\ \leqq |v_\lambda|_{\beta_2,\beta_2/2} \leqq \tilde{K}\lambda^{(\beta_2-2)/2} & \leqq \|v_\lambda\|_{\beta_2,\beta_2/2} \leqq \tilde{K}\lambda^{(\beta_2-2)/2} \end{array}$$

with some constant $\tilde{K} > 0$ independent of $\lambda \in [1, \infty)$.

In (5.2.5) v_λ is defined by

$$(5.2.6)\qquad v_\lambda(x, t) = u_\lambda(x, t)\bar{H}(\lambda t)$$

with $\bar{H} \in C^\infty[0, \infty)$,

$$\begin{aligned} &\bar{H}(\tau) = 1 \quad \text{for } 0 \leqq \tau \leqq 1,\\ &\bar{H}(\tau) = 0 \quad \text{for } \tau \geqq 2. \end{aligned}$$

Comparing the behavior for $\lambda \to \infty$ in (5.2.4) and (5.2.5) we see that no matter how we choose β_2, β_1, $\hat{\beta}_1$ *it is not true* that (5.2.1) possesses a solution u for which

$$(5.2.7) \qquad |u|_{0,0} \leqq K_0 |(0, 0, \psi)|^{cp}_{\beta_2,\beta_1,\hat{\beta}_1} \quad \Big| \quad \|u\|_{0,0} \leqq K_0 \|(0, 0, \psi)\|^{cp}_{\beta_2,\beta_1,\hat{\beta}_1}$$

for some $K_0 > 0$ independent of ψ.

So this problem (5.2.1) is very badly posed!

A result such as Theorem 5.1 clearly can not be true here. The reason is that Assumption 5.1.3 is certainly not fulfilled in this example.

6. Application of the theory to some examples. Let us consider the following examples:

$$(6.1) \qquad \begin{array}{ll} \dfrac{\partial u}{\partial t} = (\Delta - \bar{c}FP)u + f, & \text{FPDE}, \\[2ex] \dfrac{\partial u}{\partial \vec{n}} = 0, & \text{BC}, \\[2ex] u(\cdot, 0) = \Psi, & \text{IC}. \end{array}$$

Note that $p = q = 1$. The control input function is taken constant in space and time: $\bar{c} \in \mathbb{R}$, $\bar{c} > 0$.

As for F and P we shall consider the following cases.

(i) $P = \delta_y$, $(F\xi)(t) = \xi(t)$,

(ii) $P = \delta_y$, $(F\xi)(t) = \mu \int_0^t \exp(-\mu(t-\tau))\xi(\tau)\, d\tau$, $\mu \in \mathbb{R}$, $\mu > 0$,

(iii) $P = \delta_y(1 + s(\partial/\partial t))$, $(F\xi)(t) = \xi(t)$, $s \in \mathbb{R}$, $s > 0$.

Note that (i), (ii) are special cases of the examples of § 1.2, (iii) is a special case of the example considered at the end of § 1.3 (near (1.3.14)).

At the end of § 1.3 it was also shown that (iii) falls within the setting for the control operator, when the FPDE is rewritten as

$$(6.1.(\text{iii})) \qquad \frac{\partial u}{\partial t} = (\Delta + \hat{\Pi})u + g,$$

where $\hat{\Pi} = -c_1\delta_y\{1 + s\Delta\}$, $c = \bar{c}(1 + s\bar{c})^{-1}$, $g = f - s\bar{c}(1 + s\bar{c})^{-1}\delta_y f$. Let us now apply our Theorems 4.1 and 5.1 to (6.1).

Case (i), (ii): C-theory. Note that the order of Π is $\alpha = 0$ (see (1.3.11)). So we have to apply the C-theory version of Theorem 4.1 here. This leads to the following result. If $f \in C^{\varepsilon,\varepsilon/2}(\bar{Q})$, $\Psi \in C^{2+\varepsilon}(\bar{D})$ with $0 < \varepsilon < 1$ and Ψ satisfies the compatibility relation $\partial\Psi/\partial n = 0$ on ∂D; then there exists a unique solution of (6.1) in $C^{2+\varepsilon,(2+\varepsilon)/2}(\bar{Q})$ which depends continuously on f and Ψ. The compatibility relation follows from Lemma 2.1.

Case (iii): C-theory. Note that the order of $\hat{\Pi}$ is $\alpha = 2$ (see (1.3.14)). We apply now the C-theory version of Theorem 5.1 to (6.1(iii)). The Assumption 5.1.3 is fulfilled here because of the Corollary b (1 is an eigenfunction of Δ with Neumann boundary conditions corresponding to the eigenvalue 0). Assumption 5.1.4 can be omitted because of Corollary c. We find, that if $f \in C^{\varepsilon,\varepsilon/2}(\bar{Q})$, $\Psi \in C^{2+\varepsilon}(\bar{D})$ with $0 < \varepsilon < 1$ and Ψ satisfies the compatibility relation $\partial\Psi/\partial n = 0$ on ∂D; then there exists a unique solution of (6.1) in $C^{2+\varepsilon,(2+\varepsilon)/2}(\bar{Q})$ which depends continuously on f and Ψ.

Case (i), (ii): H-theory. In this situation the order of Π is $\alpha = (n/2) + \varepsilon$, $\varepsilon > 0$ arbitrarily small (see (1.3.11)). So we have to distinguish between the cases: $n \leqq 3$ and $n \geqq 4$.

In the case $n \leqq 3$, an application of the H-theory version of Theorem 4.1 yields: if $f \in L_2(Q)$ and $\Psi \in H^1(D)$, then there exists a unique solution of (6.1) in $H^{2,1}(Q)$, which depends continuously on f and Ψ. Note that no further compatibility conditions on Ψ are necessary in accordance with Lemma 2.1.

In the case $n \geqq 4$, we have to apply the H-theory version of Theorem 5.1. The Assumption 5.1.3 is fulfilled for reasons completely analogous to the ones of Case (iii), C-theory. We now use Corollary a and we obtain the result: if $f \in H^{3\alpha-6,(3\alpha-6)/2}(Q)$, $\Psi \in H^{3\alpha-5}(D)$ with $\alpha = (n/2)+\varepsilon$, $\varepsilon > 0$ sufficiently small and the data are compatible, then there exists a unique solution of (6.1) in $H^{\alpha,\alpha/2}(Q)$ which depends continuously on f and Ψ.

The compatibility of the data is always satisfied for $n = 4$. For $n = 5$ the data are compatible, if $\partial\Psi/\partial\vec{n} = 0$ on ∂D. For $n = 6, 7$ the data are compatible if $\partial\Psi/\partial\vec{n} = 0$ on ∂D and $\partial/\partial\vec{n}\{(\Delta + \Pi_0)\Psi + f|_{t=0}\} = 0$ on ∂D with $\Pi_0 = \Pi$ in Case (i), $\Pi_0 = 0$ in Case (ii). This is easily checked using Lemma 2.2.

It is left to the reader to obtain explicit expressions for the compatibility relations for general n.

Case (iii): H-theory can be dealt with in an analogous way. Details are left to the reader.

Till now we have exclusively considered the well-posedness of problems such as (1.1.1–2–3) without worrying about (i) the construction of the solution if such a solution exists (ii) the properties of the solution.

Let us now conclude this paper by showing that, at least for the examples (6.1(i), (ii), (iii)) it is easy to calculate the solution. This calculation is done by using the technique of expansion in eigenfunctions of the uncontrolled stationary problem

$$f = \sum_{n=0}^{\infty} f_n e_n, \quad \Psi = \sum_{n=0}^{\infty} \Psi_n e_n, \quad u = \sum_{n=0}^{\infty} u_n e_n. \tag{6.2}$$

Here $\{e_n; n \geqq 0\}$ denotes the sequence of, in $L_2(D)$-sense, orthonormal eigenfunctions of Δ with Neumann boundary conditions. e_0 corresponds to the eigenvalue 0, i.e., e_0 is constant $= \{\text{vol } D\}^{1/2}$.

The eigenvalue corresponding with e_n is denoted by $-\lambda_n$. The numbering is such that $n \geqq m \Rightarrow \lambda_n \geqq \lambda_m$.

Of course the coefficients f_n, Ψ_n and u_n are defined by:

$$f_n(t) = \langle f(\cdot, t), e_n \rangle, \quad \Psi_n = \langle \Psi, e_n \rangle, \quad u_n(t) = \langle u(\cdot, t), e_n \rangle.$$

Here $\langle \cdot , \cdot \rangle$ denotes the innerproduct in $L_2(D)$.

Substitution in (6.1) leads to the following system of equations for the u_n's, $n \geqq 0$:

$$\left.\begin{aligned} \frac{du_n}{dt} &= -\lambda_n u_n + f_n, \\ u_n(0) &= \Psi_n \end{aligned}\right\} \text{ for } n \geqq 1, \tag{6.3}$$

$$\begin{aligned} \frac{du_0}{dt} &= \nu F u_0 + R_0, \\ u_0(0) &= \Psi_0. \end{aligned} \tag{6.4}$$

In (6.4), we have in case (i), (ii):

$$\nu = \bar{c}, \quad R_0 = f_0 - \bar{c} \sum_{n=1}^{\infty} q_n F u_n, \quad q_n = \frac{\delta_y e_n}{\delta_y e_0},$$

and in case (iii):

$$\nu = c_1(1-sa_0), \quad R_0 = g_0 - \bar{c} \sum_{n=1}^{\infty} \hat{q}_n F u_n,$$

$$g_0 = f_0 - sc_1 \frac{\delta_y f}{\delta_y e_0}, \quad \hat{q}_n = (1-s\lambda_n)\frac{\delta_y e_n}{\delta_y e_0}.$$

These systems of equations can be solved in an elementary way. To do this for general f is left to the reader. In the case $\Psi \equiv 1$, $f \equiv 0$ we obtain:

$$\begin{array}{ll} \multicolumn{2}{l}{u(\cdot\,,t) = u_0(t)e_0, \quad \text{where } u_0(t) \text{ equals:}} \\ \text{uncontrolled:} & 1; \\ \text{case (i):} & \exp(-\bar{c}t); \\ \text{case (ii):} & \dot{v}(t) + \mu v(t) \text{ with} \\ & v(t) = \dfrac{1}{\nu_1 - \nu_2}\{\exp(\nu_1 t) - \exp(\nu_2 t)\}, \\ & \nu_{1,2} = \tfrac{1}{2}\{-\mu \pm \sqrt{\mu^2 - 4\bar{c}}\} \quad \text{if } \nu_1 \neq \nu_2, \\ & v(t) = \exp(-\tfrac{1}{2}\mu t)\{1 - \tfrac{1}{2}\mu t\} \quad \text{if } \nu_1 = \nu_2; \\ \text{case (iii):} & \exp\left(-\dfrac{\bar{c}}{1+s\bar{c}}t\right). \end{array} \tag{6.5}$$

Note that all these solutions decay for $t \to \infty$.

For a given $\bar{c}$, the most rapid decay for $t \to \infty$ which is possible under (i), (ii), (iii) arises if $\bar{c} \geqq 4$ in case (i) and if $0 < \bar{c} < 4$ in case (ii) with $\mu = 2\sqrt{\bar{c}}$.

REFERENCES

R. A. ADAMS (1975), *Sobolev Spaces*, Academic Press, New York.

N. DUNFORD AND J. T. SCHWARTZ (1963), *Linear Operators*, Wiley, New York.

A. FRIEDMAN (1975), *Stochastic Differential Equations and Applications*, Academic Press, New York.

T. KATO (1966), *Perturbation Theory for Linear Operators*, Springer, Berlin.

O. A. LADYZENSKAJA, V. A. SOLONNIKOV, N. N. URALTSEVA (1967), *Linear and quasi-linear equations of parabolic type*, Trans. Math. Mono. 23, American Mathematical Society Providence, R. I., 648 pp.

J. L. LIONS (1971), *Optimal Control of Systems Governed by Partial Differential Equations*, Springer, Berlin.

J. L.: LIONS AND E. MAGENES (1972), *Non-homogeneous Boundary Value Problems*, I, II, Springer, Berlin.

M. H. PROTTER AND H. F. WEINBERGER (1967), *Maximum Principles in Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ.

A. YOSIDA (1965), *Functional Analysis*, Springer, Berlin.

J. R. RINGROSE (1971), *Compact Non-self-adjoint Operators*, Van Nostrand–Reinhold, London.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 4, July 1980

0363-0129/80/1804-0007 $01.00/0

A POLYNOMIAL CHARACTERIZATION OF ($\mathscr{A}$, $\mathscr{B}$)-INVARIANT AND REACHABILITY SUBSPACES*

E. EMRE† AND M. L. J. HAUTUS‡

Abstract. Based on the state space model of P. Fuhrmann, a link is laid between the geometric approach to linear system theory, as developed by W. M. Wonham and A. S. Morse, and the approach based on polynomial matrices. In particular polynomial characterizations of (A, B)-invariant and reachability subspaces are given. These characterizations are used to prove the equivalence of the disturbance decoupling problem and the exact model matching problem and also to connect the polynomial matrix and the geometric approach to the construction of observers.

Finally, constructive procedures and conditions are given for computing the supremal (A, B)-invariant subspace and reachability space and for checking the solvability of the exact model matching problem.

1. Introduction. The geometric approach to linear system theory has proved very successful in solving a variety of problems (see [17] for a detailed account of this theory). The principal concepts in this theory, which are instrumental in the description of many results, are (A, B)-invariant subspaces and reachability (controllability) subspaces. An alternative approach to linear system design has been developed in [13], [14], [16]. This theory depends to a large extent on polynomial matrix techniques. It is evident that a method for translating results of one theory to another is very desirable, because such a method would yield a better understanding of the relations between the two different approaches. This would be very useful, in particular since the geometric method may be viewed as exponent of the so-called "modern control theory" and the polynomial matrix method may be considered a generalization of the classical frequency domain methods.

A number of papers with the objective of translating the results of geometric control theory into polynomial matrix terms have appeared (e.g., [1], [3], [10], [12]).

It is the purpose of this paper to show that a very useful link between the two approaches can be based on the work of P. Fuhrmann ([7], [8], [9]). Specifically, it will be shown that using the state space model associated with a system matrix, introduced by Fuhrmann, one can give characterizations of the concepts of (A, B)-invariant subspaces and reachability subspaces in terms of polynomial matrices. This will be the subject of § 3 and 6. One application of the polynomial characterization of (A, B)-invariant subspaces will be given in § 4, where it will be shown that the disturbance decoupling problem (see [17, Chap. 4]) and the exact model matching problem (see [16], [14], [5]) are equivalent problems. Another application is given in § 5, where it is shown that the equivalence of the polynomial matrix and the geometric formulation of observers can be derived from the results of § 3. In § 7, the concept of row properness defined in [14], [16] is used to formulate a necessary and sufficient condition for the existence of a solution of the exact model matching problem and, hence, of the disturbance decoupling problem in terms of degrees of polynomial matrices. Also in § 7 a constructive characterization of the supremal (A, B)-invariant subspace and reachability space contained in ker C is given. Finally, in § 8, the results of § 3 are extended to the situation where the system is described by Rosenbrock's system matrix.

The preliminary § 2 contains a short description of Fuhrmann's state space model.

* Received by the editors October 18, 1978, and in revised form October 5, 1979.

† Center for Mathematical System Theory, University of Florida, Gainesville, Florida 32611. The work of this author was supported in part by the United States Army Research Office under Grant DAA29-77-G-0225, and by the United States Air Force under Grant AFOSR76-3034 Mod. B through the Center for Mathematical Systems Theory, University of Florida, Gainesville, Florida.

‡ Department of Mathematics, Eindhoven University of Technology, Eindhoven, the Netherlands.

2. The state space model associated with a matrix fraction representation. Let K be a field. We denote by $K[s]$ the set of polynomials, and by $K(s)$ the set of rational functions over K. If $\mathscr{S}$ is any set and $p, q \in \mathbb{N}$, we denote by $\mathscr{S}^p$ the set of p-vectors with components in $\mathscr{S}$ and by $\mathscr{S}^{p\times q}$ the set of $p \times q$ matrices with entries in $\mathscr{S}$. If A is a $p \times q$ matrix, we denote by $\{A\}$ the K-linear space generated by the columns of A. If $U(s) \in K^{q\times r}[s]$ and $\mathscr{L}: K^q[s] \to K^p[s]$ is a linear map, then $\mathscr{L}U(s)$ denotes the result obtained by applying $\mathscr{L}$ to each of the columns of $U(s)$.

Let $x(s) \in K^p(s)$. We denote by $(x(s))_-$, the strictly proper part of $x(s)$ and by $(x(s))_{-1}$, the coefficient of s^{-1} in the expansion of $x(s)$ in powers of s^{-1}.

DEFINITION 2.1. *Let $T(s) \in K^{p\times q}[s]$. Then X_T denotes the* set *of $x(s) \in K^p[s]$ for which there exists a strictly proper $u(s) \in K^q(s)$ such that $T(s)u(s) = x(s)$.*

In what follows, X_T plays a fundamental role (compare the closely related concept of right rational annihilator [4]).

In particular, if $p = q$ and $T(s)$ is nonsingular, then

$$X_T = \{x(s) \in K^p[s] \mid T^{-1}(s)x(s) \text{ is strictly proper}\}.$$

In this particular situation we define the map

$$\pi_T : K^p[s] \to X_T : x(s) \mapsto T(s)(T^{-1}(s)x(s))_-.$$

(Compare [7] and [8] where further properties of this map are given.) In the following we consider X_T a K-linear space. We consider a linear system whose transfer matrix is given by the left matrix fraction representation

$$G(s) = T^{-1}(s)U(s). \tag{2.2}$$

We assume that $G(s)$ is strictly proper, $T(s) \in K^{q\times q}[s]$, $U(s) \in K^{q\times r}[s]$.

Define the linear maps

$$\begin{aligned} &\mathscr{A} : X_T \to X_T : x(s) \mapsto \pi_T(sx(s)),\\ &\mathscr{B} : K^r \to X_T : u \mapsto U(s)u,\\ &\mathscr{C} : X_T \to K^q : x(s) \mapsto (T^{-1}(s)x(s))_{-1}. \end{aligned} \tag{2.3}$$

By definition, for $x(s) \in X_T$ we have $\mathscr{A}x(s) = sx(s) - T(s)c(s)$ for some $c(s) \in K^q[s]$. Since $T^{-1}(s)x(s)$ and $T^{-1}\mathscr{A}x(s)$ are strictly proper it follows that $c(s)$ must be constant. Hence

$$\mathscr{A}x(s) = sx(s) - T(s)c \tag{2.4}$$

for some $c \in K^q$, depending on $x(s)$.

The following result is proved in [7].

THEOREM 2.5. *The system $\Sigma := (\mathscr{C}, \mathscr{A}, \mathscr{B})$ with state space X_T is an observable realization of $G(s)$. The realization is reachable iff $T(s)$ and $U(s)$ are left coprime.*

We will call this realization Σ the *T-realization* of $G(s)$.

Conversely, if we are given an observable system $\Sigma = (C, A, B)$, then we construct a left matrix fraction representation of the transfer matrix of Σ in the following way. Let

$$C(sI - A)^{-1} = T^{-1}(s)S(s), \tag{2.6}$$

where S and T are left coprime. Define

$$U(s) := S(s)B. \tag{2.7}$$

Then $G(s) = T^{-1}(s)U(s)$ is the required representation. We have the following result.

THEOREM 2.8. *The T-realization of $G(s)=T^{-1}(s)U(s)$, where T and U are defined by* (2.6) *and* (2.7) *is isomorphic to the system Σ.*

Proof. Using the dual of [11, Cor. 4.11], we see that $S(s)$ is a basis matrix of X_T. Hence the linear map

$$\mathcal{S}: K^n \to X_T : x \mapsto S(s)x$$

is an isomorphism. Using the equation

$$T(s)C = S(s)(sI-A),$$

which follows from (2.6), one derives easily the relations $\mathcal{A}\mathcal{S}=\mathcal{S}A$, $\mathcal{B}=\mathcal{S}B$, $\mathcal{C}\mathcal{S}=C$. In particular, $\mathcal{A}\mathcal{S}x=\mathcal{A}(S(s)x)=\pi_T(S(s)(sI-A)x)+\pi_T(S(s)Ax)=\pi_T(T(s)Cx)+\pi_T(S(s)Ax)=S(s)Ax=\mathcal{S}Ax$.

It follows that (C, A, B) and $(\mathcal{C}, \mathcal{A}, \mathcal{B})$ are isomorphic. □

Using Theorem 2.8, we may transform results obtained for the particular realization $(\mathcal{C}, \mathcal{A}, \mathcal{B})$ to any observable system.

3. $(\mathcal{A}, \mathcal{B})$-invariant subspaces. We give a characterization of the $(\mathcal{A}, \mathcal{B})$-invariant subspaces of the T-realization of a transfer matrix $G(s)=T^{-1}(s)U(s)$, as defined in the previous section. For the definition of $(\mathcal{A}, \mathcal{B})$-invariant subspaces we refer to [17].

THEOREM 3.1. *Let $\Psi(s)$ be a $q\times m$ polynomical matrix. Then $\{\Psi(s)\}$ is an $(\mathcal{A}, \mathcal{B})$-invariant subspace of X_T iff there exist $C_1\in K^{q\times m}$, $F_1\in K^{r\times m}$ and $A_1\in K^{m\times m}$ such that*

$$T(s)C_1+U(s)F_1=\Psi(s)(sI-A_1). \tag{3.2}$$

Proof. Suppose that $\{\Psi(s)\}$ is an $(\mathcal{A}, \mathcal{B})$-invariant subspace, i.e.,

$$\mathcal{A}\{\Psi(s)\}\subseteq\{\Psi(s)\}+\operatorname{Im}\mathcal{B}. \tag{3.3}$$

Applying (2.4) to each column of $\Psi(s)$, we fine that $\mathcal{A}\Psi(s)=\Psi_1(s)$, where

$$\Psi_1(s) := s\Psi(s)-T(s)C_1 \tag{3.4}$$

for some $C_1\in K^{q\times m}$. On the other hand, (3.3) implies

$$\Psi_1(s)=\Psi(s)A_1+U(s)F_1 \tag{3.5}$$

for some $A_1\in K^{m\times m}$ and $F_1\in K^{r\times m}$. Combining (3.4) and (3.5) yields (3.2). Conversely, if we assume (3.2), then

$$T^{-1}(s)\Psi(s)=(C_1+T^{-1}(s)U(s)F_1)(sI-A_1)^{-1} \tag{3.6}$$

is strictly proper and hence $\{\Psi(s)\}\subseteq X_T$. Furthermore, if we define $\Psi_1(s)$ by (3.4), then (3.5) follows from (3.2) and, hence, $\{\Psi_1(s)\}\subseteq X_T$. It follows that

$$\mathcal{A}(\Psi(s))=\pi_T(s\Psi(s))=\pi_T(\Psi_1(s)+T(s)C_1)=\Psi_1(s).$$

Thus, (3.5) implies (3.3). □

If the matrix $\Psi(s)$ occurring in Theorem 3.1 has full column rank, it is possible to give an interpretation to the matrices A_1, F_1, C_1. For in that case there exists a K-linear map $\mathcal{F}: X_T\to K^r$ satisfying

$$\mathcal{F}\Psi(s)=F_1. \tag{3.7}$$

Then (3.2) implies

$$(\mathcal{A}-\mathcal{B}\mathcal{F})\Psi(x)=\Psi(s)A_1. \tag{3.8}$$

It follows that $\{\Psi(s)\}$ is an $(\mathcal{A}-\mathcal{B}\mathcal{F})$-invariant subspace, and that A_1 is the matrix of the restriction of $\mathcal{A}-\mathcal{B}\mathcal{F}$ to $\{\Psi(s)\}$ with respect to the basis matrix $\Psi(s)$. In addition, F_1 is the matrix (with respect to the basis matrix $\Psi(s)$ of $\{\Psi(s)\}$ and the natural basis in K^r) of $\mathcal{F}$. Finally, we have

$$\mathcal{C}\Psi(s)=C_1 \tag{3.9}$$

so that C_1 is the matrix of the restriction of $\mathcal{C}$ to $\{\Psi(s)\}$ with respect to the basis matrix $\Psi(s)$ of $\{\Psi(s)\}$ and the natural basis of K^q.

The last result gives a characterization of $(\mathcal{A},\mathcal{B})$-invariant subspaces contained in ker $\mathcal{C}$.

COROLLARY 3.10. *Let* $\Psi(s)\in K^{q\times m}[s]$. *Then* $\{\Psi(s)\}$ *is an* $(\mathcal{A},\mathcal{B})$*-invariant subspace contained in* ker $\mathcal{C}$ *iff there exist matrices* F_1, A_1 *such that*

$$U(s)F_1=\Psi(s)(sI-A_1). \tag{3.11}$$

Proof. According to (3.9), we must have $C_1=0$ in formula (3.2). □

COROLLARY 3.12. X_U *is the largest* $(\mathcal{A},\mathcal{B})$*-invariant subspace of* X_T *contained in* ker $\mathcal{C}$.

Proof. According to (3.10), we have for an arbitrary $(\mathcal{A},\mathcal{B})$-invariant subspace $\{\Psi(s)\}$ contained in ker $\mathcal{C}$:

$$\Psi(s)=U(s)F_1(sI-A_1)^{-1}.$$

Hence $\{\Psi(s)\}\subseteq X_U$ (see Definition 2.1). It remains to be shown that X_U itself is an $(\mathcal{A},\mathcal{B})$-invariant subspace. If $\Phi(s)$ is a basis matrix of X_U then there exists a strictly proper matrix $Q(s)$ such that $U(s)Q(s)=\Phi(s)$. Let (F_1, A_1, B_1) be a reachable realization of $Q(s)$, so that

$$\Phi(s)=U(s)F_1(sI-A_1)^{-1}B_1.$$

It follows from Lemma 3.13 that

$$\Psi(s):=U(s)F_1(sI-A_1)^{-1}$$

is a polynomial matrix. By Corollary 3.10, $\{\Psi(s)\}$ is an $(\mathcal{A},\mathcal{B})$-invariant subspace. Hence $\{\Psi(s)\}\subseteq\{\Phi(s)\}$. On the other hand, since $\Phi(s)=\Psi(s)B_1$, it follows that $\{\Phi(s)\}\subseteq\{\Psi(s)\}$. Consequently, $X_U=\{\Phi(s)\}=\{\Psi(s)\}$ is an $(\mathcal{A},\mathcal{B})$-invariant subspace. □

LEMMA 3.13. *Let* $Q(s)\in K^{l\times n}[s]$, $A\in K^{n\times n}$, $B\in K^{n\times r}$, (A,B) *reachable. If* $Q(s)$ $(sI-A)^{-1}B$ *is a polynomial matrix then* $Q(s)(sI-A)^{-1}$ *is a polynomial matrix.*

Proof. We decompose the rational matrix $Q(s)(sI-A)^{-1}$ into its polynomial and its strictly proper part:

$$Q(s)(sI-A)^{-1}=P(s)+R(s);$$

then

$$R_0:=R(s)(sI-A)=Q(s)-P(s)(sI-A)$$

is a polynomial of degree zero and hence constant. It follows that

$$R_0(sI-A)^{-1}B=Q(s)(sI-A)^{-1}B-P(s)B$$

is a strictly proper polynomial and hence zero. Since (A,B) is reachable, this implies $R_0=0$ and hence

$$Q(s)(sI-A)^{-1}=P(s).$$

□

The foregoing implies that the set of $(\mathcal{A}, \mathcal{B})$-invariant subspaces in $\ker \mathcal{C}$ is uniquely determined by the numerator polynomial matrix of the matrix fraction representation of the transfer function matrix:

COROLLARY 3.14. *Let $U(s) \in K^{q\times r}[s]$, $T_i(s) \in K^{q\times q}[s]$, $i = 1, 2$, such that*

$$G_i(s) := T_i^{-1}(s)U(s)$$

is strictly proper for $i = 1, 2$. Let $(\mathcal{C}_i, \mathcal{A}_i, \mathcal{B}_i)$ be the T_i-realization of $G_i(s)$ for $i = 1, 2$. Then $M \subseteq X_U$ is an $(\mathcal{A}_1, \mathcal{B}_1)$-invariant subspace of X_{T_1} contained in $\ker \mathcal{C}_1$ *iff M is an $(\mathcal{A}_2, \mathcal{B}_2)$-invariant subspace of X_{T_2} contained in* $\ker \mathcal{C}_2$.

REMARK 3.15. Theorem 3.1 may be specialized to the case $U(s) = 0$, that is, $\mathcal{B} = 0$. In this case we have a realization of $G(s) = 0$ with the same state space X_T and the same map $\mathcal{C}$ as before. An $(\mathcal{A}, \mathcal{B})$-invariant subspace of X_T then is just an $\mathcal{A}$-invariant subspace. Thus we obtain the following characterization of $\mathcal{A}$-invariant subspaces.

PROPOSITION. *Let $\Psi(s)$ be a $q \times m$ polynomial matrix. Then, $\{\Psi(s)\}$ is an $\mathcal{A}$-invariant subspace of X_T iff there exist $C_1 \in K^{q\times m}$, $A_1 \in K^{m\times m}$ such that*

$$T(s)C_1 = \Psi(s)(sI - A_1).$$

4. Exact model matching and disturbance decoupling. If we have an observable system (C, A, B) with state space X then we may consider the problem of characterizing the (A, B)-invariant subspaces contained in $\ker C$. Using the isomorphism given in Theorem 2.8, we transform the problem to the case of a suitable T-realization. For this case we may appeal to Corollary 3.10, by which a complete characterization is given. It is important that, as already noted in Corollary 3.14, this characterization depends only on the numerator polynomial $U(s)$. Consequently, we have the following result.

THEOREM. *Let $\bar{\Sigma} = (C, A, B)$ be a realization with state space X of a transfer matrix $G(s) = T^{-1}(s)U(s)$, and let $\Sigma = (\mathcal{C}, \mathcal{A}, \mathcal{B})$ be the T-realization of $G(s)$. If $\bar{\Sigma}$ and Σ are isomorphic by the isomorphism $\mathcal{L}: X \to X_T$, then $M \subseteq X$ is an (A, B)-invariant subspace contained in* $\ker \mathcal{C}$ *iff there exist constant matrices F_1, A_1 satisfying*

$$U(s)F_1 = \Psi(s)(sI - A_1),$$

where $\Psi(s)$ is a basis matrix of $\mathcal{L}(M)$.

Thus we see how characterizations for $(\mathcal{A}, \mathcal{B})$-invariant subspaces of the particular state space model Σ can be generalized to arbitrary (observable) state space models.

In this section we use the theory developed thus far to show the equivalence of the exact model matching problem and the disturbance decoupling problem.

PROBLEM 4.1. (Disturbance decoupling problem (DDP)). *Given the system*

$$\dot{x}(t) = Ax(t) + Bu(t) + Eq(t), \qquad y(t) = Cx(t), \tag{4.2}$$

where (C, A) is observable, determine a constant matrix F such that if

$$u(t) = Fx(t), \qquad t \geqq 0,$$

the output $y(t)$ does not depend on $q(t)$, $t \geqq 0$.

The following result has been given in [17, Thm. 4.2] in a slightly different but equivalent formulation:

THEOREM 4.3. *Problem* 4.1 *has a solution iff there exists a subspace M of the state space such that*

$$AM \subseteq M + \{B\}, \qquad \{E\} \subseteq M \subseteq \ker C. \qquad \square$$

In this paper we will also consider a slightly modified problem (compare also [18]).

PROBLEM 4.4. (Modified disturbance decoupling problem (MDDP)). *Given system* (4.2), *determine constant matrices F and D such that if*

$$u(t)=Fx(t)+Dq(t),$$

the output does not depend on $q(t)$.

In the modified problem, one assumes that not only the state but also the disturbance is directly available for measurement. Similarly to (4.3) we have the following result.

THEOREM 4.5. *Problem* 4.4 *has a solution iff there exists a subspace M such that*

$$AM\subset M+\{B\},\quad \{E\}\subset M+\{B\},\quad M\subset \ker C. \qquad \square$$

The exact model matching problem is defined as follows.

PROBLEM 4.6. *Given transfer function matrices $G_1(s)$ and $G_2(s)$ determine a* (i) *strictly proper or* (ii) *proper rational matrix $Q(s)$ such that*

$$G_1(s)Q(s)=G_2(s).$$

Problem 4.6(i) will be called the *exact model matching problem* (EMMP), and Problem 4.6(ii) will be called the *modified exact model matching problem* (MEMMP). It is the purpose of this section to show that the existence of a solution of Problem 4.1 is equivalent to the existence of a solution of Problem 4.6(i). Similarly: Problem 4.4 has a solution iff Problem 4.6(ii) has a solution. We will concentrate on the modified problems. The original problems can be dealt with similarly.

First we have to indicate which MEMMP corresponds to a given MDDP and vice versa. Let us start with system (4.2). The data $G_1(s)$ and $G_2(s)$ of MEMMP are then defined by

$$G_1(s):=C(sI-A)^{-1}B,\qquad G_2(s):=C(sI-A)^{-1}E.$$

Conversely, if we are given $G_1(s)$ and $G_2(s)$ in MEMMP, we construct an observable realization $(C, A, [B, E])$ of the transfer matrix $[G_1(s), G_2(s)]$. Then C, A, B, E are the data for MDDP. Thus, we have a one to one correspondence between MEMMP's and MDDP's.

Following Theorem 2.8, we assume that

$$C(sI-A)^{-1}=T^{-1}(s)S(s)$$

with $T(s)$ and $S(s)$ relatively prime, and $U(s)=S(s)B$; and we consider the T-realization $(\mathscr{C}, \mathscr{A}, \mathscr{B})$ of $G_1(s)=T^{-1}(s)U(s)$. According to Theorem 2.8, the map $\mathscr{S}: x\mapsto S(s)x: K^n\to X_T$ is an isomorphism. Consequently, we introduce the polynomial matrix $R(s):=S(s)E$ as representative of E in X_T. Then we have $G_2(s)=T^{-1}(s)R(s)$ and we can state the following result.

THEOREM 4.7. *Let $\{\Psi(s)\}$ be an $(\mathscr{A}, \mathscr{B})$-invariant subspace in* $\ker \mathscr{C}$, *so that there exist constant matrices F_1 and A_1 satisfying*

$$U(s)F_1=\Psi(s)(sI-A_1). \tag{4.8}$$

In addition, assume that $\{R(s)\}\subseteq\{\Psi(s)\}+\{U(s)\}$, so that there exist matrices B_1 and D_1 such that

$$R(s)=\Psi(s)B_1+U(s)D_1. \tag{4.9}$$

Then $Q(s):=F_1(sI-A_1)^{-1}B_1+D_1$ is a solution of MEMMP. Conversely, let $Q(s)$ be a solution of MEMMP and let (F_1, A_1, B_1, D_1) be a reachable realization of $Q(s)$. Then there exists a polynomial matrix $\Psi(s)$ satisfying (4.8) *and* (4.9).

Proof. If $\Psi(s)$ satisfies (4.8) and (4.9) then

$$U(s)Q(s)=\Psi(s)B_1+U(s)D_1=R(s),$$

which implies $G_1(s)Q(s)=G_2(s)$. Conversely the latter equation implies $U(s)Q(s)=R(s)$. Hence,

$$U(s)F_1(sI-A_1)^{-1}B_1=R(s)-U(s)D_1. \tag{4.10}$$

Since (A_1, B_1) is reachable it follows from Lemma 3.13 that

$$\Psi(s):=U(s)F_1(sI-A_1)^{-1} \tag{4.11}$$

is a polynomial. Now (4.10) and (4.11) imply (4.9) and (4.8). □

COROLLARY 4.12. *MEMMP has a solution iff the corresponding MMDP has a solution.*

Similarly one proves

PROPOSITION 4.13. *EMMP has a solution iff the corresponding DDP has a solution.*

Thus, if we want to solve (M)EMMP, we may construct the data A, B, C, E of (M)DDP and solve the latter problem. Then we do not only obtain a solution $Q(s)$ of (M)EMMP but also a realization of this solution. In this respect, it is important to note that the solution of (M)EMMP only depends on the numerator polynomials $U(s)$ and $R(s)$. Consequently, by a suitable choice of $T(s)$ (not necessarily equal to the original denominator polynomial) we may try to obtain a simple (M)DDP; compare [3]. We will formulate this idea more explicitly in § 6. Also in § 6, we will give existence conditions for a solution of (M)EMMP and, hence, of (M)DDP in terms of $U(s)$ and $R(s)$.

The following result states that if disturbance decoupling is at all possible by a (dynamic) control depending causally upon $q(t)$, then it is possible by a feedback control of the form $u=Fx+D_1q$.

COROLLARY 4.14. *Let there exist a proper rational matrix $H(s)$ such that, if the control $u=H(s)q$ is used in* (4.2), *the output does not depend on q. Then MDDP has a solution. If there exists a strictly proper matrix $H(s)$ with this property, then DDP has a solution.*

Proof. If the control $u=H(s)q$ is used in (4.2), then the transfer function matrix from q to y is $G_1(s)H(s)+G(s)$. If y does not depend on q, then this transfer matrix must be zero, hence

$$G_1(s)H(s)=-G_2(s),$$

that is, $-H(s)$ is a solution of MEMMP. Consequently, by Corollary 4.12, MDDP has a solution. □

5. Observers. We consider several formulations of the observer problem, which is a well-known problem in linear system theory. Further references on the subject can be found in [14], [15], [16], [19], [20], [2] and [6].

Thus far two types of formulation of this problem have appeared in the literature: the geometric formulation (see [19], [20], and [2]) and the polynomial matrix formulation (see [14], [15], and [16]).

Here, our purpose is (based on the results on the connections of the geometric theory of linear systems and polynomial matrix approaches developed in § 3) to show explicitly the algebraic equivalence of the geometric and the polynomial matrix formulations of this problem, including the case where some of the inputs may be unknown.

Let $\Sigma = (C, A, B)$ be a given system over $\mathbb{R}$. Let $\mathbb{C}^-$ be a subset of $\mathbb{C}$ satisfying $\mathbb{C}^- \cap \mathbb{R} \neq \emptyset$. We call a rational function $u(s)$ *stable* (with respect to $\mathbb{C}^-$) if $u(s)$ has no poles in $\mathbb{C} \backslash \mathbb{C}^-$. In the continuous time interpretation of Σ, one might choose $\mathbb{C}^- = \{s \in \mathbb{C} | \operatorname{Re} s < 0\}$ and in discrete time $\mathbb{C}^- = \{s \in \mathbb{C} | \, |s| < 1\}$, but also different choices of $\mathbb{C}^-$ are possible.

We assume that $C \in \mathbb{R}^{q \times n}$, $B \in \mathbb{R}^{n \times r}$ and that in addition to Σ we are given a feedthrough matrix $D \times \mathbb{R}^{q \times r}$. In continuous time, the interpretation (Σ, D) reads:

$$(5.1) \qquad \dot{x} = Ax + Bu, \qquad y = Cx + Du,$$

and the transfer function of (Σ, D) is

$$G(s) = G_{\Sigma,D}(s) = C(sI - A)^{-1}B + D.$$

DEFINITION 5.2. *Let $L \in \mathbb{R}^{p \times n}$. A system $(\bar{\Sigma}, \bar{D}) = (\bar{A}, \bar{B}, \bar{C}, \bar{D})$ is an L-observer of (Σ, D) if for every initial value x_0 of Σ, $\bar{x}_0$ of $\bar{\Sigma}$ and every control function u, the output $\bar{y}$ of*

$$(5.3) \qquad \dot{\bar{x}} = \bar{A}\bar{x} + \bar{B}y, \qquad \bar{y} = \bar{C}\bar{x} + \bar{D}y$$

satisfies: *$\bar{y} - Lx$ is stable (in particular rational).*

The observer uses only the output of (Σ, D). If one wants to consider the situation in which partial or total knowledge of the input of Σ is available, one can incorporate this in the problem by a suitable choice of D. In particular, if the input is completely known, one introduces new matrices $\tilde{C}, \tilde{D}$ and a new output $\tilde{y}$ of Σ according to

$$\tilde{y} = \begin{bmatrix} y \\ u \end{bmatrix}, \quad \tilde{D} = \begin{bmatrix} D \\ I \end{bmatrix}, \quad \tilde{C} = \begin{bmatrix} C \\ 0 \end{bmatrix}$$

so that $\tilde{y} = \tilde{C}x + \tilde{D}u$ represents the total data available for the estimation of Lx.

Let us use the following notation:

$$G(s) := G_{\Sigma,D}(s) = C(sI - A)^{-1}B + D,$$

$$\bar{G}(s) := G_{\bar{\Sigma},\bar{D}}(s) = \bar{C}(sI - \bar{A})^{-1}\bar{B} + \bar{D},$$

$$G_L(s) := L(sI - A)^{-1}B.$$

Then we have the following result.

THEOREM 5.4. *Let the system Σ be reachable and let $\bar{\Sigma}$ be observable. Then the following statements are equivalent*:

(i) *$(\bar{\Sigma}, \bar{D})$ is an L-observer of (Σ, D).*
(ii) *$G_L(s) = \bar{G}(s)G(s)$ and $\sigma(\bar{A}) \subseteq \mathbb{C}^-$.*
(iii) *There exists a real matrix M such that*

$$\bar{D}D = 0,$$

$$MA = \bar{A}M + \bar{B}C,$$

$$L = \bar{C}M + \bar{D}C,$$

$$MB = \bar{B}D$$

$$\text{and } \sigma(\bar{A}) \subseteq \mathbb{C}^-.$$

Proof. (i)⇒(ii). If $x_0 = 0$, $\bar{x}_0 = 0$, then we have

$$\hat{y} - L\hat{x} = H(s)\hat{u},$$

where $H(s) = \bar{G}(s)G(s) - G_L(s)$, and $\hat{y}, \hat{x}, \hat{u}$ are Laplace transforms. Choosing, in particular, $u = t^n e^{\lambda t}u_0$ with $\lambda \in \mathbb{C}$, we find

$$\hat{y} - L\hat{x} = \gamma_n (s-\lambda)^{-n-1} H(s)u_0.$$

Since $\hat{y} - L\hat{x}$ has to be stable for all $n \in \mathbb{N}$, $\lambda \in \mathbb{C}$, $u_0 \in \mathbb{R}^r$, it follows that $H(s) = 0$.

Furthermore, if $x_0 = 0$, $u = 0$, then

$$\hat{y} - L\hat{x} = \bar{C}(sI - \bar{A})^{-1}\bar{x}_0.$$

Since $(\bar{C}, \bar{A})$ is observable, the stability of $\hat{y} - L\hat{x}$ implies that $\sigma(\bar{A}) \subseteq \mathbb{C}^-$.

(ii)$\Rightarrow$(iii). We use the matrix fraction representation

$$G_0'(s) := B'(sI - A')^{-1}C' = T^{-1}(s)U(s)$$

defined by $B'(sI - A')^{-1} = T^{-1}(s)S(s)$, $U(s) = S(s)C'$, and we consider the T-realization $(\mathscr{C}, \mathscr{A}, \mathscr{B})$ of $G_0'(s)$ (see Theorem 2.8). Then the equation $\bar{G}(s)G(s) = G_L(s)$ may be rewritten as

$$(U(s) + T(s)D')\bar{G}'(s) = S(s)L'.$$

Hence, if we write

$$\Psi(s) := (U(s) + T(s)D')\bar{B}'(sI - \bar{A}')^{-1}, \tag{5.5}$$

then

$$\Psi(s)\bar{C}' = SL' - (U + TD')\bar{D}'. \tag{5.6}$$

Since $(\bar{C}, \bar{A})$ is observable, it follows from Lemma 3.13 that $\Psi(s)$ is a polynomial matrix. Equation (5.5) implies

$$U(s)\bar{B}' + T(s)D'\bar{B}' = \Psi(s)(sI - \bar{A}').$$

Hence, $\{\Psi(s)\}$ is an $(\mathscr{A}, \mathscr{B})$-invariant subspace, and

$$\mathscr{A}\Psi(s) = s\Psi(s) - T(s)D'\bar{B}' = \Psi(s)\bar{A}' + U(s)\bar{B}' \tag{5.7}$$

(see (3.4)). We consider again the map $\mathscr{S}$ defined in the proof of Theorem 2.8: $\mathscr{S}x = S(s)x$. Then we define $M' = \mathscr{S}^{-1}\Psi(s)$, so that $S(s)M' = \Psi(s)$. It follows from (5.7) that

$$\mathscr{S}A'M' = \mathscr{A}\mathscr{S}M' = \mathscr{A}\Psi(s) = \Psi(s)\bar{A}' + U(s)\bar{B}' = \mathscr{S}M'\bar{A}' \times \mathscr{S}C'\bar{B}'.$$

Hence, $A'M' = M'\bar{A}' + C'\bar{B}'$.

Furthermore, (5.6) implies:

$$T^{-1}(s)\Psi(s)\bar{C}' - T^{-1}(s)S(s)L' + T^{-1}U\bar{D}' = -D'\bar{D}.$$

Since the left-hand side is strictly proper, it follows that $D'\bar{D} = 0$ and

$$\mathscr{S}M'\bar{C}' - \mathscr{S}L' + \mathscr{S}C'\bar{D}' = 0,$$

Hence,

$$M'\bar{C}' - L' + C'\bar{D}' = 0.$$

Finally, it follows from (5.7) that

$$T^{-1}(s)U(s)\bar{B}' + T^{-1}(s)\Psi(s)\bar{A}' = sT^{-1}\Psi(s) - D'\bar{B}'.$$

Hence,

$$B'M' = \mathscr{C}\Psi(s) = (T^{-1}(s)\Psi(s))_{-1} = D'\bar{B}'$$

since $T^{-1}(s)U(s)$ and $T^{-1}(s)\Psi(s)$ are strictly proper.

(iii)$\Rightarrow$(i). A short calculation yields

$$\bar{y}-Lx=\bar{C}(\bar{x}-Mx),$$

$$\frac{d}{dt}(\bar{x}-Mx)=\bar{A}(\bar{x}-Mx),$$

and the result follows from $\sigma(\bar{A})\subseteq\mathbb{C}^-$. □

The equivalence (ii)$\Leftrightarrow$(i) is given in [15], and the equivalence (i)$\Leftrightarrow$(iii), with the a priori assumption (in the proof of (i)$\Rightarrow$(ii)) that $\bar{x}-Mx$ is stable, is given in [2]. Notice that here the stability of $\bar{x}-Mx$ is a consequence, rather than an assumption (see the proof of (iii)$\Rightarrow$(i)). For the situation of availability of the whole input, this was also shown in [6].

REMARK 5.8. The results of this section can easily be extended to systems over an arbitrary field K, provided an appropriate definition of stable rational function has been defined. Such a definition can be given as follows: Let $\mathcal{M}$ be a multiplicative subset of $K[s]$ (i.e., $p(s)\in\mathcal{M}$, $g(s)\in\mathcal{M}\Rightarrow p(s)g(s)\in\mathcal{M}$; $1\in\mathcal{M}$). Then we say that a rational function $r(s)\in K(s)$ is stable if $r(s)$ has the representation $r(s)=p(s)/q(s)$ with $p(s)\in K[s]$, $q(s)\in\mathcal{M}$. Then the stable functions form a ring. In the situation described above we have

$$\mathcal{M}=\{p(s)\in K[s]\,|\,p(s)=0\Rightarrow s\in\mathbb{C}^-\}.$$

In the general situation Theorem 5.4 remains valid if one replaces the condition $\sigma(\bar{A})\subseteq\mathbb{C}^-$ with "$(sI-\bar{A})^{-1}$ is stable".

A particular example, which is relevant for discrete time systems, over arbitrary fields, is

$$\mathcal{M}:=\{s^n\,|\,n=0,1,\cdots\}.$$

An observer constructed according this multiplicative set is called a deadbeat observer.

6. Reachability subspaces.

Let $\Psi(s)$ be a full column rank basis matrix of an $(\mathcal{A},\mathcal{B})$-invariant subspace. Recall the interpretation of the matrices A_1, F_1, C_1 given in (3.7), (3.8) and (3.9). Let B_1 be any constant $m\times p$ matrix such that $\{\Psi(s)B_1\}\subseteq\{U(s)\}$, say

$$\Psi(s)B_1=U(s)L_1.$$

Then B_1 is the matrix of the (codomain) restriction of $\mathcal{B}L_1$ to $\{\Psi(s)\}$. It follows that

$$(\mathcal{A}-\mathcal{B}\mathcal{F})^k\mathcal{B}L_1v=\Psi(s)A_1^kB_1v$$

for every $v\in K^p$. Consequently,

$$\langle\mathcal{A}-\mathcal{B}\mathcal{F}\,|\,\mathcal{B}L_1\rangle=\{\Psi(s)[B_1,\cdots,A_1^{m-1}B_1]\}. \tag{6.1}$$

This formula immediately implies the following result.

THEOREM 6.2. *Let $\Psi(s)$ be a (full column rank) basis matrix of an $(\mathcal{A},\mathcal{B})$-invariant subspace. Then*

(i) *$\{\Psi(s)\}$ is a reachability subspace iff there exists a constant matrix B_1 such that $\{\Psi(s)B_1\}\subseteq\{U(s)\}$, and (A_1,B_1) is reachable (here A_1 is given by (3.2)).*

(ii) *If B_1 is a constant matrix such that*

$$\{\Psi(s)B_1\}=\{U(s)\}\cap\{\Psi(s)\}, \tag{6.3}$$

then $\{\Psi(s)[B_1, \cdots, A_1^{n-1}B_1]\}$ *is the supremal reachability subspace contained in* $\{\Psi(s)\}$.

Let us now consider reachability subspaces contained in ker $\mathscr{C}$. Let $\Psi(s)$ be a basis matrix of such a space. According to Corollary 3.10, there exist matrices F_1 and A_1 such that

$$(6.4) \qquad \Psi(s) = U(s)F_1(sI - A_1)^{-1}.$$

It follows from Theorem 6.2 that there exists B_1 such that (A_1, B_1) is reachable and $\{\Psi(s)B_1\} \subseteq \{U(s)\}$, say $\Psi(s)B_1 = U(s)L_1$. Hence

$$(6.5) \qquad U(s)Q(s) = U(s)L_1,$$

where $Q(s) := F_1(sI - A_1)^{-1}B_1$. Also, since $\Psi(s)$ has full column rank, (F_1, A_1) is observable, as follows from (6.4). Hence (F_1, A_1, B_1) is a minimal realization of $Q(s)$.

COROLLARY 6.6. *There exists a nontrivial reachability subspace contained in* ker $\mathscr{C}$ *iff*

$$\{U(s)\} \cap X_U \neq \{0\}.$$

Proof. If $\Psi(s)$ is a basis matrix of the $(\mathscr{A}, \mathscr{B})$-invariant subspace X_U and $\Psi(s) = U(s)F_1(sI - A_1)^{-1}$, then the supremal reachability subspace contained in X_U (or, equivalently, in ker $\mathscr{C}$) is nontrivial iff $B_1 \neq 0$, where B_1 is a matrix satisfying (6.3). □

According to (6.5), $Q(s) - L_1$ is a nontrivial right zero matrix of $U(s)$. Consequently, if the supremal reachability subspace contained in $\mathscr{C}$ is nonzero, then $U(s)$ is not left invertible. The converse, however, is not true. For example, if $U(s) = [U_1(s), 0]$ where $U_1(s)$ is left invertible, then it is easily seen that $U(s)$ is not left invertible, and $\{U(s)\} \cap X_U = \{0\}$. In order to give a necessary and sufficient condition for the existence of a maximal reachability subspace contained in ker $\mathscr{C}$, we consider the $K[s]$-module

$$(6.7) \qquad \Delta := \{v(s) \in K^r[s] \mid U(s)v(s) = 0\}.$$

This module is generated by the columns of a matrix $M(s)$ (see [5, Thm. 3.1]).

COROLLARY 6.8. *There exists a nontrivial reachability subspace contained in* ker $\mathscr{C}$ *iff the module* Δ *defined in* (6.7) *is not generated by a constant matrix.*

Proof. Let $M(s)$ be a generator matrix of minimal degree, say $M(s) = M_0 s^k + \cdots + M_k$. Then $s^{-k}M(s) = Q(s) - L_1$, where $Q(s) = M_1 s^{-1} + \cdots + M_k s^{-k}$ and $L_1 = -M_0$. We have

$$U(s)Q(s) = U(s)L_1$$

and $U(s)L_1 \neq 0$, since, otherwise, $[M(s) - s^k M_0, M_0]$ would be a generator matrix of lower degree than k. It follows that $\{U(s)L_1\} \subseteq \{U(s)\} \cap X_U$, so that $\{U(s)\} \cap X_U \neq \{0\}$.

Conversely, suppose that Δ is generated by a constant matrix, say D, and that $v \in \{U(s)\} \cap X_U$, say $v = U(s)c = U(s)r(s)$, where c is a constant vector and $r(s)$ is a strictly proper rational vector. It follows that there exists a rational vector $q(s)$ such that $c - r(s) = Dq(s)$. Decomposing $q(s)$ into a polynomial and a strictly proper part $q(s) = q_1(s) + q_2(s)$, we conclude that $c = Dq_1(s)$, so that $v = U(s)c = 0$. Hence, $\{U(s)\} \cap X_U = \{0\}$. □

Now we have a procedure for constructing reachability subspaces contained in ker $\mathscr{C}$. Choosing any matrix L_1 such that $\{U(s)L_1\} \subseteq X_U$, we have $U(s)Q(s) = U(s)L_1$ for some strictly proper $Q(s)$. If (F_1, A_1, B_1) is a minimal realization of $Q(s)$, it follows that $\Psi(x) := U(s)F_1(sI - A_1)^{-1}$ is a basis matrix of a reachability subspace, provided the columns of $\Psi(s)$ are independent. In general, it seems difficult to formulate conditions upon L_1 and $Q(s)$ that guarantee that $\Psi(s)$ has full column rank. A sufficient condition for this is that $Q(s)$ be a strictly proper rational matrix with minimal McMillan degree satisfying the equation $U(s)Q(s)=U(s)L_1$. Indeed, if in this case $\Psi(s)$ does not have full column rank, there exists $\Phi(s)$ with less columns than Ψ such that $\{\Phi(s)\}=\{\Psi(s)\}$. Since $\{\Phi(s)\}$ is an $(\mathscr{A},\mathscr{B})$-invariant subspace, there exist F_2, A_2 such that $\Phi(s)=U(s)F_2(sI-A_2)^{-1}$. Also, there exists D_1 such that $\Psi(s)=\Phi(s)D_1$. Hence,

$$U(s)Q(s)=\Psi(s)B_1=\Phi(s)D_1B_1=U(s)Q_2(s)=U(s)L_1,$$

where $Q_2(s):=F_2(sI-A_2)^{-1}D_1B_1$ has lower McMillan degree than $Q(s)$.

THEOREM 6.9. *Let L_1 be a constant matrix such that $\{U(s)L_1\}=\{U(s)\}\cap X_U$. Let $Q(s)$ be a strictly proper rational matrix of minimal McMillan degree, satisfying the equation $U(s)Q(s)=U(s)L_1$.* Let *(F_1, A_1, B_1) be a minimal realization of $Q(s)$. Then $\Psi(s):=U(s)F_1(sI-A_1)^{-1}$ is a basis matrix of the supremal reachability space contained in* ker $\mathscr{C}$.

Proof. The supremal reachability subspace contained in ker $\mathscr{C}$ is the (unique) minimal $(\mathscr{A},\mathscr{B})$-invariant subspace $\mathscr{V}$ satisfying $(\operatorname{Im}\mathscr{B})\cap\mathscr{W}\subseteq\mathscr{V}\subseteq\mathscr{W}$, where $\mathscr{W}$ is the supremal $(\mathscr{A},\mathscr{B})$-invariant subspace contained in ker $\mathscr{C}$. To see this, observe that an $(\mathscr{A},\mathscr{B})$-invariant subspace $\mathscr{V}$ satisfying $(\operatorname{Im}\mathscr{B})\cap\mathscr{W}\subseteq\mathscr{V}\subseteq\mathscr{W}$ is $(\mathscr{A}-\mathscr{B}\mathscr{F})$-invariant for every $\mathscr{F}$ such that $\mathscr{W}$ is $(\mathscr{A}-\mathscr{B}\mathscr{F})$-invariant. Indeed, $(\mathscr{A}-\mathscr{B}\mathscr{F})\mathscr{V}\subseteq(\mathscr{A}-\mathscr{B}\mathscr{F})\mathscr{W}\subseteq\mathscr{W}$ and $(\mathscr{A}-\mathscr{B}\mathscr{F})\mathscr{V}\subseteq\mathscr{V}+\operatorname{Im}\mathscr{B}$ imply

$$(\mathscr{A}-\mathscr{B}\mathscr{F})\mathscr{V}\subseteq\mathscr{W}\cap(\mathscr{V}+\operatorname{Im}\mathscr{B})=\mathscr{V}+\mathscr{W}\cap\operatorname{Im}\mathscr{B}\subseteq\mathscr{V}.$$

Since $\{U(s)\}\cap X_U=\{U(s)L_1\}=\{\Psi(s)B_1\}\subseteq\{\Psi(s)\}\subseteq X_U$, and because of the minimal McMillan degree of $Q(s)$, the result follows. □

In the next section, it will be shown how Theorem 6.9 can be used for the explicit construction of the supremal reachability subspace.

7. Constructive characterizations. Conditions for solvability and the characterization of solutions of various problems can be made explicit by the use of row and column proper matrices (see [16]). This will be the subject of this section.

If $R\in K^{p\times q}[s]$ has rows $r_1(s),\cdots,r_p(s)$, then deg $r_i(s)$ is called the ith *row degree* of $R(s)$. The coefficient vector of s^{ν_i} in $r_i(s)$, where $\nu_i=\deg r_i(s)$ is called the ith *leading coefficient row vector*, and is denoted by $[r_i]_r$. We denote by $[R]_r$ the matrix of leading coefficient row vectors, that is, the constant matrix with rows $[r_1]_r,\cdots,[r_p]_r$. Similarly, $[R]_c$ denotes the matrix of leading coefficient column vectors, that is, $[R]_c=([R']_r)'$. A matrix is called *row* (*column*) *proper* if $[R]_r([R]_c)$ is nonsingular. A row proper matrix is easily seen to be right invertible. Conversely, we have (see [16, Thm, 2.5.7])

LEMMA 7.1. *If $L(s)\in K^{p\times q}[s]$ is right invertible there exists a unimodular matrix $M(s)\in K^{p\times p}[s]$ such that $M(s)L(s)$ is row proper with row degrees $\nu_1,\cdots,\nu_p$ satisfying $\nu_1\leqq\cdots\leqq\nu_p$. If $L(s)\in K^{p\times q}[s]$ is not right invertible, there exists a unimodular matrix $M(s)$ such that*

$$M(s)L(s)=\begin{bmatrix}L_1(s)\\0\end{bmatrix},$$

where $L_1(s)$ is row proper with row degrees $\nu_1\leqq\cdots\leqq\nu_l$. The number l of rows of $L_1(s)$ equals the rank of $L(s)$.

The row degrees ν_1 are independent of $M(s)$ (which is not unique) and will be called the *row indices* of $L(s)$.

The following result (see [14, Property 2.2]) states a simple criterion for the properness of a rational matrix $T^{-1}(s)U(s)$ if the denominator polynomial matrix is row proper.

LEMMA 7.2. *Let $T(s)$ be row proper with row degrees $\nu_1, \cdots, \nu_q$. If the row degrees of $U(s)$ are $\lambda_1, \cdots, \lambda_q$ then $T^{-1}(s)U(s)$ is proper iff $\lambda_1 \leqq \nu_i$ $(i = 1, \cdots, q)$ and strictly proper iff $\lambda_i < \nu_i$ $(i = i, \cdots, q)$.*

Observe that if T is not row proper, there exists a unimodular matrix $M(s)$ such that $T_1(s) := M(s)T(s)$ is row proper. If we define $U_1(s) := M(s)U(s)$, we have $T^{-1}(s)U(s) = T_1^{-1}(s)U_1(s)$, and we may apply Lemma 7.2.

Let us now consider (M)EMMP as defined in Problem 4.6. Assume that we have a matrix fraction representation $T^{-1}(s)[U(s), R(s)]$ of $[G_1(s), G_2(s)]$. Then the equation for $Q(s)$ reads

$$U(s)Q(s) = R(s). \tag{7.3}$$

In order that this equation has a (not necessarily proper) rational solution, it is necessary and sufficient that rank $U(s) =$ rank $[U(s), R(s)]$. For the existence of a proper solution additional conditions have to be imposed. Writing down the ith row of (7.3)

$$u_i(s)Q(s) = r_i(s),$$

we note that a necessary condition for the existence of a proper solution is deg $u_i(s) \geqq$ deg $r_i(s)$. The following result shows that this is also sufficient provided that $U(s)$ has the form

$$\begin{bmatrix} U_1(s) \\ 0 \end{bmatrix},$$

with $U_1(s)$ row proper. According to Lemma 7.1, this can always be obtained by premultiplying (7.3) with a suitable unimodular matrix $M(s)$.

THEOREM 7.4. *Let $M(s)$ be a unimodular matrix such that*

$$M(s)U(s) = \begin{bmatrix} U_1(s) \\ 0 \end{bmatrix}, \qquad M(s)R(s) = \begin{bmatrix} R_1(s) \\ R_2(s) \end{bmatrix},$$

where $U_1(s)$ is row proper. Let the row degrees of $U_1(s)$ be $\nu_1, \cdots, \nu_l$ and let the row degrees of $R_1(s)$ be $\lambda_1, \cdots, \lambda_l$. Then (7.3) has a proper solution iff $R_2(s) = 0$ and $\lambda_i \leqq \nu_i$ $(i = 1, \cdots, l)$. Equation (7.3) has a strictly proper solution iff $R_2(s) = 0$ and $\lambda_i < \nu_i$ $(i = 1, \cdots, l)$.

Proof. The conditions are necessary according to the foregoing discussions. Now assume that the conditions hold. Then there exists $L \in K^{r \times l}$ such that $U_1(s)L$ is a row proper $l \times l$ matrix with row degrees $\nu_1, \cdots, \nu_l$. Define

$$Q(s) := L(U_1(s)L)^{-1}R_1(s).$$

Then $Q(s)$ satisfies (7.3). It follows from (7.2) that $Q(s)$ is proper. The proof for the strictly proper solution is similar. □

We can express the result of Theorem 7.4 in a way not involving explicitly the matrix $M(s)$:

COROLLARY 7.5. *Equation (7.3) has a proper solution iff $U(s)$ and $[U(s), R(s)]$ have the same rank and the same row indices.*

In [14], no explicit condition for the solvability is given. In [5], a condition is given in terms of the kernel of the matrix $[U(s), R(s)]$. The conditions given in Theorem 7.4 and Corollary 7.5 are directly expressed in terms of the matrices $U(s)$ and $R(s)$.

The set X_U is the largest $(\mathcal{A}, \mathcal{B})$-invariant subspace contained in ker $\mathcal{C}$. By definition $x(s) \in X_U$ iff the equation

$$U(s)v(s) = x(s)$$

has a strictly proper solution $v(s)$. Therefore, using Theorem 7.4, we can give a constructive characterization of X_U.

COROLLARY 7.6. *Let $M(s)$ be as in Theorem 7.4. Then $x(s) \in X_U$ iff $y(s) := M(s)x(s)$ satisfies the conditions*

$$\deg y_i(s) < v_i \qquad (i = i, \cdots, l),$$
$$y_i(s) = 0 \qquad (i = l+1, \cdots, q).$$

Here $y_i(s)$ denotes the ith component of $y(s)$. In particular, if we introduce the row vector $w_k(s) := [s^{k-1}, \cdots\cdot, 1]$, then $M^{-1}(s)W(s)$ is a basis matrix of X_U, where

$$W(s) := \begin{bmatrix} W_1(s) \\ 0 \end{bmatrix}$$

with $W_1(s) :=$ diag $(w_{\nu_1-1}(s), \cdots, w_{\nu_l-1}(s))$.

One way of solving (7.3), already mentioned in § 4, is the reformulation of (7.3) as a (M)DDP. In doing so, it is not necessary to use the original denominator matrix $T(s)$. One might try to find a new denominator matrix $T_1(s)$ such that $T_1^{-1}(s)U(s)$ is strictly proper and $T_1(s)$ is as simple as possible. If we choose $T_1(s)$ row proper, then according to Lemma 7.2, it suffices for the strict properness of $T_1^{-1}U$, that the row degrees of T_1 are larger than the row degrees of U. If we denote the latter by $\lambda_1, \cdots, \lambda_1$, the simplest choice of $T_1(s)$ is $T_1(s) = \operatorname{diag}(s^{\lambda_q+1}, \cdots, s^{\lambda_1+1})$.

For this computation, it is not necessary that $U(s)$ be in row proper form. But if we transform $U(s)$ such that it has the form given in Theorem 7.4, then the dimension of the state space will be minimal. These ideas are worked out in more detail in [3].

We conclude this section with a construction of the supremal reachability subspace contained in ker $\mathscr{C}$. To this end, we consider the space

$$\Lambda := \{v(s) \in K^r(s) \mid U(s)v(s) = 0\},$$

and we choose a minimal basis for Λ (see [5]), that is, a basis for Δ (see (6.7)) which is column proper. We define $L_1 := [M]_c$. Furthermore we choose any $D(s) \in K^{l \times l}[s]$ which has the same column degrees $M(s)$ and such that $[D]_c = I$. Then we observe (by Lemma 7.2) that, if

$$N(s) := L_1 D(s) - M(s),$$

then $Q(s) := N(s)D^{-1}(s)$ is strictly proper. Now we have

THEOREM 7.7. (i) $\{U(s)L_1\} = X_U \cap \{U(s)\}$,

(ii) *$Q(s)$ is a strictly proper rational matrix of minimal McMillan degree satisfying*

$$U(s)Q(s) = U(s)L_1. \tag{7.8}$$

Hence, if (F_1, A_1, B_1) is a minimal realization of $Q(s)$, then $\Psi(s) := U(s)F_1(sI - A_1)^{-1}$ is a basis of the supremal reachability subspace contained in ker $\mathscr{C}$.

Proof. (i) Since $U(s)M(s) = 0$, it is easily seen that (7.8) is satisfied. This implies that $\{U(s)L_1\} \subseteq X_U \cap \{U(s)\}$. Suppose that there exists a matrix $\bar{L}_1$ of full column rank such that $\{U(s)L_1\} \subseteq \{U(s)\bar{L}_1\}$, and $U(s)\bar{L}_1 = U(s)\bar{Q}(s)$ for some strictly proper $\bar{Q}(s)$. Let $\bar{N}, \bar{D}$ be right coprime polynomial matrices such that $\bar{Q}(s) = \bar{N}(s)\bar{D}^{-1}(s)$, and $\bar{D}(s)$ is column proper with $[\bar{D}]_c = I$. Then

$$U(s)(\bar{N}(s) - \bar{L}_1\bar{D}(s)) = 0.$$

Since $\bar{Q}(s)$ is strictly proper, the columns of $\bar{N}(s) - \bar{L}_1\bar{D}(s)$ are linearly independent over $K(s)$. But then $\bar{L}_1$ cannot have more columns than L_1. Consequently, $\{U(s)L_1\} = \{U(s)\bar{L}_1\}$.

(ii) Suppose that $\bar{Q}(s) = \bar{N}(s)\bar{D}^{-1}(s)$ has a lower McMillan degree than $Q(s)$ and that $N(s)$ and $D(s)$ are relatively prime and that $\bar{D}(s)$ is column proper with $[\bar{D}(s)]_c = I$. Then we have

$$U(s)(\bar{N}(s) - L_1\bar{D}(s)) = 0,$$

and hence, $\bar{N}(s) - L_1\bar{D}(s) = M(s)R(s)$. By the "predictable degree property" (see [5, § 3, Remark]), this implies that the sum of the column degrees of $\bar{D}(s)$, and hence deg det $\bar{D}(s)$ is not less than deg det $D(s)$, which contradicts our assumption. □

8. Generalization to systems represented by Rosenbrock's system matrix. In this section, we indicate how the result of § 3 can be generalized to the case where the system is represented by a system matrix

$$P(s) = \begin{bmatrix} T(s) & U(s) \\ -V(s) & W(s) \end{bmatrix}, \tag{8.1}$$

where $T(s) \in K^{q\times q}[s]$ is nonsingular and $P(s) \in K^{(q+l)\times(q+r)}[s]$. We assume that the transfer function matrix

$$G(s) := V(s)T^{-1}(s)U(s) + W(s)$$

and the matrix $T^{-1}(s)U(s)$ are strictly proper. If the latter condition is not satisfied, we can obtain this by strict system equivalence (see [13, § 3.1]). Indeed, if we define

$$U_1(s) := \pi_T(U(s)),$$

then

$$Q(s) := T^{-1}(s)(U(s) - U_1(s))$$

is a polynomial matrix. Therefore,

$$P_1(s) := \begin{bmatrix} T(s) & U_1(s) \\ -V(s) & W(s) + V(s)Q(s) \end{bmatrix}$$

is a polynomial system matrix with the same transfer matrix $G(s)$.

In [9], it is shown that the maps

$$\mathscr{A} : X_T \to X_T : x(s) \mapsto \pi_T(sx(s)),$$
$$\mathscr{B} : K^r \to X_T : u \mapsto U(s)u,$$
$$\mathscr{C} : X_T \to K^l : x(s) \mapsto (V(s)T^{-1}(s)x(s))_{-1}$$

yield a realization $(\mathscr{C}, \mathscr{A}, \mathscr{B})$ of $G(s)$ which is reachable iff $T(s)$ and $U(s)$ are left coprime, and observable iff $T(s)$ and $V(s)$ are right coprime.

It is easily seen that Theorem 3.1 is equally valid in this situation. Instead of Corollary 3.10 we get

THEOREM 8.2. *Let $\Psi(s)$ be a $q \times m$ polynomial matrix. Then $\{\Psi(s)\}$ is an $(\mathscr{A}, \mathscr{B})$-invariant subspace in* ker $\mathscr{C}$ *iff there exists $C_1 \in K^{q\times m}$, $F_1 \in K^{r\times m}$, $A_1 \in K^{m\times m}$ and an $l \times m$ polynomial matrix $\Phi(s)$ such that*

$$P(s)\begin{bmatrix} C_1 \\ F_1 \end{bmatrix} = \begin{bmatrix} \Psi(s) \\ \Phi(s) \end{bmatrix}(sI - A_1). \tag{8.3}$$

Proof. By Theorem 3.1, $\{\Psi(s)\}$ is an $(\mathscr{A}, \mathscr{B})$-invariant subspace of X_T iff for some

C_1, F_1, A_1 we have (3.2) and hence (3.6). But then

$$\begin{aligned}\mathscr{C}\Psi(s) &= (V(s)T^{-1}(s)\Psi(s))_{-1}\\ &=((V(s)C_1+(G(s)-W(s)F_1))(sI-A_1)^{-1})_{-1}\\ &=((V(s)C_1-W(s)F_1)(sI-A_1)^{-1})_{-1}\end{aligned}$$

since $G(s)$ and $(sI-A_1)^{-1}$ are both strictly proper. Now it follows from Lemma (8.5) that

$$\Phi(s) := (-V(s)C_1+W(s)F_1)(sI-A_1)^{-1} \tag{8.4}$$

is a polynomial iff $\mathscr{C}\Psi(s)=0$. Combining (3.2) and (8.4) yields the desired result. □

LEMMA 8.5. *Let* $Q(s)\in K^{l\times n}[s]$, $A\in K^{n\times n}$. *If*

$$(Q(s)(sI-A^{-1}))_{-1}=0,$$

then $Q(s)(sI-A)^{-1}$ *is a polynomial matrix.*

The proof is analogous to the proof of Lemma 3.13 and will be omitted.

The generalization of Corollary 3.12 can be expressed in terms of the map

$$\mathscr{P}: K^{q+l}[s]\to K^q[s]: \begin{bmatrix} x(s)\\ y(s)\end{bmatrix}\mapsto x(s).$$

COROLLARY 8.6. *The largest* $(\mathscr{A}, \mathscr{B})$*-invariant subspace of* X_T *contained in* ker $\mathscr{C}$ *is* $\mathscr{P}(X_P)$.

The proof is similar to the proof of Corollary 3.12 and will be omitted.

Acknowledgment. One of the authors (E. Emre) would like to thank the Department of Mathematics of the Eindhoven University of Technology for financial support and hospitality while this research was being done.

REFERENCES

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

[2] P. BHATTACHARYYA, *Observer design for linear systems with unknown inputs*, IEEE Trans. Automatic Control, AC-23 (1978), pp. 483–484.

[3] E. EMRE, *Nonsingular factors of polynomial matrices and* (A, B)*-invariant subspaces*, Memorandum COSOR 78-12, Dept. of Math., Eindhoven Univ. of Technology; this Journal, to appear.

[4] E. EMRE AND L. M. SILVERMAN, *Relatively prime polynomial matrices: Algorithms*, Proc. IEEE Conference on Decision and Control, Houston, TX, 1975.

[5] G. D. FORNEY, *Minimal bases of rational vector spaces, with application to multivariable linear systems*, this Journal, 13 (1977), pp. 493–520.

[6] T. E. FORTMANN AND D. WILLIAMSON, *Design for low-order observers for linear feedback control laws*, IEEE Trans. Automatic Control, AC-17 (1972), pp. 301–308.

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

[8] ———, *Linear algebra and finite dimensional linear systems*, Math. Rep. 143, Ben Gurion Univ. of the Negev, Beersheva, Israel.

[9] ———, *On strict system equivalence and similarity*, Internat. J. Control, 25 (1977), pp. 5–10.

[10] A. G. J. MACFARLANE AND N. KARCANIAS, *Relationships between state-space and frequency-response concepts*, Preprints of 7th World Congress IFAC, Pergamon Press, New York, 1978, pp. 1771–1779.

[11] M. L. J. HAUTUS AND H. HEYMANN, *Linear feedback—an algebraic approach*, this Journal, 16 (1978), pp. 83–105.

[12] A. S. MORSE, *Minimal solutions to transfer matrix equations*, IEEE Trans. Automatic Control, AC-21 (1976), pp. 131–133.

[13] H. H. ROSENBROCK, *State space and multivariable theory*, Wiley, New York, 1970.

[14] S. H. WANG AND E. J. DAVISON, *A minimization algorithm for the design of linear multivariable systems*, IEEE Trans. Automatic Control, AC-18 (1973), pp. 220–225.

[15] ———, *Observing partial states for systems with unmeasurable disturbances*, Ibid., AC-23 (1978), pp. 481–483.

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

[17] W. M. WONHAM, *Linear Multivariable Control: A Geometric Approach*, Lecture Notes in Mathematical Systems no., 101, Springer-Verlag, New York, 1974.

[18] ———, *Geometric Methods in the structural synthesis of linear multivariable controls.* Proc. of Joint Automatic Control Conference, San Francisco, June 1977.

[19] ———, *Dynamic observers: Geometric theory*, IEEE-AC-15, pp. 258–259.

[20] W. M. WONHAM AND A. S. MORSE, *Feedback invariants of linear multivariable systems*, Automatica–J.IFAC, 8 (1972), pp. 93–100.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 4, July 1980

0363–0129/80/1804–0008 $01.00/0

MINIMAL INEQUALITIES AND SUBADDITIVE DUALITY*

ACHIM BACHEM† AND RAINER SCHRADER†

Abstract. In this note, we use a duality theorem for mixed integer programs (first explicitly stated by Johnson (1973) for the one-row group problem) to characterize minimal inequalities. This characterization extends earlier results, which assumed either a rational or a bounded constraint set (Blair (1978), Jeroslow (1979), Johnson (1976)), by relaxing these assumptions either entirely or almost so. It also extends results given first by Gomory and Johnson (1969) for the group problem.

1. Introduction. Consider the Farkas theorem [6] which states: Either

$$Ax = b, \qquad x \geqq 0$$

has a sôlution or there is a u such that

$$uA \geqq 0, \qquad ub \leqq -1,$$

but not both. The importance of this theorem in linear programming theory is due to the fact that it gives a positive criterion for determining when a system of linear inequalities has no solutions.

Although it turns out that Farkas' theorem has an integer analogue (i.e., replacing "$x \geqq 0$" by "x integer", see Bachem and Randow [2]), it does not carry over to the NP-complete problem (cf. [15]): Is the system

$$Ax = b, \qquad x \geqq 0, \qquad x \text{ integer}$$

consistent? This can easily be seen to be wrong. In § 2, however, we give a positive criterion for determining the inconsistency of

$$Ax + By = b, \qquad x, y \geqq 0, \qquad x \text{ integer}$$

which yields a Farkas' type theorem for the mixed integer case. Actually this idea was motivated by a paper of Johnson [12]. Moreover, Jeroslow states in [10] a duality result for mixed integer programs from which this Farkas version can easily be derived. In § 2, we slightly extend Jeroslow's theorem to a complete analogue of the linear programming duality theorem.

Besides primal methods and branch and bound algorithms, one of the powerful tools for solving general mixed integer programs (MIP) is the cutting plane method, first mentioned by Dantzig [5] in 1963. Intuitively it is clear that the "deeper" the cut, the higher the efficiency and speed of convergence. So in order to develop necessary and sufficient conditions for a valid inequality to be a "deep" cut, one reduces the theory of valid inequalities to the theory of subadditive functions on the semigroup generated by all feasible right-hand sides for MIP (cf. Araoz [1], Blair [3], Gomory and Johnson [8], Jeroslow [9], [10], [11], Johnson [12], [13], [14]). It is known that any such subadditive function defines a valid inequality and, conversely, any valid inequality can be improved by a cut derived from a subadditive function.

The concepts of valid inequalities and subadditivity are interrelated by the strong duality theory for mixed integer programs (cf. Theorem 2) which we use in §3 to characterize valid cuts which are tight, i.e., which are at least supporting hyperplanes of

* Received by the editors June 27, 1978, and in revised form October 9, 1979. This research was supported by SFB 21 (DFG), Institut für Operations Research, Universität Bonn, Bonn.

† Institut für Ökonometrie und Operations Research, Universität Bonn, Bonn, West Germany.

the closed convex hull of the feasible points for MIP. This result extends the theorem of Blair [3], Jeroslow [10] and Johnson [13] in that our result needs no assumption on the data of MIP (cf. Theorem 3). Similar results for master mixed integer programming problems can be found in [14].

We make no notational distinction between a vector x and its transpose. Whenever it is necessary to know which is intended, the context makes it clear.

If A is an (m, n) matrix and d an n-vector, we denote by A_i the ith column of A, by d_j the jth component of d.

2. Duality in mixed integer programming. For given real matrices A and B of dimensions (m, n), (m, s), respectively, we denote by $\mathrm{RHS}(A, B)$ the set of all right-hand sides $v \in \mathbb{R}^m$ for which the set $S(v) = \{(x, y) | Ax + By = v, x, y \geqq 0, x \textit{ integer}\}$ is not empty. The upper directional derivative (at zero) of a function $f: \mathrm{RHS}(A, B) \rightarrow \mathbb{R} \cup \{-\infty\}$ is denoted by f' and defined by

$$f'(b) := \lim_{t \to 0_+} \sup f(tb)/t$$

in case $tb \in \mathrm{RHS}(A, B)$ for $0 \leqq t \leqq \varepsilon$ and some $\varepsilon > 0$. f is called *subadditive* if $f(v + w) \leqq f(v) + f(w)$ holds for all $v, w \in \mathrm{RHS}(A, B)$ (note that $\mathrm{RHS}(A, B)$ is closed under addition). As usual, we have the conventions $x + (-\infty) = -\infty$ for all $x \in \mathbb{R} \cup \{-\infty\}$, $x(-\infty) = -\infty$ for positive x and $(-\infty)0 = 0$.

We should emphasize the point that in the following we do not consider f to be a fixed function but will look at systems of constraints defining a set of subadditive functions.

THEOREM 1. *Given the data A, B, c and d either there exists a subadditive function satisfying*

$$\begin{aligned} &f(A_i) \leqq c_i, && i = 1, \cdots, n,\\ &f'(B_j) \leqq d_j, && j = 1, \cdots, s,\\ &f(0) \geqq 0, && \end{aligned} \tag{2.1}$$

or there exist real vectors x, y such that

$$\begin{aligned} &Ax + By = 0,\\ &cx + dy < 0,\\ &x, y \geqq 0, \qquad x \textit{ integer}, \end{aligned} \tag{2.2}$$

but not both.

Note that if we relax the integrality of x in (2.2) and restrict f in (2.1) to be a linear function, Theorem 1 turns out to be the well-known Farkas lemma. Additionally (2.1) and (2.2) belong most likely to different complexity classes, unless $\overline{NP} \subset NP(\overline{NP}$, the complement of NP). Before we give a proof of Theorem 1, we state some duality results (Theorem 2), most of them due to Jeroslow [9] and Johnson [13].

Given the data A, B, b, c and d, consider the following pair of dual programs:

$$\begin{aligned} &\inf cx + dy,\\ &Ax + By = b,\\ &x, y \geqq 0, \qquad x \text{ integer} \end{aligned} \tag{2.3}$$

and

$$
\begin{aligned}
&\max f(b), && \\
&f(A_i) \leqq c_i, && i=1,\cdots,n, \\
&f'(B_j) \leqq d_j, && j=1,\cdots,s, \\
&f(0)=0, && f\colon \mathrm{RHS}(A,B) \to \mathbb{R} \cup \{-\infty\} \text{ subadditive.}
\end{aligned}
\tag{2.4}
$$

Note that we require $f(0)=0$, while earlier subadditive duals required $f(0) \leqq 0$. Furthermore, let $G(v)$ be the function defined by

$$G(v) := \inf \{cx + dy \mid Ax + By = v,\, x,\, y \geqq 0,\, x \text{ integer}\},$$

where v is a feasible right-hand side for the primal program (2.3) ($v \in \mathrm{RHS}(A, B)$). We call G the value function of (2.3).

For parts of the following theorem we shall assume that either

(2.5) $s = 0$, i.e., there are no continuous variables,

or

(2.6) A and B have rational entries,

or

(2.7) there is a bound $K \in \mathbb{R}$ such that $(x, y) \in S(b)$ implies $\|x\| \leqq K$ for some norm $\|x\|$

holds. Often rationality of the data A and B or boundedness of the constraint set $S(b)$ in (2.3) are assumed, because in general the convex hull of $S(b)$ is not closed, hence not a polyhedron. This was noted by Meyer [17], [18] and Noltemeier [20]. Meyer [17] proved that the assumption of one of the three conditions (2.5)–(2.7) guarantees the validity of the following trichotomy:

(2.8) Either (2.3) is inconsistent or the objective of (2.3) is unbounded or (2.3) has an optimal solution.

Clearly, (2.5)–(2.7) are only sufficient conditions for the validity of (2.8).

THEOREM 2. *Consider the pair of dual programs* ((2.3) *and* (2.4)), *where the data A, B, b, c, and d are given.*

(i) *Assume one of* (2.5)–(2.7) *holds. Then* (2.3) *and* (2.4) *both have optimal solutions iff they both have feasible solutions and in this latter case,* (2.4) *has an optimal solution f with only finite values.*

(ii) *If (x, y) and f are optimal solutions to* (2.3) *and* (2.4) *resp., then $cx + dy = f(b) = G(b)$.*

(iii) *Let (x, y) be an optimal solution to* (2.3). *If $x_i > 0$ then $G(A_i) = G(b) - G(b - A_i)$ holds and analogously if $y_j > 0$, $G(\varepsilon B_j) = G(b - \varepsilon B_j)$ holds for all $0 \leqq \varepsilon \leqq y_j$.*

(iv) *Let $f(A)$ ($f'(B)$) denote a vector with components $f(A_i)$, $i = 1, \cdots, n$, ($f'(B_j)$, $j = 1, \cdots, s$). If (x, y) and f are optimal solutions to* (2.3) *and* (2.4), *then the complementary slackness condition holds, i.e., $x(c - f(A)) + y(d - f'(B)) = 0$. This condition remains true if f is replaced by G.*

Proof. We shall require:

LEMMA 1 (Jeroslow [10]). *For every feasible f of* (2.9),

$$\begin{aligned} &f(A_i)\leqq c_i, \qquad i=1,\cdots,n,\\ (2.9)\qquad &f'(B_j)\leqq d_j, \qquad j=1,\cdots,s,\\ &f\colon \mathrm{RHS}(A,B)\to \mathbb{R}\cup\{-\infty\} \textit{ subadditive,}\end{aligned}$$

and all $v\in$ RHS(A, B), *$f(v)\leqq G(v)$ holds. Moreover, G itself is feasible for* (2.9).

LEMMA 2 (see also Jeroslow [9], Johnson [13]). *If G is the value function for* (2.3) *we have*

(i) *G is subadditive on* RHS(A, B);

(ii) $G(x_iA_i)\leqq x_iG(A_i)$, $\forall x_i$ *integer* $x_i>0$;

(iii) $G(\delta B_j)\leqq \delta G'(B_j)$, $\forall\delta>0$;

(iv) *If $G'(b)$ is finite for $b\in$* RHS(A, B)
$G'(b)=\lim_{t\to 0_+} G(tb)/t$.

LEMMA 3. *If one of* (2.5)–(2.7) *holds at least for one $v\in$* RHS(A, B), *we have*

$$G(0)=0 \textit{ iff } G(v)>-\infty \textit{ for all } v\in \mathrm{RHS}(A,B).$$

Proof. If $S(v)$ is bounded for some v (case (2.7)), then $S(0)$ is bounded. Hence, for all $v\in$ RHS(A, B), (2.7) holds and $S(v)$ is the union of finitely many polyhedral sets (see [17]). If $G(v)=-\infty$ for some v, then (2.3) in unbounded over some polyhedral set with fixed x, which implies $G(0)=-\infty$.

In the other two cases, conv $(S(v))$ is polyhedral (in case (2.5) see [19], [20] and in case (2.6) see [17]) and thus the arguments of [17] (cf. Theorem 5) can be used to prove

$$(2.10)\qquad \inf\{cx+dy\,|\,Ax+By=0,\,x,\,y\geqq 0\}\neq-\infty \quad \text{iff} \quad G(0)\neq-\infty.$$

Hence if $v\in$ RHS (A, B) with $G(v)=-\infty$, we obtain $\inf\{cx+dy|Ax+By=v, x, y\geqq 0\}=-\infty$, which implies $\inf\{cx+dy|Ax+By=0, x, y\geqq 0\}=-\infty$, which in turn gives (using (2.10)) $G(0)=-\infty$. Conversely, if $G(v)>-\infty$ for all $v\in$ RHS(A, B), we obtain (since $(0, 0)\in S(0)$ and $G\not\equiv-\infty$) $0\leqq G(0)\leqq 0$ (cf. [9]) and thus $G(0)=0$.

Proof of Theorem 1. Assume (2.2) is consistent, i.e., there is an $(x, y)\in S(0)$ and $cx+dy<0$ which implies $k(x, y)\in S(0)$, $k\in\mathbb{N}$, hence $G(0)=-\infty$. If (2.1) is also consistent, we can use Lemma 1 to derive $0\leqq f(0)\leqq G(0)$ which contradicts $G(0)=-\infty$; thus both (2.1) and (2.2) can not be consistent. Assume (2.2) is inconsistent, i.e., $G(0)\geqq 0$. Using Lemma 1, we obtain G is feasible for (2.1).

Proof of Theorem 2. If (2.3) and (2.4) both have feasible solutions $(\tilde{x}, \tilde{y})$ and $\tilde{f}$, we can use Lemma 1 to derive (note that $(0, 0)\in S(0)$) $0=\tilde{f}(0)\leqq G(0)\leqq 0$, i.e., using Lemma 3 we obtain $G(b)>-\infty$. Assuming one of the conditions (2.5)–(2.7), the validity of (2.8) holds which proves the existence of an optimal solution (x, y) of (2.3), and Lemma 1 implies G is dual optimal and by definition of G: $G(b)\geqq f(b)\geqq G(b)=cx+dy$. Let $i\in\{1,\cdots,n\}$ be fixed and let $e_i\in\mathbb{R}^n$ denote the ith unit vector of $\mathbb{R}^n$, then $((x-ke_i), y)\in S(b-kA_i)$ for all $k\in\mathbb{Z}$ with $k\leqq x_i$. Therefore $G(b-kA_i)\leqq c(x-ke_i)+dy$ and $c_ik=G(b)-c(x-ke_i)-dy\leqq G(b)-G(b-kA_i)\leqq G(kA_i)\leqq kG(A_i)\leqq c_ik$, hence equality holds all over. If $x_i>0$ set $k=1$, which yields $G(A_i)=c_i$ and $G(b)-G(b-A_i)=G(A_i)$. On the other hand, if $G(A_i)<c_i$, x_i must be zero. Using Lemma 2 (iii), the same argument holds for the continuous variables. The argumentation goes through if G is replaced by f, which proves the theorem.

Note that for the complementary slackness result we do not need any assumption on the data A and B. It is exactly this technique that we use later on to prove necessary conditions for minimal inequalities.

3. Minimal inequalities without using the trichotomy. We say that an inequality $cx+dy \geqq h$ with real data c, d and h is valid for $S(b)$ if it is satisfied by all $(x, y) \in S(b)$. A valid inequality is called minimal if there is no other valid inequality $\tilde{c}x+\tilde{d}y \geqq \tilde{h}$ of $S(b)$ with $\tilde{c} \leqq c$, $\tilde{d} \leqq d$ and $\tilde{h} \geqq h$. Using the coefficients (c, d) of a given valid inequality as the objective coefficients in (2.3), we call G the value function of that valid inequality.

Clearly if $cx+dy \geqq h$ is a minimal inequality then $cx+dy \geqq G(b)$ is a valid inequality, hence $G(b) \leqq h$. Since for all positive ε there is an $(x, y) \in S(b)$ such that $G(b) \leqq h \leqq cx+dy \leqq G(b)+\varepsilon$ we obtain for $\varepsilon \to 0_+$,

$$G(b)=h. \tag{3.1}$$

Investigations of minimal inequalities have been done in various settings, (cf. [3], [7]–[14]). Recently Jeroslow [9] and Blair [2] gave a characterization which holds even for the general mixed integer problem assuming a bounded feasible set $S(b)$ [10] or rational data A and B [13] (see also [14] for related results).

In this section we prove that the characterization given in [10] can be stated without any assumption on the feasible set $S(b)$.

THEOREM 3. *If $cx+dy \geqq h$ is a minimal inequality for $S(b)$ and G is its value function, then*

$$G(b)=h, \tag{3.2}$$

$$G(A_i)=G(b)-G(b-A_i)=c_i, \qquad i=1,\cdots,n, \tag{3.3}$$

$$G'(B_j)=d_j=\lim_{\delta\to 0_+}(G(b)-G(b-\delta B_j))/\delta, \qquad j=1,\cdots,s. \tag{3.4}$$

Clearly as mentioned above, (3.2) is an easy consequence of the minimality.

LEMMA 4. *If $cx+dy \geqq h$ is a minimal inequality, then $S(b-x_iA_i)\emptyset$ for some integer $x_i \geqq 1$, $i=1,\cdots,n$ and $G(b-x_iA_i)>-\infty$, $G(x_iA_i)>-\infty$ for every $x_i \in \mathbb{N}$ such that $S(b-x_iA_i) \neq \emptyset$. Analogously $S(b-y_jB_j) \neq \emptyset$ for some $y_j>0$, $j=1,\cdots,s$, and $G(b-y_jB_j)>-\infty$, $G(y_jB_j)>-\infty$ for every $y_i>0$ such that $S(b-y_jB_i) \neq \emptyset$.*

Proof. If $x_i=0$ for every feasible $(x, y) \in S(b)$, $cx+dy \geqq h$ cannot be minimal because c_i can be lowered to, say, $\tilde{c}_i=c_i-1$. $G(b-x_iA_i)=-\infty$ or $G(x_iA_i)=-\infty$ imply $G(b)=-\infty$ which contradicts $G(b)=h>-\infty$; hence one part of Lemma 4 holds. The same argument carries over to the second part.

LEMMA 5. *If $cx+dy \geqq h$ is a minimal inequality and $S(b-x_i^0A_i) \neq \emptyset$ for some integer $x_i^0>0$, then $(G(b)-G(b-x_i^kA_i))/x_i^k$ is monotonically increasing for every monotonically decreasing sequence $(x_k^i \mid k \in \mathbb{N})$ with $x_i^0>x_i^k>0$ and integer. Similarly if $S(b-\delta^0B_j) \neq \emptyset$ for some $\delta^0>0$, then $(G(b)-G(b-\delta^kB_j))/\delta^k$ is monotonically increasing for every monotonically decreasing sequence $(\delta^k \mid k \in \mathbb{N})$ with $\delta^0>\delta^k>0$.*

Proof. Without loss of generality $x_i^0 \geqq 2$. Since $S(b-x_i^kA_i) \neq \emptyset$ for all integer $x_i^0>x_i^k>0$, we have by Lemma 4 $G(b-x_i^kA_i)>-\infty$ and $G(x_i^kA_i)>-\infty$. Consider the function $F(x_i):=G(b)-G(b-x_iA_i)-c_ix_i$. Using the subadditivity of G and Lemma 1, we obtain for integer x_i, $z_i>0$ such that $x_i+z_i \leqq x_i^0$:

$$\begin{aligned}
F(x_i+z_i) &= G(b)-G(b-(x_i+z_i)A_i)-c_i(x_i+z_i)\\
&\leqq G(b)-G(b-x_iA_i)+G(z_iA_i)-c_i(x_i+z_i) \quad \text{(subadditivity)}\\
&\leqq G(b)-G(b-x_iA_i)+c_iz_i-c_i(x_i+z_i) \quad \text{(Lemmas 1, 2)}\\
&\leqq G(b)-G(b-x_iA_i)-c_ix_i\\
&= F(x_i).
\end{aligned}$$

Since $x_i + z_i \geqq x_i$, this yields $F(x_i + z_i)/(x_i + z_i) \leqq F(x_i)/x_i$, which establishes one part of the lemma. But the other part can be proven exactly in the same way.

LEMMA 6. *If $cx + dy \geqq h$ is a minimal inequality, we have*:

$$\sup \{(G(b) - G(b - x_iA_i))/x_i | (x, y) \in S(b), x_i > 0\} = c_i, \tag{3.5}$$

$$\sup \{(G(b) - G(b - y_jB_j))/y_j | (x, y) \in S(b), y_j > 0\} = d_j. \tag{3.6}$$

Proof. Because the proof of (3.5) and (3.6) use exactly the same arguments let us only show (3.5). Let $(x, y) \in S(b)$ with $x_i > 0$, then $G(b) \leqq G(b - x_iA_i) + G(x_iA_i)$ (using subadditivity of G). Applying Lemma 2, we obtain $(G(b) - G(b - x_iA_i))/x_i \leqq x_iG(A_i)/x_i = G(A_i) \leqq c_i$; hence if $\tilde{c}_i$ denotes the supremum of the left hand side of (3.5), $\tilde{c}_i \leqq c_i$ must hold. Let $\tilde{c}_k := c_k$ for $k = 1, \cdots, n$ and $k \neq i$, then

$$\begin{aligned} \tilde{c}x + dy &= \tilde{c}_ix_i + c(x_1, \cdots, x_{i-1}, 0, x_{i+1}, \cdots, x_n) + dy \\ &\geqq \tilde{c}_ix_i + G(b - x_iA_i) \\ &\geqq G(b) - G(b - x_iA_i) + G(b - x_iA_i) \quad \text{(definition of } \tilde{c}_i\text{)} \\ &= G(b) = h \end{aligned}$$

holds for every $(x, y) \in S(b)$, i.e., $\tilde{c}x + dy \geqq h$ is still valid for $S(b)$. The minimality of $cx + dy \geqq h$ implies $\tilde{c} \geqq c$, hence $\tilde{c} = c$ and (3.5) is established.

Proof of Theorem 3. Lemma 5 implies that the supremum in (3.5) will be attained at some $\bar{x}_i > 0$. Thus

$$\begin{aligned} c_i\bar{x}_i &= G(b) - G(b - \bar{x}_iA_i) = G(b) - G(b - A_i - (\bar{x}_i - 1)A_i) \\ &\leqq G(b) - G(b - A_i) + (\bar{x}_i - 1)G(A_i) \quad \text{(subadditivity)} \\ &\leqq G(b) - G(b - A_i) + c_i(\bar{x}_i - 1) \quad \text{(Lemma 1)} \\ &\leqq c_i\bar{x}_i \quad \text{(subadditivity and Lemma 1),} \end{aligned}$$

i.e., equality holds everywhere and we obtain $c_i \doteq G(b) - G(b - A_i) \leqq G(A_i) \leqq c_i$ which proves (3.3). To prove (3.4) we use again Lemma 5 and Lemma 6 which prove

$$\begin{aligned} d_j &= \sup \{(G(b) - G(b - y_jB_j))/y_j | (x, y) \in S(b), y_j > 0\} \\ &= \lim_{\delta \to 0_+} (G(b) - G(b - \delta B_j))/\delta. \end{aligned}$$

Using Lemma 1 we obtain

$$\begin{aligned} d_j &= \lim_{\delta \to 0_+} (G(b) - G(b - \delta B_j))/\delta \\ &\leqq \lim_{\delta \to 0_+} G(\delta B_j)/\delta \\ &= G'(B_j) \leqq d_j, \end{aligned}$$

which proves the theorem.

Remark. The converse of Theorem 3 holds: if $cx + dy \geqq h$ is a valid inequality for $S(b)$, and G is its value function, then (3.2), (3.3) and (3.4) imply that $cx + dy \geqq h$ is a minimal inequality. The proof in [10] goes through without using the trichotomy or any more restrictive assumption.

Acknowledgment. The authors wish to thank Professor R. Jeroslow and Professor E. Johnson for reading a first draft of the manuscript and suggesting several improvements.

REFERENCES

[1] A. A. Araoz, *Polyhedral neopolarities*, Ph.D. dissertation, Univ. of Waterloo, Waterloo, Ontario, Canada, November, 1973.

[2] A. Bachem and R. V. Randow, *Integer theorems of Farkas lemma type*, Operations Research Verfahren Vol. 32, R. Henn et al., eds., Athenäum, Hain, Scriptor, Hanstein, Meisenheim, Federal Republic of Germany, 1979, pp. 19–28.

[3] C. E. Blair, *Minimal inequalities for mixed integer programs*, Discrete Math., 24, (1978), pp. 147–151.

[4] C. A. Burdet and E. L. Johnson, *A subadditive approach to the group problem of integer programming*, Math. Programming Stud. (1974), pp. 51–71.

[5] G. B. Dantzig, *Linear Programming and Extensions*, Princeton University Press, Princeton, NJ, 1963.

[6] J. Farkas, *Theorie der einfachen Ungleichungen*, J. Reine Angew. Math., 124 (1902), pp. 1–27.

[7] R. E. Gomory, *Some polyhedra related to combinatorial problems*, Linear Algebra Appl. 2 (1969), pp. 451–558.

[8] R. E. Gomory and E. L. Johnson, *Some continuous functions related to corner polyhedra: I and II*, Math. Programming, 3 (1972) pp. 23–85 and pp. 359–389.

[9] R. Jeroslow, *Cutting planes for relaxations of integer programs*, Management Science Research Rep. 347, Graduate School of Industrial Administration, Carnegie–Mellon University, Pittsburgh, PA 1974.

[10] ———, *Minimal inequalities*, Math. Programming, 17 (1979), pp. 1–15.

[11] ———, *Cutting-plane theory: Algebraic methods*, Discrete Math. 23 (1978), pp. 121–150.

[12] E. L. Johnson, *Cyclic groups, cutting planes, shortest paths*, Mathematical Programming, T.C. Hu and S. M. Robinson, eds. Academic Press, New York, 1973, pp. 185–211.

[13] ———, *The group problem for mixed integer programming*, Math. Programming Stud., 2 (1974), pp. 137–179.

[14] ———, *Faces of Polyhedra of Mixed Integer Programming Problems*, Proceedings of the Conference of Mathematical Programming and its Applications, Symposia Mathematica, Istituto Nazionale di Alto Matematica Roma; Academic Press, New York, 1976.

[15] R. Kannan and C. L. Monma, *On the Computational Complexity of Integer Programming Problems*, Optimization and Operations Research, R. Henn, B. Korte and W. Oettli, eds., Lecture Notes in Economics and Mathematical Systems 157, Springer, Heidelberg, 1978.

[16] L. Kronecker, *Näherungsweise ganzzahlige Auflösung linearer Gleichungen*, Leopold Kroneckers Werke Band III_1, K. Hensel ed., Teubner, Leipzig, 1899, pp. 47–110.

[17] R. R. Meyer, *On the existence of optimal solutions to integer and mixed-integer programming problems*, Math. Programming, 7 (1974), pp. 223–235.

[18] ———, *Integer and mixed-integer programming models: General properties*, J. Optimization Theory Appl., 16 (1975), pp. 191–206.

[19] R. R. Meyer and M. L. Wage, *On the polyhedrality of the convex hull of the feasible set of an integer program*, this Journal, 16 (1978), pp. 682–687.

[20] H. Noltemeier, *Sensitivitätsanalyse bei diskreten linearen Optimierungsproblemen*, Lecture Notes in Operations Research and Mathematical Systems 30, Springer, Berlin, 1970.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 4, July 1980

0363–0129/80/1804–0009 $01.00/0

EXTENSIONS OF HILDRETH'S ROW-ACTION METHOD FOR QUADRATIC PROGRAMMING*

ARNOLD LENT† AND YAIR CENSOR‡

Abstract. An extended version of *Hildreth's iterative quadratic programming algorithm* is presented, geometrically interpreted, and proved to produce a sequence of iterates that (i) converges to the solution, and (ii) has an important *intermediate optimality property*. This extended Hildreth algorithm is cast into a new form which more pronouncedly brings out its *primal-dual* nature. The application of the algorithm may be governed by an index sequence which is more general than a cyclic sequence, namely, by an *almost cyclic control*, and a sequence of *relaxation parameters* is incorporated without ruining convergence. The algorithm is a *row-action* method which is particularly suitable for handling *large* (or *huge*) and *sparse systems.*

1. Introduction. Linearly constrained quadratic optimization problems appear in various fields of applications and it is not rare to encounter large-scale problems or even huge-scale problems (i.e., problems in which the number of variables run into the range of 10^5 and more, with an even larger number of constraints). Usually, the matrix describing the constraints will be sparse, but all too often no special structure pattern is detectable in it. In such cases the use of *row-action methods* is strongly suggested. A row-action method is any iterative procedure which uses the rows of the matrix one at a time; see Censor [4], Censor and Herman [5].

As one example of an application field in which such a situation arises, we name the subject of image reconstruction from projections, which has had, during the last years, an overwhelming impact on diagnostic radiology through the introduction of the techniques of computerized transaxial tomography; see, e.g., Gordon and Herman [14], Herman and Lent [15].

Hildreth's quadratic programming procedure [19] is indeed a row-action method, and therefore, it deserves the attention of solvers of large and sparse quadratic programming problems. The capabilities of this algorithm were demonstrated in Herman and Lent [18] where results of a numerical experiment on a large (ca. 1,600 variables, 4,800 interval constraints) are presented.

In the present paper we cast Hildreth's algorithm into a "compact" form and introduce into it a sequence $\{r^{(k)}\}$ of *relaxation parameters* (typically, $0 < r^{(k)} < 2$). The option of using relaxation parameters has been shown to be a very important tool in practical implementations of other row-action methods (see the discussion in § 6 of Herman and Lent [15] and the experimental results of Herman et al. in [17]), and, undoubtedly, one would like to have the assurance that the convergence of the algorithm is not ruined by the introduction of the relaxation parameters.

Another feature introduced here into the Hildreth algorithm is the *almost cyclic control*, meaning that the rows of the matrix may be taken up, as the iterations proceed, in a manner which is less restrictive than cyclic control. The introduction of the almost cyclic control, though it causes some difficulties which are laboriously overcome in the

*Received by the editors August 23, 1978, and in revised form October 26, 1979. This research was supported in part by the National Institutes of Health under Grants HL4664, HL18968 and RR7, and by the National Cancer Institute under Contract CB-84235.

†Medical Image Processing Group, Department of Computer Science, State University of New York at Buffalo, Amherst, New York 14226 and Biodynamics Research Unit, Mayo Clinic, Rochester, Minnesota 55901. Now at Department of Quantitative Methods, College of Business, University of Illinois at Chicago Circle, Chicago, Illinois 60680. The work of this author was supported in part by a grant from the Waters Instrument Company, Rochester, Minnesota.

‡Medical Image Processing Group, Department of Computer Science, State University of New York at Buffalo, Amherst, New York 14226. Now at Department of Mathematics, University of Haifa, Haifa 31999, Israel.

proof of convergence, lays the foundations for the method of quadratic optimization over pairs of inequalities (i.e., interval constraints) developed in Herman and Lent [18], where almost cyclicality is essential. It also constitutes a contribution towards the question raised by D'Esopo in [9, p. 41], regarding generalizations of cyclic controls.

Thus, our main task here is to supply a proof of Hildreth's procedure with relaxation parameters under almost cyclic control. The proof presented here does not rely on any full-row-rank assumptions. In this way the difficulties, discovered by D'Esopo and reported in the erratum [19], with Hildreth's original proof are completely circumvented. Although D'Esopo recognized the difficulty, it has not been taken care of in a satisfactory manner in his work on Hildreth's procedure [9]. There exists yet another proof of the convergence of Hildreth's algorithm given by Bregman [3]; but being completely differently structured, Bregman's proof covers only the cyclic case and does *not* lend itself to a modification which will allow for the incorporation of relaxation parameters. In our proof, which differs from previously given proofs, we also pick up along the line an interesting intermediate optimality property of the iterates of the algorithm, which has important consequences due to the fact that, in practice, the algorithm is always stopped after a finite number of iterations, and a particular iterate is taken as the approximate solution to the problem.

As far as the literature goes, it is worthwhile to mention that only a handful of textbooks on optimization present the Hildreth algorithms at all (see Hadley [13], Kunzi and Krelle [21], Luenberger [23]). In Gorenflo and Kovetz [11], Herman and Lent [18], and Wendler [30] applications of Hildreth's algorithms were made, while rate-of-convergence results were reported in Oettli [27]. We would like to believe that this paper, along with the development in [18] and the presentations in [4] and [5], sheds more light on the benefits which Hildreth's procedure promises for large and sparse systems applications.

In § 2, almost cyclicality and relaxation parameters are introduced and the extended Hildreth algorithm is presented and discussed along with a geometric interpretation. In § 3, our convergence theorem is formulated, and duality theory is invoked to yield some preliminary results concerning the dual sequence. Also, the convergence of the differences of iterates to zero, a cornerstone in the convergence proof, is established.

In § 4, the convergence of a sequence of constraint sets is proved, and then the proof of convergence of the sequence of iterates produced by the extended Hildreth algorithm is concluded. In § 5, an intermediate optimality property is illuminated, and a concluding discussion is presented in § 6.

2. Almost cyclicality, relaxation parameters, and the extended Hildreth algorithm. Hildreth's quadratic programming procedure [19] is an iterative method for finding an approximate solution of the problem:

$$\begin{aligned} &\min \tfrac{1}{2}\langle By, y\rangle + \langle y, d\rangle \\ &\text{such that } Gy \leqq h, \end{aligned} \tag{2.1}$$

where B is a positive-definite $n \times n$ matrix, G is an $m \times n$ matrix, $y \in R^n$, $d \in R^n$ and $h \in R^m$; $\langle \cdot, \cdot \rangle$ stands for the inner product in the n-dimensional Euclidean space R^n.

Using the Choleski decomposition with $B = D^T D$ and letting $y = D^{-1}x - B^{-1}d$, the problem (2.1) is transformed into the following:

Standard Problem

$$\begin{aligned} &\min \tfrac{1}{2}\|x\|^2 \\ &\text{such that } Ax \leqq b, \end{aligned} \tag{2.2}$$

where $A = GD^{-1}$ and $b = h + GB^{-1}d$; $\|\cdot\|$ stands for the Euclidean norm in R^n.

The application of Hildreth's algorithm to a given quadratic optimization problem is governed by a sequence of indices which specifies the row of the matrix to be taken up at a given iterative step. This sequence will be called the *control* of the algorithm. The following definition was introduced in Lent [22].

DEFINITION 2.3. Let $I = \{1, 2, \cdots, m\}$ be a finite set. A sequence $\{i_k\}_{k=0}^{\infty}$ is *almost cyclic on I* if

(i) $i_k \in I$ for all $k \geqq 0$, and

(ii) there exists an integer C, called an *almost cyclicality constant*, such that for all $k \geqq 0$, $I \subseteq \{i_{k+1}, \cdots, i_{k+C}\}$.

An almost cyclic sequence on $\{1, 2, \cdots, \mathrm{m}\}$ is called *cyclic* if $C = m$.

Almost cyclic controls are less restrictive than the cyclic control, and they add another important option as to how the application of the method to a particular problem will be carried out. In Herman and Lent [18], almost cyclicality is essential. Various other almost cyclic controls can be employed as, for example, the Aitken's double sweep method see, e.g., Luenberger [24, p. 158].

We now give the extended version of Hildreth's algorithm for solving the standard problem (2.2). Remarks concerning the application of the algorithm to the general quadratic programming problem (2.1) are given in the discussion presented in § 6, below.

We will assume throughout that $a_i \neq 0$ for all $i \in I = \{1, 2, \cdots, m\}$, where $a_i \in R^n$ forms the ith row of the matrix A. Denoting $S \equiv \{x | Ax \leqq b\}$, the feasible set of (2.2), it will be assumed throughout that $S \neq \emptyset$.

ALGORITHM 2.4 (Extended Hildreth algorithm).

(2.4.1) *Initialization*: $z^{(0)} \in R_+^m$ arbitrary (R_+^m stands for the nonnegative orthant of R^m) and $x^{(0)} \equiv -A^T z^{(0)}$.

(2.4.2) *Typical step*: $x^{(k+1)} = x^{(k)} + c^{(k)} a_{i_k}$, $z^{(k+1)} = z^{(k)} - c^{(k)} e_{i_k}$ with

$$c^{(k)} \equiv \min\left(z_{i_k}^{(k)}, r^{(k)} \frac{b_{i_k} - \langle a_{i_k}, x^{(k)} \rangle}{\|a_{i_k}\|^2}\right),$$

where e_i is the vector with 1 in the ith place and zeros elsewhere, and $\{r^{(k)}\}$ is the sequence of *relaxation parameters*, all of which are assumed to be positive.

(2.4.3) *Control*: The sequence $\{i_k\}_{k=0}^{\infty}$ is almost cyclic on $I = \{1, 2, \cdots, m\}$.

This compact presentation of the extended Hildreth algorithm paves the way for a nice geometrical interpretation which was reported in Herman and Lent [16]. First, let us show that for the special case of a cyclic control (i.e., $i_k = k \pmod{m} + 1$) and with unity relaxation (i.e., $r^{(k)} = 1$ for all k) the extended Hildreth algorithm (Algorithm 2.4), as formulated above, does coincide with the original (and traditional) representation of Hildreth's algorithm; see, e.g., Hildreth [19], Luenberger [23], Oettli [27], Wendler [30].

To do this, observe that $a_i = A^T e_i$ and that $x^{(k)} = -A^T z^{(k)}$ for all k. This last fact is easily seen to be true because $x^{(0)} = -A^T z^{(0)}$ from (2.4.1) and by induction, $x^{(k+1)} = x^{(k)} + c^{(k)} a_{i_k} = -A^T z^{(k)} + c^{(k)} A^T e_{i_k} = -A^T z^{(k+1)}$. Given k, we make here the temporary abbreviation $i = i_k$, so that

$$\begin{aligned} z^{(k+1)} &= z^{(k)} - c^{(k)} e_i \\ &= z^{(k)} - \left[\min\left(z_i^{(k)}, \frac{b_i - \langle a_i, x^{(k)}\rangle}{\|a_i\|^2}\right)\right] e_i \qquad (2.5) \\ &= z^{(k)} + \left[\max\left(0, z_i^{(k)} - \frac{b_i - \langle a_i, x^{(k)}\rangle}{\|a_i\|^2}\right) - z_i^{(k)}\right] e_i. \end{aligned}$$

Now,

$$\begin{aligned} -\langle a_i, x^{(k)}\rangle &= \langle a_i, A^T z^{(k)}\rangle \\ &= \left\langle a_i, \sum_{j=1}^{m} a_j z_j^{(k)}\right\rangle = \sum_{j=1}^{m} \langle a_i, a_j\rangle z_j^{(k)} \qquad (2.6) \\ &= \sum_{j=1}^{m} d_{ij} z_j^{(k)}, \end{aligned}$$

where $d_{ij} = \langle a_i, a_j\rangle$ is a typical element of the $m \times m$ matrix AA^T, and $d_{ii} = \langle a_i, a_i\rangle \neq 0$ for all $i \in I$.

Substituting this in (2.5) and writing it componentwise we get

$$(2.7) \qquad \begin{aligned} z_j^{(k+1)} &= z_j^{(k)} \quad \text{if } j \neq i, \\ z_j^{(k+1)} &= \max\left(0, z_i^{(k)} - \frac{1}{d_{ii}}\left(b_i + \sum_{j=1}^{m} d_{ij} z_j^{(k)}\right)\right) \quad \text{if } j = i. \end{aligned}$$

It is in the form (2.7) that Hildreth's procedure has been presented in the literature with the initialization $z^{(0)} \geqq 0$ and the update rule $x^{(k)} = -A^T z^{(k)}$, and with $z^{(k)}$, the vector of *dual variables*, "leading" the execution of the algorithm at each iteration.

The form in which the extended Hildreth Algorithm 2.4 is cast here emphasizes its primal-dual nature (see Luenberger [23]) and enables us to give the following geometric interpretation. Assume for the time being that $r^{(k)} = 1$ for all k. Observe that, because of the initialization (2.4.1) and by induction, $z^{(k)} \in R_+^m$ for all k. Now consider the nature of the kth step, which produces $x^{(k+1)}$ from $x^{(k)}$. Again, abbreviate $i = i_k$ for the control index taken at this step and let $H_i \equiv \{x | \langle a_i, x\rangle \leqq b_i\}$ be the half-space associated with the ith constraint of the standard problem. We distinguish between two possibilities.

If $x^{(k)} \notin H_i$, i.e., $x^{(k)}$ violates the ith constraint, then $b_i - \langle a_i x^{(k)}\rangle < 0$ and $c^{(k)} = (b_i - \langle a_i, x^{(k)}\rangle)/\|a_i\|^2$. This implies that $x^{(k+1)}$ is the *orthogonal projection* of $x^{(k)}$ onto the bounding hyperplane ∂H_i of the half space H_i (see Fig. 1).

If $x^{(k)} \in H_i$, then again an orthogonal move towards ∂H_i is made. The move is given by $z_i^{(k)} a_i$ unless this would result in crossing the hyperlane ∂H_i, in which case we "stop" at the hyperplane (see Fig. 2). Observe that if $x^{(k)} \in \partial H_i$ then $x^{(k+1)} = x^{(k)}$

In Hildreth's algorithm, $x^{(k+1)} \in H_i$ always and the absolute value of $z_i^{(k+1)}$ has been made as small as possible—consistent with that condition.

The Hildreth algorithm is specifically designed for approximating the minimum-norm (shortest, least-distance) element of a polyhedral convex set when the set is described in terms of large numbers of intersecting hyperplanes. *Different* descriptions of the feasible set suggest the use of different algorithms. When the set is defined as the *convex* (respectively, *conical*) *hull* of a point set, then the method of Wolfe [32] (respectively, Wilhelmsen [31]) might be used, provided that storage for the necessary tableaux is available. If the *vertices* of the convex set are known, then the method of Bazaraa, Good and Rardin [2] is applicable.

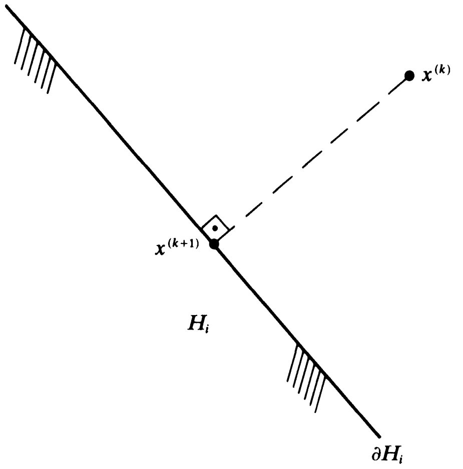


FIG. 1

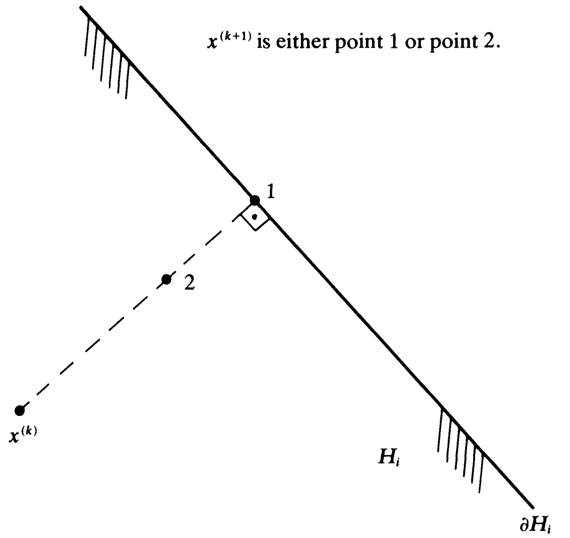


FIG. 2

For the minimization of more complicated quadratic functions (e.g., least squares) over the canonical set R^n_+, the (iterative) S.O.R.-like method of Cryer [6], or its extension of Mangasarian [25], is a possibility. These algorithms update the iterates on a component-by-component basis, and do not require any auxiliary tableaux.

It is not simple, then, to compare the performance of the Hildreth algorithm with that of other large-scale programming algorithms. The three classes of algorithms mentioned correspond to three different classes of problems, and there seems to be no real competition between them.

3. A convergence theorem for the extended Hildreth algorithm, the dual problem and convergence to zero of the differences. The following theorem for the extended Hildreth algorithm will be proved.

THEOREM 3.1. *Assumptions*: (1) $S=\{x|Ax\leqq b\}\neq\emptyset, a_i\neq 0$ *for all* $i\in I=\{1,2,\cdots,m\}$.

(2) $\{i_k\}_{k=0}^{\infty}$ *is an almost cyclic sequence on I, and*

(3) *there exist* $\varepsilon_1, \varepsilon_2>0$ *such that* $\varepsilon_1\leqq r^{(k)}\leqq 2-\varepsilon_2$ *for all k.*

Conclusion: *The sequence* $\{x^{(k)}\}$ *produced by the extended Hildreth algorithm* (2.4) *converges to the solution of* (2.2).

In the sequal, the necessary theory is developed and it is not until (4.15), in § 4, that the proof of this theorem is completed.

The dual problem to the standard problem (2.2) is

$$\begin{aligned}&\max \Phi(z)\\ &\text{such that } z\in R_+^m,\end{aligned}\tag{3.2}$$

where $\Phi(z)\equiv\min_{x\in R^n}L(x,z)$, with the Lagrangian $L(x,z)=\frac{1}{2}\sum_{j=1}^n x_j^2+\sum_{i=1}^m z_i(\sum_{j=1}^n a_{ij}x_j-b_i)$, and therefore

$$\Phi(z)=-\tfrac{1}{2}\|A^Tz\|^2-\langle b,z\rangle.\tag{3.3}$$

The standard problem (2.2) and its dual (3.2) are related by the duality theorem (see, e.g., [24])

$$\min_{x\in S}\tfrac{1}{2}\|x\|^2=\max_{z\in R_+^m}\Phi(z).\tag{3.4}$$

Next, we show that the extended Hildreth algorithm (2.4) defines a feasible ascent method in the dual variables, i.e., the sequence of dual vectors $\{z^{(k)}\}$ produced by (2.4) is feasible for (3.2) and the values of $\Phi(z^{(k)})$ increase monotonically and converge. From this, the convergence to zero of the differences of the iterates (primal and dual) follows, as is shown by

LEMMA 3.5. *If* $0<r^{(k)}\leqq 2-\varepsilon$, $\varepsilon>0$, *then for the sequences* $\{z^{(k)}\}$, $\{x^{(k)}\}$ *and* $\{c^{(k)}\}$ *produced by* (2.4), *we have*

(a) $z^{(k)}\in R_+^m$ *for all k,*

(b) $\Phi(z^{(k+1)})\geqq\Phi(z^{(k)})$,

(c) $\lim_{k\to\infty}[\Phi(z^{(k+1)})-\Phi(z^{(k)})]=0$,

(d) $c^{(k)}\xrightarrow[k\to\infty]{}0$, *and*

(e) $(x^{(k+1)}-x^{(k)})\xrightarrow[k\to\infty]{}0, (z^{(k+1)}-z^{(k)})\xrightarrow[k\to\infty]{}0.$

Proof. Again, abbreviate $i=i_k$.

(a) By induction. $z^{(0)}\in R_+^m$ from (2.4.1), $c^{(k)}\leqq z_i^{(k)}$ from (2.4.2), so that $z_i^{(k+1)}=z_i^{(k)}-c^{(k)}\geqq 0$ and all other components of $z^{(k+1)}$ remain unchanged.

(b) From (3.3) and from $x^{(k)}=-A^Tz^{(k)}$ for all k, we get

$$\Phi(z^{(k+1)})-\Phi(z^{(k)})=\|a_i\|^2c^{(k)}\left[\frac{b_i-\langle a_i,x^{(k)}\rangle}{\|a_i\|^2}-\frac{c^{(k)}}{2}\right].\tag{3.6}$$

To bound the right-hand side, we go back to (2.4.2). Note that

$$c^{(k)}\leqq r^{(k)}\frac{b_i-\langle a_i,x^{(k)}\rangle}{\|a_i\|^2},\tag{3.7}$$

where strict inequality can hold only if $c^{(k)}=z_i^{(k)}$, which is nonnegative because of (a) above. Hence,

$$\Phi(z^{(k+1)})-\Phi(z^{(k)})\geqq\|a_i\|^2[c^{(k)}]^2\left(\frac{1}{r^{(k)}}-\frac{1}{2}\right)\geqq 0 \tag{3.8}$$

because $0<r^{(k)}<2$.

(c) From (b), $\{\Phi(z^{(k)})\}$ is monotonically increasing, and (3.4) shows that it is bounded from above, so the conclusion follows.

(d) Let $\|a_i\|^2\geqq a>0$ for all i. Then, from (3.8) it follows that

$$\Phi(z^{(k+1)})-\Phi(z^{(k)})\geqq\tfrac{1}{4}a\varepsilon[c^{(k)}]^2$$

so that (c) implies $c^{(k)}\to 0$.

(e) This is immediate from (2.4.2) and (d). The interested reader may wish to consult Daniel's book [7, §§ 4.2, 6.2 and 6.3] for a general discussion of the significance of convergence to zero of the differences sequence of a sequence of iterates. □

Remark 3.9. Conditions on the relaxation sequence $\{r^{(k)}\}$, similar to those appearing here, are abundant in various iterative methods; see, e.g., Gubin et al. [12], Polyak [28].

4. A convergent sequence of perturbed constraint sets and the convergence of Hildreth's extended algorithm. Here we introduce a sequence of slack vectors $\{q^{(k)}\}$ which serve to construct a sequence of perturbed constaint sets $S^{(k)}$. These sets converge (see, e.g., Valentine [29, p. 39] for a definition of convergence of a sequence of sets) to the constraint set of (2.2). The intermediate optimality property of the iterates produced by (2.4) is established with respect to the sequence $\{S^{(k)}\}$.

DEFINITION 4.1. $q^{(0)}=0$, and

$$q_j^{(k+1)}\equiv\begin{cases} q_j^{(k)} & \text{if } j\neq i_k,\\ -\dfrac{c^{(k)}\|a_i\|^2}{r^{(k)}}+b_i-\langle a_i,x^{(k)}\rangle & \text{if } j=i_k\equiv i,\end{cases}$$

where i_k is the constraint index taken up at the kth step, abbreviated by i.

Next, define the vectors $b^{(k)}$ by

$$b^{(k)}\equiv q^{(k)}+Ax^{(k)}, \tag{4.2}$$

and use them to construct a sequence of *perturbed constraint sets* $S^{(k)}$ by

$$S^{(k)}\equiv\{x\,|\,Ax\leqq b^{(k)}\}. \tag{4.3}$$

Note that by construction, $S^{(k)}\neq\varnothing$ for all k. (The nonnegativity of $q^{(k)}$ follows by induction using (2.4.2).) The next result proves the convergence, in some sense, of the perturbed constraint sets.

LEMMA 4.4. *Let $\{i_k\}_{k=0}^{\infty}$ be almost cyclic on $I=\{1,2,\cdots,m\}$ and assume the existence of $\varepsilon_1,\varepsilon_2>0$ such that $\varepsilon_1\leqq r^{(k)}\leqq 2-\varepsilon_2$ for all k. Then, $b^{(k)}\to b$ as $k\to\infty$.*

Proof. We shall show the convergence componentwise. Let $t\in I$ be fixed. Taking any given k, denote by $l\equiv l(k)$ the most recent iteration, $l\leqq k$, with the property $i_l=t$. Because of the almost cyclicality of $\{i_k\}_{k=0}^{\infty}$ we know that $k-C\leqq l$, where C is an almost cyclicality constant. (We have assumed implicitly that $k>C$.)

From (4.1) we learn that

$$q_t^{(k)}=q_t^{(k-1)}=\cdots=q_t^{(l+1)},$$

so that substituting into (4.2) we get

$$\begin{aligned} b_t^{(k)} &= q_t^{(k)} + \langle a_t, x^{(k)}\rangle \\ &= q_t^{(l+1)} + \langle a_t, x^{(k)}\rangle \\ &= -\frac{c^{(l)}\|a_t\|^2}{r^{(l)}} + b_t + \langle a_t, x^{(k)} - x^{(l)}\rangle. \end{aligned} \tag{4.5}$$

The first summand on the right-hand side of (4.5) tends to zero as $k \to \infty$ because l also goes to infinity and Lemma 3.5 applies. The inner product can be expanded as

$$\langle a_t, x^{(k)} - x^{(l)}\rangle = \sum_{j=l}^{k-1} \langle a_t, x^{(j+1)} - x^{(j)}\rangle$$

with *at most C* terms in the sum, therefore, and because of the convergence of the differences $x^{(k+1)} - x^{(k)}$ to zero (Lemma 3.5(e)), this inner product tends to vanish with $k \to \infty$, and we are left with $b_t^{(k)} \to b_t$ as $k \to \infty$. □

We conclude the proof of the convergence of $\{x^{(k)}\}$ by appealing to some theorems of convex analysis. We will need the following

THEOREM 4.6. *Let* $T = \{y|Ay \leqq d\}$ *and* $T' = \{y'|Ay' \leqq d'\}$ *and assume* T' *is nonempty. Then, there exists a constant* α, *which depends only on* A, *such that for any* $y \in T$ *there exists a* $y' \in T'$ *with the property*

$$\|y - y'\| \leqq \alpha \|(d - d')^+\|, \tag{4.7}$$

where the upper plus notation means, for any vector v, $(v^+)_i \equiv \max(0, v_i)$.

This is Daniel's [8] statement of Hoffman's [20] theorem.

We adopt some additional notations: x^* will denote the point in S with minimum norm, $\hat{x}^{(k)}$ will denote the point in S closest to $x^{(k)}$, and $\hat{x}_*^{(k)}$ will denote the point in $S^{(k)}$ (see (4.3)) closest to x^*.

THEOREM 4.8. *Under the assumptions of Theorem* 3.1 *the following hold*:

(a) $\|x^{(k)}\| \to \|x^*\|$ *as* $k \to \infty$,

(b) $\|\hat{x}^{(k)} - x^{(k)}\| \to 0$ *as* $k \to \infty$, *and*

(c) $\|\hat{x}^{(k)}\| \to \|x^*\|$ *as* $k \to \infty$.

Proof. (a) Write

$$\|x^*\| \leqq \|\hat{x}^{(k)}\| \leqq \|\hat{x}^{(k)} - x^{(k)}\| + \|x^{(k)}\|. \tag{4.9}$$

Applying Theorem 4.6 with $S^{(k)}$ as T, S as T', and $x^{(k)}$ as y, we have

$$\|x^{(k)} - \hat{x}^{(k)}\| \leqq \alpha_1 \|(b^{(k)} - b)^+\|; \tag{4.10}$$

and because of Lemma 4.4 there exists, for any positive ε, a $K_1 = K_1(\varepsilon)$ such that

$$\|x^*\| \leqq \varepsilon + \|x^{(k)}\|, \qquad k \geqq K_1(\varepsilon). \tag{4.11}$$

To get an inequality in the opposite direction, write

$$\|x^{(k)}\| \leqq \|x_*^{(k)}\| \leqq \|x_*^{(k)} - x^*\| + \|x^*\|, \tag{4.12}$$

where the leftmost inequality in (4.12) is justified by Theorem 5.1, which describes the intermediate optimality property of the iterates. Now, apply again Theorem 4.6, this time with S as T, $S^{(k)}$ as T' and x^* as y, to get

$$\|x_*^k - x^*\| \leqq \alpha_2 \|(b - b^{(k)})^+\|. \tag{4.13}$$

Once more, the right-hand side can be made less than $\varepsilon > 0$ because of Lemma 4.4, and so (4.12) becomes,

$$\|x^{(k)}\| \leqq \varepsilon + \|x^*\|, \qquad k \geqq K_2(\varepsilon), \tag{4.14}$$

and the required result follows from (4.11) and (4.14).

(b) Follows now from (4.10) and Lemma 4.4.

(c) Follows from (4.9) and from (a) and (b). □

Proof of Theorem 3.1. We show here that $\lim_{k\to\infty} x^{(k)} = x^*$. Write

$$\|x^* - x^{(k)}\| \leqq \|x^* - \hat{x}^{(k)}\| + \|x^{(k)} - \hat{x}^{(k)}\|. \tag{4.15}$$

The second term on the right-hand side is taken care of by (b) of Theorem 4.8. Concerning the first summand, (c) of Theorem 4.8 ensures that, for the sequence $\{\hat{x}^{(k)}\} \subset S$,

$$\|\hat{x}^{(k)}\| \to \|x^*\| \quad \text{as } k \to \infty.$$

Since S is a closed and convex set and x^* is the *only* point in S which has minimum norm, it follows readily that $\hat{x}^{(k)} \to x^*$ as $k \to \infty$, thereby completing the proof. □

5. An intermediate optimality property. Here we use an argument due to Everett [10] to show that the iterates $x^{(k)}$ produced by applying Algorithm 2.4 to the standard problem (2.2) are optimal in the sense that each $x^{(k)}$ minimizes the objective function $f(x) = \frac{1}{2}\|x\|^2$ over the perturbed constraint set $S^{(k)}$.

Convergence of the sequence of iterates, or even rate-of-convergence results do not tell us too much about the nature of any particular iterate, and it is well-known to practitioners that sometimes results which are acceptable in terms of the real-world problem can be obtained by picking an iterate from a sequence generated by some algorithm whose very convergence is in doubt. Therefore, we consider the intermediate optimality property, giving information on the nature of the iterates $\{x^{(k)}\}$ individually, to be of some interest to the user of the algorithm who stops the iterations at some instant and takes a certain iterate as the approximate solution to the problem. (By the real-world problem we mean a problem that is modeled by the optimization problem, for example, the problem of image reconstruction from projections—see Herman and Lent [15]).

THEOREM 5.1. *$x^{(k)}$ produced by the extended Hildreth algorithm* (2.4) *minimizes $f(x)$ over the perturbed constraint set $S^{(k)}$.*

Proof. $x^{(k)}$ optimizes the Lagrangian $L(x, z^{(k)})$ (see § 3), therefore,

$$f(x^{(k)}) + \langle Ax^{(k)} - b, z^{(k)}\rangle \leqq f(x) + \langle Ax - b, z^{(k)}\rangle \quad \text{for all } x \in R^n, \tag{5.2}$$

that is,

$$f(x^{(k)}) \leqq f(x) + \langle Ax, z^{(k)}\rangle - \langle Ax^{(k)}, z^{(k)}\rangle.$$

It is not difficult to see that, for every $k = 0, 1, 2 \cdots$ and for every $j \in I = \{1, 2, \cdots, m\}$,

$$q_j^{(k)} \cdot z_j^{(k)} = 0, \tag{5.3}$$

where $q^{(k)}$ are the slack vectors of Definition 4.1 and $z^{(k)}$ are the dual vectors. (Using induction on k with (2.4.2) and Definition 4.1, it follows from the definition of $c^{(k)}$ that $q_j^{(k+1)} \cdot z_j^{(k+1)} = 0$).

This enables us to write

$$\begin{aligned} f(x^{(k)}) &\leqq f(x) + \langle Ax, z^{(k)}\rangle - \langle Ax^{(k)} + q^{(k)}, z^{(k)}\rangle \\ &= f(x) + \langle Ax - b^{(k)}, z^{(k)}\rangle \quad \text{for all } x \in R^n. \end{aligned} \tag{5.4}$$

For $x \in S^{(k)}$, we have $Ax - b^{(k)} \leqq 0$, whereas $z^{(k)} \in R_+^m$ for all k (Lemma 3.5). Therefore,

$$f(x^{(k)}) \leqq f(x) \quad \text{for all } x \in S^{(k)}. \qquad \square$$

6. Conclusion. Hildreth published his algorithm more than twenty years ago [19], but it was not until very recently that it was applied to a highly significant, huge-scale, sparse problem [18]. The essential simplicity of the extended Hildreth algorithm presented here makes it an easily programmable and low storage demanding method. Assuming row generation capability, core storage need be provided only for the x vector, while the dual vector z and the data b can be stored on disc. The extension to almost cyclic control and the introduction of a relaxation sequence, along with its row-action nature, make this algorithm particularly attractive for large and sparse quadratic programming problems. A convergence result for the application of the extended Hildreth algorithm to the general problem (2.1) follows readily from Theorem 3.1, but this is not to say that the algorithm is recommended for use on general quadratic problems. The presence of the matrix B^{-1} in the transformation from (2.1) to (2.2) obviously limits the applicability of the method. Even if the problem involves a simple, known B^{-1} it is unlikely that sparseness will be preserved through the Cholesky factorization.

The Cholesky factorization is, therefore, used here only as a theoretical tool and it is only for problems of the form (2.2) with large and sparse A, or problems that can be reduced to this form, that the extended Hildreth algorithm presented here is recommended.

Acknowledgments. The constant encouragement of Professor Gabor Herman is gratefully acknowledged. A preliminary report on the results of this paper can be found in [22a]. The manuscript was efficiently typed by Ms. Barbara S. Dane.

REFERENCES

[1] S. Agmon, *The relaxation method for linear inequalities*, Canad. J. Math., 6 (1954), pp. 382–392.

[2] M. S. Basaraa, J. J. Goode and R. L. Rardin, *An algorithm for finding the shortest element of a polyhedral set with application to Lagrangian duality*, J. Math. Anal. Appl., 65 (1978), pp. 278–288.

[3] L. M. Bregman, *The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming*, U.S.S.R. Computational, Math and Math. Phys., 7 (1967), pp. 200–217.

[4] Y. Censor, *Row-action methods for huge and sparse systems and their applications*, Tech. Rep. MIPG 28, Medical Image Processing Group, Dept. of Computer Science, State Univ. of New York, Amherst, NY, 1979; SIAM Rev., to appear.

[5] Y. Censor and G. T. Herman, *Row-generation methods for feasibility and optimization problems involving sparse matrices and their applications*, Sparse Matrix Proceedings 1978, I. S. Duff and G. W. Stewart, eds., Society for Industrial and Applied Mathematics, Philadelphia, 1979, pp. 197–219.

[6] C. W. Cryer, *The solution of a quadratic programming problem using systematic overrelaxation*, this Journal, 9 (1971), pp. 385–392.

[7] J. W. Daniel, *The approximate minimization of functionals*, Prentice-Hall, Englewood Cliffs, NJ, 1971.

[8] ———, *On perturbations of systems of linear inequalities*, SIAM J. Numer. Anal., 10 (1973), pp. 229–307.

[9] D. A. D'Esopo *A convex programming procedure*, Naval Res. Logist. Quart., 6 (1959), pp. 33–42.

[10] H. Everett, III, *Generalized Lagrange multiplier method for solving problems of optimum allocations of resources*, Operations Res., 11 (1963), pp. 399–417.

[11] R. Gorenflo and Y. Kovetz, *Solution of an Abel-type integral equation in the presence of noise by quadratic programming*, Numer. Math., 8 (1966), pp. 392–406.

[12] L. G. GUBIN, B. T. POLYAK AND E. V. RAIK, *The method of projections for finding the common point of convex sets*, U.S.S.R. Computational, Math. and Math. Phys., 7 (1967), pp. 1–24.

[13] G. HADLEY, *Nonlinear and Dynamic Programming*, Addison–Wesley, Reading, MA, 1964.

[14] R. GORDON AND G. T. HERMAN, *Three-dimensional reconstruction from projections: A review of algorithms*, Internat. Rev. Cytology, 38 (1974), pp. 111–151.

[15] G. T. HERMAN AND A. LENT, *Iterative reconstruction algorithms*, Computers in Biology and Medicine, 6 (1976), pp. 273–294.

[16] ———, *A relaxation method with application in diagnostic radiology*, Proc. of the IX International Symposium on Mathematical Programming, Budapest, Hungary, August 23–27, 1976.

[17] G. T. HERMAN, A. LENT AND P. H. LUTZ, *Relaxation methods for image reconstruction*, Comm. ACM, 21 (1978), pp. 152–158.

[18] G. T. HERMAN AND A. LENT, *A family of iterative quadratic optimization algorithms for pairs of inequalities, with application in diagnostic radiology*, Math. Programming Stud., 9 (1978), pp. 15–29.

[19] C. HILDRETH, *A quadratic programming procedure*, Naval Res. Logist. Quart., 4 (1957), pp. 79–85; Erratum, Ibid., p. 361.

[20] A. J. HOFFMAN, *On approximate solutions of systems of linear inequalities*, J. Res. Nat. Bur. Standards, 49 (1952), pp. 263–265.

[21] H. P. KUNZI AND W. KRELLE, *Nichtlineare Programmierung*, Springer, Berlin, 1962.

[22] A. LENT, *Maximum entropy and multiplicative ART*, Image Analysis and Evaluation, SPSE Conference Proceedings, R. Shaw, ed., July 1976, Toronto, Canada, pp. 249–257.

[22a] ———, *On the convergence of Hildreth's quadratic programming algorithm*, Tech. Rep. 122, Dept. of Computer Science, State University of New York, Buffalo, NY, 1977.

[23] D. G. LUENBERGER, *Optimization by Vector Space Methods*, Wiley, New York, 1969.

[24] ———, *Introduction to Linear and Nonlinear Programming*, Addison-Wesley, Reading, MA, 1973.

[25] O. L. MANGASARIAN, *Solution of symmetric linear complementary problems by iterative methods*, J. Optimization Theory Appl., 22 (1977), pp. 465–485.

[26] T. S. MOTZKIN AND I. J. SCHOENBERG, *The relaxation method for linear inequalities*, Canad. J. Math., 6 (1954), pp. 393–404.

[27] W. OETTLI, *Einzelschrittverfahren zur Losung konvexer und dual-konvexer Minimierungsprobleme*, Z. Angew. Math. Mech., 54 (1974), pp. 343–351.

[28] B. T. POLYAK, *Minimization of unsmooth functionals*, U.S.S.R. Computational Math. Math. Phys., 9 (1969), pp. 14–29.

[29] F. A. VALENTINE, *Convex Sets*, McGraw-Hill, New York, 1964.

[30] K. WENDLER, *Die ε-Storung linearer und quadratischer Optimierungsaufgaben und ihre Anwendung auf das Hildreth-Verfahren*, Unternehmensforschung, 15 (1971), pp. 1–14.

[31] D. WILHELMSEN, *A nearest point algorithm for convex polyhedral cones and applications to positive linear approximation*, Math. Comp., 30 (1976), pp. 48–47.

[32] P. WOLFE, *Algorithm for a least-distance programming problem*, Math. Programming Stud., 1 (1974), pp. 190–205.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 4, July 1980

0363-0129/80/1804-0010 $01.00/0

EXISTENCE AND UNIQUENESS OF REALIZATIONS OF NONLINEAR SYSTEMS*

BRONISŁAW JAKUBCZYK†

Abstract. Necessary and sufficient conditions for the existence of either analytic or smooth symmetric realizations of nonlinear input-output maps are given. It is also shown that two minimal realizations are diffeomorphic.

1. Introduction. Let G be a group with a multiplication $R \times G \to G$, and let p be a function $p: G \to R$. In the present paper, we state a theorem (Theorem 3) which gives a necessary and sufficient condition of G to be represented as a group of diffeomorphisms $\{\phi_a\}_{a \in G}$ on some (canonically defined) manifold X such that $p(a) = h(\phi_a(x_0))$, where h is a function $h: X \to R$ and $x_0 \in X$. We apply this theorem for solving the nonlinear realization problem.

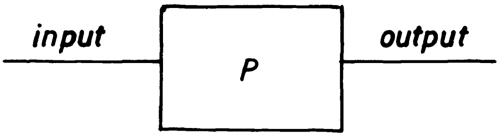


FIG. 1

The realization problem can be formulated as follows. Given a system ("black box") with an "input" and "output" (see Fig. 1), we put some functions (controls), with values in a set Ω, at the input and get some functions (with values in R^r) at the output. The set of pieces of input functions (on finite intervals) can be regarded as a semigroup S with multiplication being concatenation. We assume that the output function y at time $t > 0$ is uniquely defined by the piece u of a control function on the interval $[0, t]$; i.e., the following map is given

$$p: S \to R^r. \tag{1}$$

The triple (S, p, R^r) will be called an input-output system.

We ask whether there exists a manifold X (state space) with a distinguished point $x_0 \in X$ and functions $f: X \times \Omega \to TX$, $h: X \to R^r$ such that the control system

$$\dot{x} = f(x, u), \quad x(0) = x_0, \quad y = h(x) \tag{2}$$

realizes our system (S, p, R^r). This means that for any $u \in S$,

$$p(u) = h(\phi^f_u(x_0)). \tag{3}$$

Here and below, ϕ^f_u is a diffeomorphism $X \to X$, $\phi^f_u(x) = \varphi(t_u, x)$, where $\varphi(t, x)$ is the solution of $\dot{\varphi} = f(\varphi, u)$, $\varphi(0, x) = x$, and t_u denotes the period of action of u.

There is a wide literature known as "linear realization theory" devoted to the linear case of the problem (cf. Kalman et al. [4]).

A realization theory was also developed for some special classes of nonlinear systems, such as bilinear systems (let us only mention the paper by d'Alessandro, Isidori and Ruberti [2]) or systems on groups (cf. Brockett [1]).

Recently, an important step in the extension of linear theory to the general, nonlinear case was made by Sussmann [10] (cf. also Sussmann [8], and Hermann and

* Received by the editors January 16, 1979, and in revised form September 23, 1979.

† Institute of Mathematics, Polish Academy of Sciences, 00-950 Warsaw, Sniadeckich 8, Poland.

Krener [5]). Roughly speaking, the main result of [10] can be formulated as follows: if the system (S, p, R^r) has a realization (symmetric in the nonanalytic case), then it has a minimal realization which is unique up to a diffeomorphism (for the definition of a minimal realization see § 2).

In this paper, we give sufficient and necessary conditions for existence of realizations of the system (S, p, R^r). We also show that two minimal realizations are diffeomorphic. Our construction gives a minimal realization; that is, essentially, our result implies the result of Sussmann [10].

In our approach, we use some ideas of Lobry [6] and Sussmann [9] on group theoretical approach to control systems.

Throughout the paper we use the convention that for any set A, the symbol A^k denotes the Cartesian product $A^k = A \times \cdots \times A$, k-times, and $A^\infty = \bigcup_{k=1}^\infty A^k$. The symbols A^k, A^∞ have the set theoretical meaning only—even if A has some algebraic structure. Composition of functions φ, ψ is denoted by $\varphi \circ \psi$, and we often write $\varphi \circ \psi(x)$ instead of $(\varphi \circ \psi)(x)$.

2. Piecewise constant controls. Let Ω be a set of admissible values of controls (its elements will be denoted α, β). First we shall consider the case of piecewise constant controls, so we do not assume any structure in Ω.

Denote by

$$a = (t_k\alpha_k) \cdots (t_2\alpha_2)(t_1\alpha_1) \tag{4}$$

the function $[0, \sigma_k) \to \Omega$, $a(\tau) = \alpha_i$ for $\tau \in [\sigma_{i-1}, \sigma_i)$, where $\sigma_i = \sum_{j=1}^i t_j$, $(\sigma_0 = 0)$, $t_i \in R_+ = [0, \infty)$ and $k \geqq 0$. The set of all such functions will be denoted by S and its elements by a, b, c, d. S has a natural structure of semigroup with multiplication

$$ba = (\tau_m\beta_m) \cdots (\tau_1\beta_1)(t_k\alpha_k) \cdots (t_1\alpha_1), \tag{5}$$

where $b = (\tau_m\beta_m) \cdots (\tau_1\beta_1)$. The identity e in S is an empty sequence (4).

There is a natural action of R_+ on S,

$$ta = ((tt_k)\alpha_k) \cdots ((tt_1)\alpha_1) \tag{6}$$

(expansion). We identify $\alpha = (1\alpha)$.

The semigroup S can be extended to a group denoted by G_S, which is defined in the following way (cf. [6]). The elements of G_S are formal sequences of the form (4) with $t_i \in R$ and multiplication defined by (5), where we identify $(t_1\alpha)(t_2\alpha) = (t_1 + t_2)\alpha$ and $(0\alpha) = e$. The element ta is defined by (6) for $t > 0$, and by $ta = ((tt_1)\alpha_1) \cdots ((tt_k)\alpha_k)$ for $t < 0$.

Define the *input–output mapping* of a system as a function $p: S \to R^r$.

The triple (S, p, R^r) will be called an *input–output system* or simply a *system*.

Let us denote $\underline{b} = (b_1, \cdots, b_m)$, $m \geqq 1$, $\underline{a} = (a_1, \cdots, a_p)$, $p \geqq 1$, $\underline{t} = (t_1, \cdots, t_p)$, $t_i \in R_+$, and $\psi_{\underline{a}}^{\underline{b}}: R_+^p \to R^{mr}$ as $\psi_{\underline{a}}^{\underline{b}} = (\psi_{\underline{a}}^{b_1}, \cdots, \psi_{\underline{a}}^{b_m})$, where

$$\psi_{\underline{a}}^{b_i}(\underline{t}) = p(b_i(t_pa_p) \cdots (t_1a_1)). \tag{7}$$

It is useful to view $(t_pa_p) \cdots (t_1a_1)$ as a basic control, and b_i as measure experiments.

We shall say that the system (S, p, R^r) is of class C^k (C^k-smooth) for $k = 0, 1, \cdots, \infty, \omega$ if

(A1) the functions $\psi_{\underline{a}}^{\underline{b}}$ are of class C^k, $\forall \underline{a}, \underline{b} \in S^\infty$
(have analytic extensions onto R^p in the case of $k = \omega$).

Above and throughout the paper "analytic" will mean real analytic. Define

$$\text{(8)} \qquad \operatorname{rank} p = \sup_{\underline{a},\underline{b},\underline{t}} \operatorname{rank} D\psi_{\underline{a}}^{\underline{b}}(\underline{t}), \qquad \underline{a}, \underline{b} \in S^{\infty}, \quad \underline{t} \in R_{+}^{\infty}.$$

In order to obtain the existence of realizations we shall assume that

$$\text{(A2)} \qquad \operatorname{rank} p = n < \infty.$$

We shall call the system (S, p, R^r) *time-invertible* if

$$\forall \alpha, \exists \beta, \forall a, b, \forall t > 0; \qquad p(b(t\beta)(t\alpha)a) = p(ba) = p(b(t\alpha)(t\beta)\alpha).$$

Let s be a function $\Omega \to \Omega$ and let $c \in S$ be defined on the interval $[0, t)$. By c_s we shall denote the function from S defined by $c_s(\tau) = s(c(t-\tau))$ a.e. in the interval $[0, t)$. The above condition can then be expressed in the following equivalent form.

(A3) There is a function $s: \Omega \to \Omega$ such that
$p(bc_sca) = p(ba) = p(bcc_sa)$ for any $a, b, c \in S$.

By a C^k *realization* $(k = 2, \cdots, \infty, \omega)$ of the system (S, p, R^r), we shall mean a quadruple (X, f, h, x_0), where X is a C^k manifold (Hausdorff, without boundary), and f is a function $f: X \times \Omega \to TX$ such that, for any $\alpha \in \Omega$, the function $\phi^f_{(t\alpha)}(x)$ (see (3)) is well-defined and of class C^k with respect to $(t, x) \in R \times X$ (i.e., the vector field $f(\cdot, \alpha)$ is complete and $\phi^f_{(t\alpha)}$ is a C^k flow). The function $h: X \to R^r$ is of class C^k, and x_0 is a point of X such that the input–output mapping of the control system (2) is equal to p, i.e., $p(a) = h(\phi^f_a(x_0))$ for every $a \in S$.

The realization (X, f, h, x_0) will be called *reachable* (*weakly reachable*) if for any $x_1 \in X$, there is $a \in S$ $(a \in G_S)$ such that $\phi^f_a(x_0) = x_1$. Here, for $a \in G_S$ of the form (4), we define $\phi^f_a = \phi^f_{(t_k\alpha_k)} \circ \cdots \circ \phi^f_{(t_1\alpha_1)}$ with $\phi^f_{(t\alpha)} = (\phi^f_{(-t\alpha)})^{-1}$ for $t < 0$ (inversion of time). The realization will be called *observable* if for any $x_1, x_2 \in X$, $x_1 \neq x_2$, there is $b \in S$ such that $h(\phi^f_b(x_1)) \neq h(\phi^f_b(x_2))$. A reachable and observable realization will be called *minimal.* A weakly reachable and observable C^ω realization will be called C^ω*-minimal* (minimal in the class C^ω). The realization is *symmetric* if for any $\alpha \in \Omega$, there is $\beta \in \Omega$ such that $f(\cdot, \alpha) = -f(\cdot, \beta)$.

We shall say that two realizations (X, f, h, x_0) and (X', f', h', x_0') of the system (S, p, R^r) are C^k*-diffeomorphic* if there is a C^k diffeomorphism $\chi: X \to X'$, which carries (X, f, h, x_0) to (X', f', h', x_0'), i.e., $f' = (D\chi \cdot f) \circ \chi^{-1}$, $h' = h \circ \chi^{-1}$, $x_0' = \chi(x_0)$.

THEOREM 1. a) *Every* C^k*-smooth,* $k = 2, \cdots, \infty$, *time-invertible system* (S, p, R^r) *with finite rank has a minimal, symmetric,* C^k *realization* (X, f, h, x_0), *where* $\dim X = \operatorname{rank} p$.

Any two minimal, C^k *realizations of the same input–output system are* C^k*-diffeomorphic.*

b) *Every* C^ω *system* (S, p, R^r) *with finite rank has a* C^ω*-minimal realization such that* $\dim X = \operatorname{rank} p$.

Any two C^ω*-minimal realizations of the same input–output system are* C^ω*-diffeomorphic.*

Remark 1. Conditions (A1) and (A2) are necessary for the existence of a C^k realization. In fact, (A1) follows from the definition of a C^k realization, and (A2) is a consequence of the inequalities

$$\text{(9)} \qquad \operatorname{rank} p \leqq \max_{\underline{a},\underline{b},\underline{t}} \max \{\operatorname{rank} D\psi_{\underline{a}}(\underline{t}), \operatorname{rank} D\psi^{\underline{b}}(\psi_{\underline{a}}(\underline{t}))\} \leqq \dim X,$$

where $\psi_{\underline{a}}: R_+^p \to X$ and $\psi^{\underline{b}}: X \to R^{mr}$ are defined by

$$\psi_{\underline{a}}(\underline{t}) = \phi^f_{(t_p a_p)\cdots(t_1 a_1)}(x_0), \tag{10}$$

$$\psi^{\underline{b}} = (\psi^{b_1}, \cdots, \psi^{b_m}), \qquad \psi^{b_i}(x) = h(\phi^f_{b_i}(x)). \tag{11}$$

(By (3) we have $\psi^{\underline{b}}_{\underline{a}} = \psi^{\underline{b}} \circ \psi_{\underline{a}}$.)

Axiom (A3) is necessary for the symmetry of a realization.

Remark 2. Any C^k minimal realization of the system (S, p, R^r) satisfying (A3) is symmetric, i.e., $\forall \alpha$, $\exists \beta$, $f(\cdot, \alpha) = -f(\cdot, \beta)$, where $\beta = s(\alpha)$. In fact, if $f(x, \alpha) \neq -f(x, \beta)$, then there is $a \in S$ and $t > 0$ such that $\phi^f_{(t\alpha)} \circ \phi^f_{(t\beta)} \circ \phi^f_a(x_0) \neq \phi^f_a(x_0)$ (by reachability), and so there is $b \in S$ such that $h(\phi^f_b \circ \phi^f_{(t\alpha)} \circ \phi^f_{(t\beta)} \circ \phi^f_a(x_0)) \neq h(\phi^f_b \circ \phi^f_a(x_0))$ (by observability). This means that $p(b(t\alpha)(t\beta)a) \neq p(ba)$, which contradicts (A.3).

3. General controls. Now we shall consider systems with more general classes of control functions. To have solutions of realization (2) to be well-defined, we should choose controls regular enough and have some regularity of $f(x, \alpha)$ with respect to α.

Assume that Ω is a metric space. Let $\mathfrak{M}_c([-1, 1]; \Omega)$ denote the set of measurable functions $\varphi: [-1, 1] \to \Omega$ which have the closure of the set of values $\{\varphi(t), t \in [-1, 1]\}$ compact. Let $\mathcal{U} \subset \mathcal{M}_c([-1, 1]; \Omega)$ be a class of functions containing constant functions. We define a semigroup $\tilde{S}$ as the set of functions which are piecewise of the form $u(t) = \varphi(\mu t)$, where $\varphi \in \mathcal{U}$ and $\mu > 0$. More precisely $u \in \tilde{S}$ iff

$$u = (t_k \varphi_k) \cdots (t_1 \varphi_1), \tag{12}$$

where $t_i \geqq 0$, $\varphi_i \in \mathcal{U}$, and

$$u(\tau) = \varphi_i\left(\frac{2\tau - (\sigma_{i-1} + \sigma_i)}{t_i}\right) \quad \text{for } \tau \in [\sigma_{i-1}, \sigma_i) \tag{13}$$

($\sigma_i = \sum_{j=1}^i t_j$). We denote elements of $\tilde{S}$ by u, v. Note that the representation (12) of the function u is not necessarily unique (e.g., if $\mathcal{U}$ is the set of continuous functions and $\varphi_1(1) = \varphi_2(-1)$, then there is a continuous function $\varphi \in \mathcal{U}$ such that $(t_1 + t_2)\varphi = (t_2\varphi_2)(t_1\varphi_1)$).

Multiplication and action of R_+ on $\tilde{S}$ can be defined by (5) and (6), analogously to the case of S. There is a natural imbedding $S \subset \tilde{S}$.

The semigroup $\tilde{S}$ can be extended to a group $G_{\tilde{S}}$ in the following way. We define elements of $G_{\tilde{S}}$ of the form (12) (formally), where $t_i \in R$ and $\varphi \in \mathcal{U}$. Multiplication and action of R_+ on $G_{\tilde{S}}$ are defined by (5) and (6). For u of the form (12) and $t < 0$ we define

$$tu = ((tt_1)\varphi_1) \cdots ((tt_k)\varphi_k), \qquad t_i \in R.$$

We identify elements u of the form (12) with $t_i \in R_+$ if they define the same functions on the same interval. We also identify the product $(tu)((-t)u)$ with the identity, i.e., the inverse is defined by $u^{-1} = (-1)u$.

Similarly as in the previous case, the triple $(\tilde{S}, p, R^r)$, where p is a map $\tilde{S} \to R^r$, is called an input–output system. We use the notation and repeat all the definitions from § 2, replacing $\alpha \in \Omega$ by $\varphi \in \mathcal{U}$, S by $\tilde{S}$, and G_S by $G_{\tilde{S}}$. Time-invertibility is defined by (A3) (replacing S by $\tilde{S}$), where we additionally assume that for any function $\varphi \in \mathcal{U}$ the function φ^* defined by $\varphi^*(\tau) = s(\varphi(-\tau))$ belongs to $\mathcal{U}$.

Additionally, we shall say that the system $(\tilde{S}, p, R^r)$ is *continuous* if the following two conditions are satisfied.

(A4a) For any functions u, u_k, $k \geqq 1$, from $\tilde{S}$ defined on the interval $[0, T)$ and such that $u_k(t) \to u(t)$ almost everywhere on $[0, T)$, we have $p(u_k) \to p(u)$.

(A4b) For any $\underline{u}, \underline{v} \in \tilde{S}^{\infty}$ the function $g(\underline{t}, \alpha) = (d/dt)\psi^{\underline{v}}_{(\underline{u},\alpha)}(\underline{t}, t)/t = 0^+$ is continuous with respect to the variables $(\underline{t}, \alpha) \in R^{p}_{+} \times \Omega$ (jointly) together with the first derivative with respect to $\underline{t}$ (here $(\underline{u}, \alpha)$ denotes the sequence $(u_1, \cdots, u_p, \alpha)$, $\alpha = (1a) \in S$).

For the case of the system $(\tilde{S}, p, R^r)$ of class C^{ω}, the map: $\tilde{S} \rightarrow R^r$ can be extended to $G_{\tilde{S}}$, and functions $\psi^{\underline{b}}_{\underline{a}}$ can be defined by (7) for $\underline{a}, \underline{b}, \in G^{\infty}_{\tilde{S}}$ (we shall use the notation $\underline{u}, \underline{v} \in G^{\infty}_{\tilde{S}}$ and $\psi^{\underline{v}}_{\underline{u}}$). The extended map $\bar{p}$ is defined as follows. For $u = (t_k\varphi_k) \cdots (t_1\varphi_1)$, we define

$$(14) \qquad \bar{p}(u) = \psi^{e}_{\underline{\varphi}}(\underline{t}),$$

where $\underline{t} = (t_1, \cdots, t_k) \in R^k$, $\underline{\varphi} = ((1\varphi_1), \cdots, (1\varphi_k)) \in \tilde{S}^{\infty}$, and e denotes identity (one-element sequence).

Although the representation $u = (t_k\varphi_k) \cdots (t_1\varphi_1)$ is not unique, the definition of the extended map $\bar{p}$ is correct by the uniqueness of analytic extensions and the fact that a restriction onto a hyperplane of an analytic function is analytic. The same arguments imply that the formula (7) holds for $\underline{a}, \underline{b} \in \tilde{S}^{\infty}$ and $\underline{t} \in R^{\infty}$.

We shall call an analytic system $(\tilde{S}, p, R^r)$ *completely continuous* if the conditions (A4a) and (A4b) are satisfied where, in (A4b), $\tilde{S}^{\infty}$ is replaced by $G^{\infty}_{\tilde{S}}$, R^{p+1}_{+} by R^{p+1}, and p by $\bar{p}$.

We shall say that a realization (X, f, h, x_0) of the system $(\tilde{S}, p, R^r)$ is *continuous* if $f(x, \alpha)$ is continuous with respect to $(x, \alpha) \in X \times \Omega$ together with the first order derivative with respect to x. The notion "realization of $(\tilde{S}, p, R^r)$" includes the fact that the system (2) has a (unique) solution on any interval $[0, T]$ for any control function $\theta \in \tilde{S}$ defined on [0, T]. It is well-known that the following "approximating property" holds for a function $f(x, \alpha)$ of the above described regularity. If u_k is a sequence of functions from $\tilde{S}$ defined on the interval $[0, T)$, and $u_k(t) \rightarrow u(t)$ almost everywhere on $[0, T)(u \in \tilde{S})$, then $\phi^f_{u_k}(x_0) \rightarrow \phi^f_u(x_0)$.

THEOREM 2. a) *Every continuous C^k-smooth ($k = 2, 3, \cdots, \infty$), time-invertible system $(\tilde{S}, p, R^r)$ with finite rank has a minimal, continuous, symmetric C^k realization (X, f, h, x_0), where* $\dim X = \operatorname{rank} p$.

Any two minimal, continuous C^k realizations of the same input–output system are C^k-diffeomorphic.

b) *Every completely continuous C^{ω} system $(\tilde{S}, p, R^r)$ with finite rank has a C^{ω}-minimal, continuous realization (X, f, h, x_0), where* $\dim X = \operatorname{rank} p$.

Any two C^{ω}-minimal, continuous realizations of the same input–output system are C^{ω}-diffeomorphic.

Remark 3. It is not difficult to see that conditions (A1), (A2), (A4a) and (A4b) are necessary for existence of a continuous, C^k realization. Axiom (A3) is necessary for the symmetry of the realization. In the case of minimal realization (A3) is also sufficient for its symmetry (by arguments as in Remark 2).

4. Minimal representations of abstract systems. In this section, we state a general theorem of the type of Theorems 1, 2 with semigroups S, $\tilde{S}$ replaced by a group G. We shall derive both Theorems 1 and 2 from this theorem.

Let G be a group with a surjective map $R \times G \rightarrow G$. For the notational convenience we denote this map by $(t, a) \rightarrow ta$, though it is not assumed to have any properties of multiplication (e.g., we do not assume that $t_1(t_2a) = (t_1t_2)a$). Let p be a map $p: G \rightarrow R^r$. The triple (G, p, R^r) will be called an *abstract system*. A system of class C^k and rank p are defined as in § 2, where R_+ is replaced by R and S by G.

By a C^k *representation* ($k = 1, \cdots, \infty, \omega$) of the system (G, p, R^r), we shall mean a quadruple $(X, \{\phi_a\}_{a \in G}, h, x_0)$, where

(i) X is a C^k manifold (Hausdorff, without boundary) with a distinguished point $x_0 \in X$;

(ii) ϕ_a, $a \in G$, are C^k diffeomorphisms of X, such that

$$\phi_{ba} = \phi_b \circ \phi_a, \tag{15}$$

and the map $\psi_{\underset{\sim}{a}}: R^p \to X$ defined by

$$\psi_{\underset{\sim}{a}}(\underset{\sim}{t}) = \phi_{(t_pa_p)\cdots(t_1a_1)}(x_0) \tag{16}$$

is of class C^k for any $\underset{\sim}{a} \in G^\infty$;

(iii) the function $h: X \to R^r$ is of class C^k;

(iv) the "input–output map" of the representation is equal to p, i.e.,

$$p(a) = h(\phi_a(x_0)) \quad \text{for } a \in G. \tag{17}$$

The representation is called *transitive* if for any $x_1, x_2 \in X$ there is $a \in G$ such that $\phi_a(x_1) = x_2$. It will be called *distinguishable* if for any $x_1, x_2 \in X$, $x_1 \neq x_2$ there is $a \in G$ such that $h(\phi_a(x_1)) \neq h(\phi_a(x_2))$. A transitive and distinguishable representation is called *minimal.*

We call two representations $(X, \{\phi_a\}_{a\in G}, h, x_0)$ and $(X', \{\phi'_a\}_{a\in G}, h', x'_0)$ C^k-*diffeomorphic* if there is a C^k diffeomorphism $\chi: X \to X'$ such that $\varphi'_a \circ \chi = \chi \circ \phi_a$ for $a \in G$, $h = h' \circ \chi$ and $x'_0 = \chi(x_0)$.

THEOREM 3. *Every C^k-smooth, $k = 1, 2, \cdots, \infty, \omega$, abstract system (G, p, R^r) with finite rank has a minimal C^k representation $(X, \{\phi_a\}_{a\in G}, h, x_0)$ with* $\dim X = \operatorname{rank} p$.

Any two minimal C^k representations of (G, p, R^r) are C^k-diffeomorphic.

Remark 4. It is easy to see that C^k-smoothness of (G, p, R^r) and finiteness of the rank are also necessary for the existence of a C^k representation.

5. Proof of Theorem 3.

Existence. We shall construct our representation in 3 steps: set theoretical (without using axioms (A1), (A2)), topological (using only (A1) with $k = 0$) and smooth (using (A1) and (A2)). Similarly, Theorems 1 and 2 could be considered at these three levels (replacing vector fields by one parameter groups of transformations).

A. Set theoretical representation. We define the following equivalence relation in G,

$$a \sim b \Leftrightarrow p(ca) = p(cb) \quad \forall c \in G.$$

Our state space X is defined as the quotient space

$$X = G/\sim,$$

and $[a]$ denotes the equivalence class of the element $a \in G$. We define

$$\phi_a(x) = [ab], \quad h(x) = p(b), \quad x_0 = [e], \tag{18}$$

where $x = [b]$. It is easy to see that these definitions are correct, conditions (15) and (17) are satisfied, and the representation $(X, \{\phi_a\}_{a\in G}, h, x_0)$ is minimal (we shall call this representation canonical). Note that the mappings ϕ_a are invertible since $\phi_a \circ \phi_{a^{-1}}(x) = [aa^{-1}b] = x = [a^{-1}ab] = \phi_{a^{-1}} \circ \phi_a(x)$.

B. Topological representation. We shall introduce a topology in X and prove that ϕ_a are homeomorphisms, and h is continuous.

Define $\psi_{\underline{a}}: R^p \to X$ and $\psi^{\underline{b}}: X \to R^{mr}$ by

$$\psi_{\underline{a}}(\underline{t}) = \phi_{(t_pa_p)\cdots(t_1a_1)}(x_0) \tag{19}$$

and

$$\psi^{\underline{b}} = (\psi^{b_1}, \cdots, \psi^{b_m}), \tag{20}$$

where

$$\psi^{b_i}([a]) = h(\phi_{b_i}([a])) = p(b_ia). \tag{21}$$

Obviously, we have $\psi^{\underline{b}} \circ \psi_{\underline{a}} = \psi^{\underline{b}}_{\underline{a}}$, where $\psi^{\underline{b}}_{\underline{a}}$ is defined by (7).

We define the topology in X as the strongest topology such that $\psi_{\underline{a}}$ are continuous for every $\underline{a} \in G^{\infty}$. A subset $U \subset X$ is open in this topology iff $\psi_{\underline{a}}^{-1}(U)$ is open for any $\underline{a}$. The mappings $\psi^{\underline{b}}: X \to R^{mr}$ are continuous with respect to this topology since, by (A1), the mappings $\psi^{\underline{b}}_{\underline{a}} = \psi^{\underline{b}} \circ \psi_{\underline{a}}$ are continuous.

To prove that the topology is Hausdorff, take $x_1, x_2 \in X, x_1 \neq x_2$. From the definition of X there is $b \in G$ such that $\psi^b(x_1) \neq \psi^b(x_2)$. The conclusion follows from continuity of the map ψ^b. The continuity of h is a consequence of $h = \psi^e$, which follows from (18) and (21). To show that $\phi_a: X \to X$ is continuous, we prove that if a subset $U \subset X$ is open in X, then the set $V = \phi_a^{-1}(U)$ is open in X, i.e., $\psi_{\underline{a}}^{-1}(V)$ is open for any $\underline{a}$. In fact, from (15) and (16) we have $\phi_a \circ \psi_{\underline{a}}(\underline{t}) = \psi_{\underline{a}'}(\underline{s}')$, where $\underline{a}' = (a_1, \cdots, a_p, a')$, $t' = (t_1, \cdots, t_p, t')$ and $a = (t'a')$. Therefore, the set $\psi_{\underline{a}}^{-1}(V) \subset R^p$ may be identified with the set $\psi_{\underline{a}'}^{-1}(U) \cap (R^p \times \{t'\})$, which is open in $R^p \times \{t'\}$ by openess of U. Continuity of the inverse follows from $\phi_a^{-1} = \phi_{a^{-1}}$.

C. Representation of class C^k. Structure of n-manifold on X. We shall show that any point x of X has a neighborhood homeomorphic to an open subset of R^n. By transitiveness of the group of mappings $\{\phi_a\}_{a \in G}$, and the fact that $\phi_a, a \in G$ are homeomorphisms, all we have to do is to prove the existence of such a neighborhood for one point $x \in X$.

Take $\underline{a}, \underline{b}, \underline{t}_0$ such that

$$\operatorname{rank} D\psi^{\underline{b}}_{\underline{a}}(\underline{t}_0) = n = \operatorname{rank} p.$$

The function $\psi^{\underline{a}}_{\underline{b}} = \psi^{\underline{b}} \circ \psi_{\underline{a}}$ has p independent variables $\underline{t} = (t_1, \cdots, t_p)$, where p is, in general, greater than n. We fix $p - n$ of these variables (the fixed variables will be equal to their values at the point $\underline{t}_0$) and leave the remaining n of them, $\underline{\tau} = (t_{i_1}, \cdots, t_{i_n})$, varying. We choose the variables $\underline{\tau}$ in such a way that the differential $D\psi^{\underline{b}}_{\underline{a}}(\underline{t}_0)$, taken along these variables only, has the full rank n. (In other words, the $rm \times n$ submatrix of $D\psi^{\underline{b}}_{\underline{a}}(\underline{t}_0)$ which consists of columns with indices $i_1, \cdots, i_n$ has the rank equal to n.) We obtain the reduced functions $\tilde{\psi}_{\underline{a}}: R^n \to X$ and $\tilde{\psi}^{\underline{b}}_{\underline{a}} = \psi^{\underline{b}} \circ \tilde{\psi}_{\underline{a}}: R^n \to R^{rm}$, where

$$\tilde{\psi}_{\underline{a}}(\underline{\tau}) = \psi_{\underline{a}}(\underline{t}(\underline{\tau})).$$

Clearly, the map $\tilde{\psi}^{\underline{b}}_{\underline{a}}$ is of class C^k and rank $D\tilde{\psi}^{\underline{b}}_{\underline{a}}(\underline{\tau}_0) = n$ for $\underline{\tau}_0$ corresponding to $\underline{t}_0$. Therefore, there is a neighborhood $U \subset R^n$ of the point $\underline{\tau}_0$ such that $D\tilde{\psi}^{\underline{b}}_{\underline{a}}$ has the full rank on U, the set $M = \tilde{\psi}^{\underline{b}}_{\underline{a}}(U)$ is an n-dimensional submanifold of R^{rm} of class C^k and $\tilde{\psi}^{\underline{b}}_{\underline{a}}/U: U \to M$ is a C^k diffeomorphism between U and M.

Denote $\psi = \tilde{\psi}_{\underline{a}}/U$, $\psi: U \to X$. The function ψ is injective which follows from the injectivity of $\tilde{\psi}^{\underline{b}}_{\underline{a}}/U = \psi^{\underline{b}} \circ \psi$. Thus ψ is a continuous (by continuity of $\psi_{\underline{a}}$) bijection between $U \subset R^n$ and $V = \psi(U)$. We shall show that ψ is a homeomorphism between U and the open subset $V = \psi(U)$ of X. By the remark at the beginning of this paragraph, this will imply that X is locally homeomorphic to R^n.

We have only to prove that ψ is an open mapping, i.e., for any open subset $U' \subset U$, the set $V' = \psi(U')$ is open in X. We show this for $U' = U$ and $V' = V$ (the proof in the general case is analogous).

Assume that $V = \psi(U)$ is not open in X. From the definition of the topology in X it follows that there exist $q \geqq 1$ and $\underline{c} \in G^q$ such that the set $\psi_{\underline{c}}^{-1}(V)$ is not open in R^q. Then there is a point $\underline{s}' \in R^q$ such that

$$(22) \qquad \underline{s}' \in U_1 = \psi_{\underline{c}}^{-1}(V) \quad \text{and} \quad s' \notin \operatorname{int} U_1.$$

Denote $c = (s'_q c_q) \cdots (s'_1 c_1)$. By the assumption that the map $R \times G \to G$ is surjective, we have that $c^{-1} = (s'c')$ for some $s' \in R$ and $c' \in G$. Consider the map

$$\tilde{\psi}_{\underline{a}c'\underline{c}}(\underline{\tau}, \underline{s}) = \psi_{\underline{a}c'\underline{c}}(\underline{t}(\underline{\tau}), s', \underline{s}),$$

where $\underline{a}c'\underline{c}$ is the sequence $(a_1, \cdots, a_p, c', c_1, \cdots, c_q)$, and $\underline{t}(\underline{\tau})$ is the function $R^n \to R^p$ defined before. Since $\underline{s}' \in \psi_{\underline{c}}^{-1}(V)$, then there is a point $\underline{\tau}' \in R^n$ such that

$$\tilde{\psi}_{\underline{a}}(\underline{\tau}') = \psi_{\underline{a}}(\underline{t}(\underline{\tau}')) = \psi_{\underline{c}}(\underline{s}') = \phi_c(x_0)$$

and so

$$\phi_{(s'c')}(\tilde{\psi}_{\underline{a}}(\underline{\tau}')) = \phi_{c^{-1}}(\tilde{\psi}_{\underline{a}}(\underline{\tau}')) = x_0.$$

Therefore, by the definition of $\tilde{\psi}_{\underline{a}c'\underline{c}}(\underline{\tau}, \underline{s})$ and (16) we find that

$$(23) \qquad \tilde{\psi}_{\underline{a}c'\underline{c}}(\underline{\tau}', \underline{s}) = \psi_{\underline{c}}(\underline{s}).$$

Similarly, from the fact that $(s'_q c_q) \cdots (s'_1 c_1)(s'c') = e$, we obtain

$$(24) \qquad \tilde{\psi}_{\underline{a}c'\underline{c}}(\underline{\tau}, \underline{s}') = \tilde{\psi}_{\underline{a}}(\underline{\tau}).$$

From (24), the fact that $\operatorname{rank} D\tilde{\psi}^b_{\underline{a}}(\underline{\tau}') = n$ and assumption (A2), it follows that

$$\operatorname{rank} D\tilde{\psi}^b_{\underline{a}c'\underline{c}}(\underline{\tau}, \underline{s}) = n$$

on a neighborhood $W \subset R^{n+q}$ of the point $(\underline{\tau}', \underline{s}')$, where $\tilde{\psi}^b_{\underline{a}c'\underline{c}} = \psi^b \circ \tilde{\psi}_{\underline{a}c'\underline{c}}$. This means that the level sets $\{(\underline{\tau}, \underline{s}) \in W | \tilde{\psi}_{\underline{a}c'\underline{c}}(\underline{\tau}, \underline{s}) = \text{const}\}$ are submanifolds of $W \subset R^{n+q}$ of dimension q (actually, they form a foliation of W of codimension n). The level submanifold passing through $(\underline{\tau}', \underline{s}')$ intersects the n-dimensional submanifold $U \times \{\underline{s}'\}$ transversally (cf. Fig. 2) by the fact that $\operatorname{rank} D\tilde{\psi}^b_{\underline{a}}(\underline{\tau}') = n$, (24) and the formulas

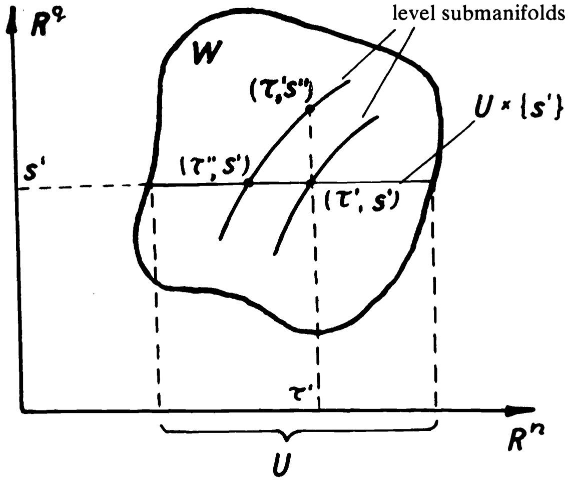


FIG. 2

$\tilde{\psi}^{\underline{b}}_{\underline{a}} = \psi^{\underline{b}} \circ \tilde{\psi}_{\underline{a}}$, $\tilde{\psi}^{\underline{b}}_{\underline{a}c'\underline{c}} = \psi^{\underline{b}} \circ \tilde{\psi}_{\underline{a}c'\underline{c}}$. Similarly, the level submanifolds, which are close to it, intersect $U \times \{\underline{s}'\}$ transversally. We shall use this fact to get a contradiction.

Take a point $\underline{s}'' \in R^q$ close enough to $\underline{s}'$ such that $\psi_{\underline{c}}(\underline{s}'') \notin V$ and the level submanifold passing through $(\underline{\tau}', \underline{s}'')$ intersects $U \times \{\underline{s}'\}$ (this is possible by (22) and the transversality of the level submanifolds and $U \times \{\underline{s}'\}$). We denote the point of intersection by $(\underline{\tau}'', \underline{s}')$. We can assume that $(\underline{\tau}'', \underline{s}')$ and $(\underline{\tau}', \underline{s}'')$ lie in the same connected component of the level submanifold.

We have that the point $x_1 = \tilde{\psi}_{\underline{a}c'\underline{c}}(\underline{\tau}', \underline{s}'') = \psi_{\underline{c}}(\underline{s}'')$ (cf. (23)) does not belong to V, and the point $x_2 = \tilde{\psi}_{\underline{a}c'\underline{c}}(\underline{\tau}'', \underline{s}') = \psi_{\underline{a}}(\underline{\tau}'')$ (cf. (24)) belongs to V. Therefore, $x_1 \neq x_2$.

Now, by the distinguishability of the representation $(X, \{\phi_a\}_{a \in G}, h, x_0)$, there is an element $b \in G$ such that

$$h \circ \phi_b(x_1) \neq h \circ \phi_b(x_2), \quad \text{that is,} \quad \psi^b(x_1) \neq \psi^b(x_2).$$

This implies that

$$\tilde{\psi}^{\underline{b}}_{\underline{a}c'\underline{c}}(\underline{\tau}'', \underline{s}') = \tilde{\psi}^{\underline{b}}_{\underline{a}c'\underline{c}}(\underline{\tau}', \underline{s}'')$$

and

$$\tilde{\psi}^{\underline{b}'}_{\underline{a}c'\underline{c}}(\underline{\tau}'', \underline{s}') \neq \tilde{\psi}^{\underline{b}'}_{\underline{a}c'\underline{c}}(\underline{\tau}', \underline{s}''),$$

where $\underline{b}' = (b_1, \cdots, b_m, b)$.

To complete the proof of the openness of V, it is enough to note that the latter sentence contradicts the following lemma.

LEMMA 1. *Let $f: W \to R^k$ and $\tilde{f}: W \to R^{k+k'}$ be maps of class C^1, where $W \subset R^m$ is an open subset and $\tilde{f}$ is of the form $\tilde{f} = (f, f')$. Assume, that* rank $D\tilde{f}(x) =$ rank $Df(x) =$ const $= n$ *on W. Then the connected components of the level submanifolds of f and $\tilde{f}$ coincide. In other words, if $f(x') = f(x'')$ and x', x'' are from the same connected component of the set $N = \{x \in U, f(x) = f(x')\}$, then $\tilde{f}(x') = \tilde{f}(x'')$.*

Proof. We have to show that the set of points $x \in N$ such that $\tilde{f}(x) = \tilde{f}(x')$ is closed and open in N. Its closedness follows from continuity of the function $\tilde{f}$. To show that it is open, we take a point x'' belonging to N such that $\tilde{f}(x'') = \tilde{f}(x')$. Since Df has the constant rank, we may change the coordinates in a neighborhood of x'' in such a way that the function f will depend only on the first n variables. Therefore, the differential Df will be of the form $Df = (f_{x_1}, \cdots, f_{x_n}, \cdots, 0)$ and so

$$D\tilde{f} = \begin{pmatrix} f_{x_1}, \cdots, f_{x_n}, 0, & \cdots, 0 \\ f'_{x_1}, \cdots, f'_{x_n}, f'_{x_{n+1}}, & \cdots, f'_{x_m} \end{pmatrix}$$

in a neighborhood of $x'' \in W$ and, locally, $N = \{x \mid x_i = x''_i, i = 1, \cdots, n\}$.

By the fact that rank $D\tilde{f}(x) =$ rank $Df(x)$, we find that $f'_{x_i} = 0$ for $i > n$, i.e., the functions $\tilde{f}$ do not depend on the variables x_i, $i > n$. This and the equality $\tilde{f}(x'') = \tilde{f}(x')$ imply that locally around x'', $\tilde{f}(x) = \tilde{f}(x')$ for $x \in N$ (by the form of N). The proof is complete.

C^k-differentiable structure on X. We endow X with a structure of a C^k manifold by introducing a differentiable structure $\mathscr{F}$ of class C^k (see Sternberg [7]). Define $\mathscr{F}$ as the set of real valued functions φ, each defined on an open subset of X, such that the function $\varphi \circ \psi_{\underline{a}}$ (see (19)) is of class C^k for any $\underline{a} \in G^\infty$. It is easy to see that φ satisfies the C^k differentiable structure axioms [7]:

(i) if $\psi \in \mathscr{F}$ is defined on U and $V \subset U$ is open, then $\varphi / V \in \mathscr{F}$;

(ii) if $V = \bigcup_\alpha U_\alpha$ and φ is a function defined on V such that $\varphi / U_\alpha \in \mathscr{F}$ for any α, then $\varphi \in \mathscr{F}$;

(iii) for any $x \in X$ there are open subsets $W \subset X$, $U \subset R^n$ $(x \in W)$ and a homeomorphism $g: W \to U \subset R^n$ such that for each open subset $V \subset W$ and for any function $\varphi: V \to R$, we have: $\varphi \in \mathscr{F}$ iff $\varphi \circ g^{-1}$ is of class C^k. The neighborhoods W, U and the homeomorphism g in condition (iii) can be constructed as in the proof that X is locally homeomorphic to $R^n (g = \psi^{-1}, W = \psi(U))$. We shall prove the nontrivial part of this condition: if $\varphi \circ g^{-1} \in C^k$ (i.e., $\varphi \circ \psi \in C^k$), then $\varphi \in \mathscr{F}$.

Take any $\underline{a} \in G^\infty$ and note that $\varphi \circ \psi_{\underline{a}} = \varphi \circ \psi \circ \psi^{-1} \circ \psi_{\underline{a}}$ (on the set $\psi_{\underline{a}}^{-1}(V)$). We have $\varphi \circ \psi \in C^k$ (by the assumption) and $\psi^{-1} \circ \psi_{\underline{a}} \in C^k$ by the equality $\psi^{-1} \circ \psi_{\underline{a}} = \tilde{\psi}_{\underline{a}}^{-1} \circ (\psi^b)^{-1} \circ \psi^b \circ \psi_{\underline{a}} = (\tilde{\psi}_{\underline{a}}^b)^{-1} \circ \psi_{\underline{a}}^b$ and the assumption (A1). Here ψ^b is treated as a map into the submanifold $M = \tilde{\psi}_{\underline{a}}^b(U)$ (see the proof that X is n-manifold) and all the mappings are defined on suitable chosen subsets determined by V. Therefore, $\varphi \circ \psi_{\underline{a}} \in C^k$ and so $\varphi \in \mathscr{F}$.

It remains to prove that ϕ_a, $a \in G$, and h are C^k-smooth. To prove that ϕ_a is of class C^k we show that for any function $\varphi: X \to R$, $\varphi \in C^k$, we have $\varphi \circ \phi_a \in C^k$. In fact, if $\varphi \in C^k$, then $\varphi \circ \psi_{\underline{a}} \in C^k$ for any $\underline{a}$ (by the definition of the differentiable structure on X). Take $\varphi \circ \phi_a \circ \psi_{\underline{a}}$ and note that $\phi_a \circ \psi_{\underline{a}}(t) = \psi_{\underline{a}'}(\underline{t}')$, where $\underline{a}' = (a_1, \cdots, a_p, a')$, $\underline{t}' = (t_1, \cdots, t_p, t')$ and $a = (t'a')$. Since $\varphi \circ \psi_{\underline{a}'} \in C^k$, then $\varphi \circ \phi_a \circ \psi_{\underline{a}} \in C^k$ and so $\varphi \circ \phi_a \in C^k$. The rank condition $\operatorname{rank} D\phi_a(x) = n$ follows from the equality $D\phi_a(x) \circ D\phi_{a^{-1}}(y) = Id$, where $y = \phi_a(x)$. The property $h \in C^k$ follows immediately from (A1) by the equality $h \circ \psi_{\underline{a}} = {}^{e}_{\underline{a}}$, which is a consequence of (18), (19) and (21). The fact that $\psi_{\underline{a}} \in C^k$ (see the definition of a C^k representation) is an immediate consequence of the definition of the C^k structure $\mathscr{F}$.

Uniqueness. It is enough to prove that any minimal, C^k representation $(X', \{\phi'_a\}_{a \in G}, h', x'_0)$ of the abstract system (G, p, R^r) is C^k-diffeomorphic to the canonical representation $(X, \{\phi_a\}_{a \in G}, h, x_0)$ constructed in the existence part of the proof.

We define our candidate for diffeomorphism $\chi: X \to X'$ by

$$\chi(\phi_a(x_0)) = \phi'_a(x'_0). \tag{25}$$

χ is well-defined by the distinguishability of $(X', \{\phi'_a\}_{a \in G}, h', x'_0)$. In fact, let $\phi_a(x_0) = \phi_b(x_0)$ and $\phi'_a(x'_0) \neq \phi'_b(x'_0)$. Then there is c such that $h(\phi'_c \circ \phi'_a(x_0)) \neq h(\phi'_c \circ \phi'_b(x_0))$, i.e., $h(\phi_c \circ \phi_a(x_0)) \neq h(\phi_c \circ \phi_b(x_0))$ and so $\phi_a(x_0) \neq \phi_b(x_0)$—a contradiction. If $\chi(\phi_a(x_0)) = \chi(\phi_b(x_0))$, then for any c, $h(\phi_c \circ \phi_a(x_0) = h(\phi'_c \circ \phi'_a(x'_0)) = h(\phi'_c \circ \phi'_b(x'_0)) = h(\phi_c \circ \phi_b(x_0))$, and by the distinguishability of the canonical representation $\phi_a(x_0) = \phi_b(x_0)$, i.e., χ is an injection. By the transitiveness of $(X', \{\phi'_a\}_{a \in G}, h', x'_0)$, the map χ is onto X', which implies that χ is a bijection $X \to X'$.

From the definition of χ, we have

$$\phi'_a \circ \chi(\phi_b(x_0)) = \phi'_a \circ \phi'_b(x'_0) = \phi'_{ab}(x'_0) = \chi(\phi_{ab}(x_0)) = \chi \circ \phi_a(\phi_b(x_0)),$$

$$h' \circ \chi(\phi_b(x_0)) = h' \circ \phi'_b(x'_0) = p(b) = h(\phi_b(x_0)),$$

i.e., $\phi'_a \circ \chi = \chi \circ \phi_a$, $h = h' \circ \chi$, and $\chi(x_0) = \chi(\phi_e(x_0)) = \phi'_e(x'_0) = x'_0$.

It remains to prove that χ is of class C^k, and $D\chi$ is nonsingular (i.e., χ is a C^k diffeomorphism). To prove the first fact we take any function $\varphi: X' \to R$ of class C^k and also show that $\varphi \circ \chi \in C^k$, i.e., $\varphi \circ \chi \circ \psi_{\underline{a}} \in C^k$ for any $\underline{a} \in G^\infty$. This follows from the form of the last function in:

$$\begin{aligned} \varphi \circ \chi(\psi_{\underline{a}}(\underline{t})) &= \varphi \circ \chi(\phi_{(t_p a_p) \cdots (t_1 a_1)}(x_0)) \\ &= \varphi(\phi'_{(t_p a_p) \cdots (t_1 a_1)}(x'_0)) \end{aligned}$$

and from the C^k-smoothness of the representation $(X', \{\phi'_a\}_{a \in G}, h', x'_0)$.

Nonsingularity of $D\chi$ follows from the equality $\psi^b_{\underline{a}} = \psi'^b \circ \chi \circ \psi_{\underline{a}}$, which is a consequence of (7), (15), (16), (17) and (25), where $\psi'^b = (\psi'^{b_1}, \cdots, \psi'^{b_m})$ and $\psi'^{b_i}(x) = h'(\phi'_{b_i}(x))$. In fact, for any $x \in X$ we may choose $\underline{a}, \underline{b}, \underline{t}$ such that rank $D\psi^b_{\underline{a}}(\underline{t}) = n$ and $\psi_{\underline{a}}(\underline{t}) = x$ (see the proof that X is n-manifold). From this and the C^k-smoothness of $\psi_{\underline{a}}$, χ and ψ'^b, we obtain that rank $D\chi(x) = n$. The proof is complete.

COROLLARY 1. *Theorem* 1 *holds with replacing R^r by a C^k finite dimensional manifold Y.*

Proof. The corollary follows from the proof of Theorem 3 as there we do not use the linear structure of R^r.

COROLLARY 2. *Let (G, p, Y) be a C^k abstract system with finite rank. If $(X, \{\phi_a\}_{a\in G}, h, x_0)$ and $(X', \{\phi'_a\}_{a\in G}, h', x'_0)$ are, respectively, minimal and distinguishable (transitive and minimal) representations, then there is a C^k one-to-one immersion (submersion onto) $\chi: X \to X'$ such that $\chi \circ \phi_a = \phi'_a \circ \chi$, $h = h' \circ \chi$* and $\chi(x_0) = x'_0$.

Proof. By Theorem 3 we may assume that $(X, \{\phi_a\}_{a\in G}, h, x_0)$ is the canonical representation of the system (G, p, Y), constructed in the proof of Theorem 3. We define χ by $\chi(\phi_b(x_0)) = \phi'_b(x'_0)$. The proof that χ has the required properties is analogous to that for the uniqueness part of Theorem 3.

In the proof of the dual version, we cannot assume that the representation $(X, \{\phi_\alpha\}_{a\in G}, h, x_0)$ is canonical. However, analogous arguments can be used since the property that a function $\varphi: X \to R$ is of class C^k iff $\varphi \circ \psi_{\underline{a}} \in C^k$, $\forall \underline{a} \in G^\infty$ holds for any transitive C^k representation. This property follows from the fact that for any $x \in X$, there are $\underline{a}$ and $\underline{t}^*$ such that rank $D\psi_{\underline{a}}(t^*) = \dim X$ (see e.g., Corollary 3) and $\psi_{\underline{a}}(\underline{t}^*) = x$; i.e., $\psi_{\underline{a}}$ is a C^k submersion in a neighborhood of $\underline{t}^*$.

In [9], Sussmann has proved that the orbit of a family of vector fields is an immersed submanifold. From Theorem 3 and Corollary 2 we obtain a generalization of this result in the case of complete vector fields.

COROLLARY 3. *Let G be a group with a surjective map $R \times G \to G$ (see the beginning of* § 4) *which acts C^k-smoothly on a manifold X. The latter means that there is a family of C^k diffeomorphisms $\{\phi_a\}_{a\in G}$ such that* (15) *is satisfied, and the mappings $\psi_{\underline{a}}: R^p \to X$ defined by* (10) *are C^k-smooth. Then each orbit of this group $Gx_0 = \{x \mid x = \phi_a(x_0),\ a \in G\}$ is a C^k immersed submanifold of X and* $\dim Gx_0 = \sup_{\underline{a},\underline{t}}$ rank $D\psi_{\underline{a}}(\underline{t})$.

Proof. Let (G, p, X) be an abstract system with $p(a) = \phi_a(x_0)$. Clearly it is C^k-smooth and has finite rank. Define its representation by $(X, \{\phi_a\}_{a\in G}, h, x_0)$, where $h = id: X \to X$ (it is C^k-smooth and distinguishable). We take any minimal representation $(X', \{\phi'_a\}_{a\in G}, h', x'_0)$ of the system (G, p, X) (it exists by Theorem 3) and define an immersion $\chi: X' \to X$ by Corollary 2. Obviously, we have that $\chi(X') = Gx_0$. To complete the proof note that for some $\underline{a}, \underline{b}, \underline{t}$, rank $D\psi^b_{\underline{a}}(\underline{t}) = \dim X'$, i.e., (by $\psi^b_{\underline{a}} = \psi^b \circ \psi_{\underline{a}}$) rank $D\psi_{\underline{a}}(\underline{t}) \geqq \dim X' = \dim Gx_0$. The converse inequality is obvious since $\psi_{\underline{a}}$ is a map into Gx_0.

Remark 5. The following property follows from the definition of the canonical manifold X and the definition (18) of the diffeomorphism ϕ_a.

If $p(bca) = p(bc'a)$ for any $a, b \in G$, then $\phi_c = \phi_c{}'$.

6. Proof of Theorem 1.

A. Existence. The system (S, p, R^r) can be extended to an abstract system $(G_S, \tilde{p}, R^r)$. We define

$$\tilde{p}(a) = p(\tilde{a}),$$

where $\tilde{a} \in S$ is obtained from $a = (t_k\alpha_k) \cdots (t_1\alpha_1) \in G_S$ by replacing the elements $(t_i\alpha_i)$ with $t_i < 0$ by $(-t, s(\alpha_i))$. By the definition of $\tilde{p}$, the C^k-smoothness of the system (S, p, R^r) implies C^k-smoothness of $(G_S, \tilde{p}, R^r)$ and rank $\tilde{p} = $ rank p.

We define X, h, x_0 as elements of a minimal C^k representation $(X, \{\phi_a\}_{a\in G_S}, h, x_0)$ of the abstract system $(G_S, \tilde{p}, R^r)$ (we use Theorem 3). Consider elements of the group G_S of the form (t, α), $t\in R$. They form one parameter subgroup of $G_S:(t_1\alpha)(t_2\alpha)=(t_1+t_2)\alpha$ and so $\phi_{(t\alpha)}$ is a flow. From the C^k-smoothness of the representation or from the definition of the canonical C^k structure on X (see the proof of Theorem 3), it follows that $\phi_{(t\alpha)}$ is a C^k flow. We define $f(\cdot, \alpha)$ to be the vector field corresponding to the flow $\phi_{(t\alpha)}$. By the definition, we have that $\phi^f_{(t\alpha)}=\phi_{(t\alpha)}$ and so $\phi^f_a=\phi_a$ for any $a\in S$. Therefore, $p(a)=\tilde{p}(a)=h\circ\phi_a(x_0)=h\circ\phi^f_a(x_0)$ for any $a\in S$. This gives that (X, f, h, x_0) is a C^k-smooth realization of the system (S, p, R^r).

By the definition of the map $\tilde{p}$ and from (A3), we obtain that $\tilde{p}(bc_sca)=\tilde{p}(ba)$ for any $a, b\in G_s$. This and Remark 5 imply that $\phi^f_{c_sc}=id$, i.e., $\phi^f_{c_s}=(\phi^f_c)^{-1}$. In particular, $\phi^f_{(t\beta)}=(\phi^f_{(t\alpha)})^{-1}$ for $\beta=s(\alpha)$ and $t\in R$. This means that $f(\cdot, \beta)=-f(\cdot, \alpha)$, i.e., the realization is symmetric.

From the above it follows that for any $a\in G_S$ we have $\phi^f_a=\phi^f_{\tilde{a}}$, where $\tilde{a}\in S$. this fact and the minimality of the representation $(X, \{\phi_a\}_{a\in}G, h, x_0)$ give that the realization (X, f, h, x_0) is minimal.

Uniqueness. A C^k minimal realization (X, f, h, x_0) of (S, p, R^r) defines a C^k minimal representation $(X, \{\phi^f_a\}_{a\in G_S}, h, x_0)$ of the abstract system $(G_S, \tilde{p}, R^r)$, where $\phi^f_a=\phi^f_{(t_k\alpha_k)}\circ\cdots\circ\phi^f_{(t_1\alpha_1)}$ for $a=(t_k\alpha_k)\cdots(t_1\alpha_1)$; $\phi^f_{(t\alpha)}$ is a flow generated by $f(\cdot, \alpha)$ and $\tilde{p}$ is a map: $G_S\rightarrow R^r$ defined by

$$\tilde{p}(a)=h\circ\phi^f_a(x_0).$$

It is enough to prove that the abstract system, defined as above, is the same for any other minimal C^k realization (X', f', h', x_0') of (S, p, R^r). The conclusion will then follow from Theorem 3.

The equality

$$h\circ\phi^f_a(x_0)=h'\circ\phi^{f'}_a(x_0'), \qquad a\in G_S \tag{26}$$

will be proved by induction with respect to the number of negative t_i in $a=(t_k\alpha_k)\cdots(t_1\alpha_1)$. Suppose that this was proved for any $a\in G_S$ with the number of negative t_i less than or equal to s. We shall show that this implies

$$h\circ\phi^f_{c(-t\alpha)a}(x_0)=h'\circ\phi^{f'}_{c(-t\alpha)a}(x_0) \tag{27}$$

for any $t>0$ and $c\in S$. From the reachability of the first realization, there is a $b\in S$ such that

$$\phi^f_b(x_0)=\phi^f_{(-t\alpha)a}(x_0). \tag{28}$$

We also claim that

$$\phi^{f'}_b(x_0)=\phi^{f'}_{(-t\alpha)a}(x_0). \tag{29}$$

In fact, if this does not hold, then

$$\phi^{f'}_{(t\alpha)b}(x_0')=\phi^{f'}_{(t\alpha)}\circ\phi^{f'}_b(x_0')\neq\phi^{f'}_a(x_0')$$

since $\phi^{f'}_{(t\alpha)}$ is one-to-one. Therefore, by observability of the second realization there is $d\in S$ such that $h'\circ\phi^{f'}_{d(t\alpha)b}(x_0')\neq h'\circ\phi^{f'}_{da}(x_0')$. Now we can use the induction assumption for both sides of the last inequality ($d(t\alpha)b\in S$ and da has the same number of $t_i<0$ as a). We obtain that $h\circ\phi^f_{d(t\alpha)b}(x_0)\neq h\circ\phi^f_{da}(x_0)$, which contradicts (28) and proves that (29) holds.

From (28) and (29), we find that $\phi^f_{cb}(x_0) = \phi^f_{c(-t\alpha)a}(x_0)$ and $\phi^{f'}_{cb}(x'_0) = \phi^{f'}_{c(-t\alpha)a}(x'_0)$ for any $c \in S$. The latter two equalities and $h \circ \phi^f_{cb}(x_0) = h' \circ \phi^{f'}_{cb}(x'_0)$ (since $cb \in S$) imply that (27) holds. This completes the proof of (26).

B. Existence. We extend the system (S, p, R^r) to an abstract system $(G_S, \bar{p}, R^r)$, where

$$\bar{p}(a) = \psi^e_{\underline{\alpha}}(\underline{t}).$$

Here $a = (t_k\alpha_k) \cdots (t_1\alpha_1)$, $\underline{\alpha} = (\alpha_1, \cdots, \alpha_k)$ and $\underline{t} = (t_1, \cdots, t_k)$, and the function $\psi^e_{\underline{\alpha}}$ is a particular case of $\psi^{\underline{b}}_{\underline{a}}$ with $\underline{a} = \underline{\alpha}$ and $\underline{b} = e(m = 1)$.

By the fact that a restriction into a hyperplane of an analytic function is analytic and by uniqueness of analytic extensions, the map is well-defined (it agrees with the identifications $(t_1 + t_2)\alpha = (t_1\alpha)(t_2\alpha)$). The same arguments imply that the equalities

$$\psi^{b_i}_{\underline{a}}(\underline{t}) = \bar{p}(b_i(t_pa_p) \cdots (t_1a_1)) \quad \text{for } a_i, b_j \in S$$

can be extended to R^p. We define the functions $\psi^{\underline{b}}_{\underline{a}}$, with $a_i, b_j \in G_s$ by (7). The analyticity of the system implies that the rank of the abstract system $(G_S, \bar{p}, R^r)$, which is of class C^ω, is equal to the rank of (S, p, R^r). To prove this we show that every minor M of $D\psi^{\underline{b}}_{\underline{a}}(\underline{t})$ of order $n + 1$ is equal to zero, for any $\underline{a}, \underline{b} \in G_S^\infty$ and $\underline{t} \in R^\infty$. In fact, by the definition (7) of the functions $\psi^{\underline{b}}_{\underline{a}}$ and the form (4) of elements of G_S, the minor M can be expressed as a linear combination of products of expressions $(\partial/\partial\tau_i)\psi^e_{\underline{\alpha}}(\underline{\tau})$, where $\psi^e_{\underline{\alpha}}(\underline{\tau}) = p((\tau_q\alpha_q) \cdots (\tau_1\alpha_1))$. From the fact that rank of the system (S, p, R^r) is equal to n, we obtain that $M = 0$ for any $\underline{\tau} \in R^q_+$ in the expressions $(\partial/\partial\tau_i)\psi^e_{\underline{\alpha}}(\underline{\tau})$. By analyticity of the functions $\psi^e_{\underline{\alpha}}$, this implies that $M = 0$ for any value $\underline{\tau} \in R^q$, i.e., $M = 0$ for any $\underline{a}, \underline{b} \in G_S^\infty$ and $\underline{t} \in R^\infty$. We define X, h and x_0 as elements of a minimal C^ω representation $(X, \{\phi_a\}_{a \in G_s}, h, x_0)$ of the system $(G_S, \bar{p}, R^r)$ (Theorem 3). The vector fields $f(\cdot, \alpha)$, $\alpha \in \Omega$, are defined as the infinitesimal vector fields of the flows $\phi_{(t,\alpha)}$ defined by this representation. We have $\phi^f_a = \phi_a$ for any $a \in G_S$. Obviously, the realization (X, f, h, x_0) is of class C^ω and weakly reachable (this follows from the transitiveness of the representation $(X, \{\phi\}_{a \in G_S}, h, x_0)$).

To show that it is observable, take $x_1, x_2 \in X$, $x_1 \neq x_2$ and assume that $h \circ \phi^f_a(x_1) = h \circ \phi^f_a(x_2)$ for any $a \in S$. By transitiveness of the representation, there are $b_1, b_2 \in G_S$ such that $\phi_{b_1}(x_0) = x_1$ and $\phi_{b_2}(x_0) = x_2$, which gives $h \circ \phi_{ab_1}(x_0) = h \circ \phi_{ab_2}(x_0)$ for any $a \in S$. This means that

$$\psi^e_{(b_1, \alpha_1, \cdots, \alpha_k)}(1, t_1, \cdots, t_k) = \psi^e_{(b_2, \alpha_1, \cdots, \alpha_k)}(1, t_1, \cdots, t_k)$$

for any $t_j \in R_+$, $j = 1, \cdots, k$. By uniqueness of analytic extensions, this equality holds for any $t_j \in R$, $j = 1, \cdots, k$ which implies that $h \circ \phi_{ab_1}(x_0) = h \circ \phi_{ab_2}(x_0)$, i.e., $h \circ \phi_a(x_1) = h \circ \phi_a(x_2)$ for any $a \in G_S$. But this contradicts the distinguishability of our representation and so the realization is observable.

Uniqueness. Let (X, f, h, x_0) and (X', f', h', x'_0) be two minimal C^ω realizations of a system (S, p, R^r). As in the case of $k = 2, \cdots, \infty$, we show that the extended maps

$$\tilde{p}(a) = h \circ \phi^f_a(x_0), \qquad \tilde{p}'(a) = h' \circ \phi^{f'}_a(x'_0), \qquad a \in G_S$$

are equal. This follows immediately from uniqueness of analytic extensions (from R^k_+ to R^k) of the functions $h \circ \phi^f_{(t_k\alpha_k)} \circ \cdots \circ \phi^f_{(t_1\alpha_1)}(x_0)$ and $h' \circ \phi^{f'}_{(t_k\alpha_k)} \circ \cdots \circ \phi^{f'}_{(t_1\alpha_1)}(x'_0)$ as functions of the variable $\underline{t} = (t_1, \cdots, t_k)$ (they are equal on R^k_+). Therefore, we have C^ω-minimal representations $(X, \{\phi^f_a\}_{a \in G_S}, h, x_0)$ and $(X', \{\phi^{f'}_a\}_{a \in G_S}, h', x'_0)$ of the same abstract system $(G_S, \tilde{p}, R^r)$. By Theorem 3, they are C^ω-diffeomorphic and so our realizations are also C^ω-diffeomorphic. The proof is complete.

The uniqueness part of Theorem 1 can be slightly generalized as the following corollary states.

COROLLARY 4. *Let (X, f, h, x_0) and (X', f', h', x_0') be two C^k, $k = 2, 3, \cdots, \infty, \omega$, realizations of an input–output system (S, p, R^r). If they are, respectively, minimal and observable (reachable and minimal), then there is a C^k one-to-one immersion (submersion) onto $\chi : X \to X'$, such that $D\chi f = f' \circ \chi$, $h = h' \circ \chi$ and $x_0' = \chi(x_0)$.*

If $k = \omega$, then reachability can be replaced by weak reachability.

Proof. The proof is analogous to the uniqueness part of the proof of Theorem 1, where, instead of Theorem 1, we apply Corollary 2.

The dual versions of Corollaries 2 and 4 relate to the problem of quotients of manifolds (cf. Sussmann [11]). Namely, every system of complete (forward and backward) C^k, $k = \infty, \omega$, vector fields $\{f(\cdot, \alpha)\}_{\alpha \in \Omega}$ on a C^k manifold X and a finite system of real valued functions $h_1, \cdots, h_r$ of class C^k define an equivalence relation of "indistinquishability" on X by

$$x_1 \sim x_2 \Leftrightarrow \forall b \in G_S, \qquad h_i \circ \phi_b^f(x_1) = h_i \circ \phi_b^f(x_2), \qquad i = 1, \cdots, r.$$

COROLLARY 5. *If $\{f(\cdot, \alpha)\}_{\alpha \in \Omega}$ is a transitive system of vector fields (i.e., $\forall x_1, x_2, \exists b \in G_S, \phi_b(x_1) = x_2$) and X is connected, then the equivalence relation is regular, i.e., the quotient $X/\sim$ admits a (unique) structure of a C^k manifold such that the canonical projection $X \to X/\sim$ is a C^k submersion.*

Proof. Let x_0 be any point of X and denote $h = (h_1, \cdots, h_r)$. The quadruple $(X, \{\phi_a^f\}_{a \in G_S}, h, x_0)$ is a C^k transitive representation of its own abstract input–output system (G_S, p, R^r), where $p(a) = h \circ \phi_a^f(x_0)$. By Theorem 3, this system has a minimal, C^k representation $(X', \{\phi_a'\}_{a \in G_S}, h', x_0')$. From Corollary 2, there is a C^k submersion $\chi : X \to X'$ such that $\phi_a' \circ \chi = \chi \circ \phi_a^f$ and $h = h' \circ \chi$. It is easy to see that if $\chi(x_1) = \chi(x_2)$, then $x_1 \sim x_2$. The converse is also true, which follows from the distinguishability of the second representation. Therefore, χ factorizes through a canonical map $\tilde{\chi} : X/\sim \to X'$, which is a bijection and carries the structure of a C^k manifold of X' onto $X/\sim$. The proof is complete.

The solution of the existence problem given by Theorem 1 can be regarded as complete only in the analytic case. In the case $2 \leqq k \leqq \infty$ our condition is necessary for the existence of symmetric realizations only. A necessary and sufficient condition for the existence of any C^k realizations is given by the following theorem (it is an easy consequence of Theorem 3).

THEOREM 4. *Let $k = 2, \cdots, \infty, \omega$. The input–output system (S, p, R^r) has a C^k realization if and only if the map p can be extended to a map $\bar{p} : G_S \to R^r$, which satisfies* (A1) *and* (A2) *with S replaced by G_S, and R_+ replaced by R.*

Proof. Suppose the map can be extended to a map $\bar{p} : G_S \to R^r$, which satisfies (A1) and (A2). Thus the triple $(G_S, \bar{p}, R^r)$ is a C^k abstract system. By Theorem 3, it has a C^k representation $(X, \{\phi_a\}_{a \in G_S}, h, x_0)$. This representation defines a C^k realization (X, f, h, x_0) of the system (S, p, R^r), where $f(\cdot, \alpha)$ is the infinitesimal vector field of the flow $\phi_{(t\alpha)}$.

Conversely, if the quadruple (X, f, h, x_0) is a C^k realization of the system (S, p, R^r), then the map p can be naturally extended to G_S by the formula

$$\bar{p}(a) = h \circ \phi_a^f(x_0) = h \circ \phi_{(t_k \alpha_k)}^f \circ \cdots \circ \phi_{(t_1 \alpha_1)}^f,$$

where $a = (t_k \alpha_k) \cdots (t_1 \alpha_1)$, and $\phi_{(t\alpha)}^f$ is the C^k flow defined by $f(\cdot, \alpha)$. Then condition (A1) follows from the definition of a C^k realization, and condition (A2) can be proved as in Remark 1.

7. Proof of Theorem 2.

A. Existence. Analogously, as in the proof of Theorem 1, we extend our system $(\tilde{S}, p, R^r)$ to an abstract system $(G_{\tilde{S}}, \tilde{p}, R^r)$. The extended map $\tilde{p}: G_{\tilde{S}} \to R^r$ is defined by

$$\tilde{p}(u) = p(\tilde{u}),$$

where $\tilde{u} \in \tilde{S}$ is obtained from $u \in G_{\tilde{S}}$ by replacing the elements $(t_i\varphi_i)$ with $t_i < 0$ by $(-t_i\varphi_i^*)$, where for $\varphi \in \mathcal{U}$, $\varphi^*(\tau) = s(\varphi(-\tau))$. By the assumptions on the system $(\tilde{S}, p, R^r)$ the abstract system $(G_{\tilde{S}}, \tilde{p}, R^r)$ is of class C^k and rank $\tilde{p} =$ rank p. Let $(X, \{\phi\}_{u \in G_{\tilde{S}}}, h, x_0)$ be a C^k, minimal representation of this system, which exists by Theorem 3. We define a realization of the system $(\tilde{S}, p, R^r)$ as a quadruple (X, f, h, x_0), where $f(\cdot, \alpha)$ is defined as a vector field corresponding to the flow $\phi_{(t\alpha)}$.

Before showing that the quadruple (X, f, h, x_0) has the desired properties, let us prove that $\phi_u = \phi_{\tilde{u}}$ for any $u \in G_{\tilde{S}}$, where $\phi_u, \phi_{\tilde{u}}$ are diffeomorphisms from the representation $(X, \{\phi_u\}_{u \in G_{\tilde{S}}}, h, x_0)$. By the definition of the element $\tilde{u} \in \tilde{S}$, it is enough to show that $\phi_{(t\varphi)} = \phi_{(-t\varphi^*)}$ for $t < 0$ and $\varphi \in \mathcal{U}$.

But

$$\phi_{(-t\varphi^*)} = \phi_{(t\varphi^*)^{-1}} = \overset{-1}{\phi}_{(t\varphi^*)} = \overset{-1}{\phi}_{(t\varphi)_s},$$

where $(t\varphi)_s$ is defined as in (A3). By (A3) and Remark 5, we find that $\phi_{u_s u} = \phi_{u u_s} = id$, i.e., $\phi_{u_s} = \overset{-1}{\phi}_u$. Therefore, $\overset{-1}{\phi}_{(t\varphi)_s} = \phi_{(t\varphi)}$ and so $\phi_{(-t\varphi^*)} = \phi_{(t\varphi)}$.

Now we shall prove that $f(x, \alpha)$ is continuous with respect to (x, α), together with the first derivative with respect to x. We fix x and show this locally around x. Take $\underline{u}, \underline{v} \in \tilde{S}^\infty$ and $\underline{t} \in R_+^\infty$ such that rank $D\psi_{\underline{u}}^{\underline{v}}(\underline{t}) = \dim X = n$. We have $\psi_{\underline{u}}^{\underline{v}} = \psi^{\underline{v}} \circ \psi_{\underline{u}}$, where $\psi_{\underline{u}}: R_+^p \to X$, $\psi^{\underline{v}}: X \to R^{rm}$ are defined by $\psi_{\underline{u}}(\underline{t}) = \phi_{(t_p u_p) \cdots (t_1 u_1)}(x_0)$ and $\psi^{\underline{v}}(x) = (h \circ \phi_{v_1}(x), \cdots, h \circ \phi_{v_m}(x))$. The sequences $\underline{u}, \underline{v}, \underline{t}$ may be chosen such that $\psi_{\underline{u}}(\underline{t}) = x$. In fact, if $\psi_{\underline{u}}(\underline{t}) = x_1 \neq x$, then there is $u \in G_{\tilde{S}}$ such that $\phi_u(x_1) = x$. By the fact that $\phi_u = \phi_{\tilde{u}}$, we have also $\phi_{\tilde{u}}(x_1) = x$, where $\tilde{u} \in \tilde{S}$. Take $v = \tilde{u}^{-1}$ and note that $\phi_{\tilde{v}} = \phi_v = \overset{-1}{\phi}_{\tilde{u}}$. We define new $\underline{u}, \underline{v}, \underline{t}$ by $\underline{u}' = (u_1, \cdots, u_p, \tilde{u})$, $\underline{t} = (t_1, \cdots, t_p, 1)$, $\underline{v}' = (v_1\tilde{v}, \cdots, v_m\tilde{v})$ and achieve the desired property.

The map $\psi^{\underline{v}}: X \to R^{rm}$ may be treated as a substitute of local coordinates in a neighborhood of the point x. In fact, it is C^k-smooth (by C^k-smoothness of the representation); and, by rank $D\psi_{\underline{u}}^{\underline{v}}(\underline{t}) = n$, it is a local diffeomorphism into an n-dimensional submanifold of R^{rm}. The map $\psi_{\underline{u}}: R_+^p \to X$ can be treated (locally around $\underline{t}$) as a local chart, after reducing the number of variables to n (fixing $p - n$ variables) as in the proof that X is n-manifold. We denote the reduced function by $\tilde{\psi}_{\underline{u}}(\underline{\tau})$. Define the function

$$g(\underline{\tau}, \alpha) = \frac{d}{dt}(\psi^{\underline{v}} \circ \psi_{(t\alpha)} \circ \tilde{\psi}_{\underline{u}}(\underline{\tau}))/_{t=0} = D\psi^{\underline{v}}(y) \cdot f(y, \alpha),$$

where $y = \tilde{\psi}_{\underline{u}}(\underline{\tau})$. From the assumption (A4b) it follows that $\tilde{g}$ is continuous with respect to $(\underline{\tau}, \alpha)$, together with the first derivative with respect to $\underline{\tau}$. This and the fact that $\underline{\tau} = \tilde{\psi}_{\underline{u}}^{-1}(y)$ can be treated as local coordinates around $x \in X$ give that $f(y, \alpha)$ is continuous with respect to (y, α), together with the first derivative with respect to y (note that $D\psi^{\underline{v}}(y)$ is invertible).

The diffeomorphisms ϕ_u, $u \in \tilde{S}$ defined by the representation and the diffeomorphisms ϕ_u^f defined by the equation $\dot{x} = f(x, u)$ coincide. For $u \in S$, this follows from the definition of $f(x, \alpha)$. An analogous fact for $u \in \tilde{S}$ follows from the continuity assumption.

In fact, let $\phi^f_u(x) \neq \phi_u(x)$ for some $u \in \tilde{S}$ and $x \in X$. By the minimality of the representation, there are $v, w \in G_{\tilde{S}}$ such that $x = \phi_w(x_0)$ and

$$h \circ \phi_v \circ \phi^f_u(x) \neq h \circ \phi_v \circ \phi_u \circ \phi_w(x_0) = \tilde{p}(vuw).$$

Since $\tilde{p}(vuw) = p(\tilde{v}u\tilde{w})$ with $\tilde{v}, \tilde{w} \in \tilde{S}$, thus we obtain

$$(h \circ \phi_v)(\phi^f_u(x)) \neq p(\tilde{v}u\tilde{w}). \tag{30}$$

Now, we approximate the function u by a piecewise constant function. By the approximation property and the assumption (A4a), we find that $h \circ \phi_v \circ \phi^f_a(x) \neq p(\tilde{v}a\tilde{w}) = \tilde{p}(vaw)$ for some $a \in S$. This means that $h \circ \phi_v \circ \phi^f_a \neq h \circ \phi_v \circ \phi_a$, which contradicts the equality $\phi^f_a = \phi_a$ and proves that $\phi^f_u = \phi_u$ for any $u \in \tilde{S}$.

Having $\phi^f_u = \phi_u$ for $u \in \tilde{S}$, we conclude the proof easily. The equality $h \circ \phi^f_u(x_0) = p(u), u \in \tilde{S}$ follows from the corresponding equality for the representation $(X, \{\phi_u\}_{u \in G_{\tilde{S}}}, h, x_0)$. The C^k-smoothness of the realization (X, f, h, x_0) follows easily from the C^k-smoothness of the representation. Minimality of (X, f, h, x_0) is a consequence of the minimality of the representation and the fact that for any $u \in G_{\tilde{S}}$, there is $\tilde{u} \in \tilde{S}$ such that $\phi_u = \phi_{\tilde{u}}$. Symmetry of the realization follows from the fact that $\phi^f_{(-t\alpha)} = \phi^f_{(t\beta)}$ for $t \in R_+$ and $\beta = s(\alpha)$ (note that $(\widetilde{-t\alpha}) = (t\beta)$).

Uniqueness. The proof is analogous to part A of the proof of Theorem 1, where $\alpha \in \Omega$ should be replaced by $\varphi \in \mathscr{U}$, S by $\tilde{S}$, and G_S by $G_{\tilde{S}}$.

B. Existence. We define an extension of the system $(\tilde{S}, p, R^r)$ as a triple $(G_{\tilde{S}}, \bar{p}, R^r)$, where the map $\bar{p}: G_{\tilde{S}} \to R^r$ is the extension of p defined in § 3. Clearly, $(G_{\tilde{S}}, \bar{p}, R^r)$ is an abstract system of class C^ω and rank $\bar{p}$ = rank p (the latter can be showed analogously as in part B of the proof of Theorem 1).

Let $(X, \{\phi_u\}_{u \in G_{\tilde{S}}}, h, x_0)$ be a minimal C^ω representation of the system $(G_{\tilde{S}}, \bar{p}, R^r)$ (it exists by Theorem 3). We define a realization of the system $(\tilde{S}, p, R^r)$ as a quadruple (X, f, h, x_0), where $f(\cdot, \alpha)$ is defined as the vector field corresponding to the flow $\phi_{(t\alpha)}$.

The proof that the realization is continuous is analogous to that in part A of this proof. The only modification is that we need only $\underline{u}, \underline{v} \in G^\infty_{\tilde{S}}$ (except $\underline{u}, \underline{v} \in S^\infty$), because the assumption (A4b) holds on the group $G_{\tilde{S}}$. Therefore, we take $\underline{u}' = (u_1, \cdots, u_p, u)$, $\underline{t}' = (t_1, \cdots, t_p, 1)$ and $\underline{v}' = (v_1u^{-1}, \cdots, v_mu^{-1})$. The proof that $\phi_u = \phi^f_u$ for $u \in \tilde{S}$ is also a modification of part A of this proof (replacing inequality (30) by $(h \circ \phi_v)(\phi^f_u(x)) \neq \bar{p}(vuw)$ with $v, w \in G_{\tilde{S}}$, and using (A4a) for the approximation of u by $a \in S$). It also follows that $\phi_u = \phi^f_u$, for $u \in G_{\tilde{S}}$ by the fact that $\phi_{v^{-1}} = \phi_v^{-1}$ and $\phi^f_{v^{-1}} = (\phi^f_v)^{-1}$ for $v \in \tilde{S}$.

Now from the fact that $(X, \{\phi_u\}_{u \in G_{\tilde{S}}}, h, x_0)$ is a C^ω minimal representation of $(G_{\tilde{S}}, \bar{p}, R^r)$, we obtain easily that (X, f, h, x_0) is a C^ω-minimal realization of the system $(\tilde{S}, p, R^r)$. (Observability follows analogously as in part B of the proof of Theorem 1).

Uniqueness. The proof is analogous to the proof of Theorem 1, where $\alpha_i \in \Omega$ should be replaced by $\varphi_i \in \mathscr{U}$, and $a \in G_S$ by $u \in G_{\tilde{S}}$.

REFERENCES

[1] R. W. BROCKETT, *System theory on group manifolds and coset spaces*, this Journal, 10 (1972), pp. 265–284.

[2] P. D'ALESSANDRO, A. ISIDORI AND A. RUBERTI, *Realization and structure theory of bilinear dynamical systems*, this Journal, 12 (1974), pp. 517–535.

[3] B. JAKUBCZYK, *Existence and uniqueness of nonlinear realizations*, Proc. Conf. Analyse des Systèmes, Bordeaux 1978, Asterisque, to appear.

[4] R. E. Kalman, P. L. Falb and M. A. Arbib, *Topics in Mathematical System Theory*, McGraw-Hill, New York, 1969.

[5] R. Hermann and A. J. Krener, *Nonlinear controllability and observability*, IEEE Trans. Automatic control, AC-22 (1977), pp. 728–740.

[6] C. Lobry, *Dynamical polysystems and control theory*, Geometric Methods in System Theory, D. Q. Mayne and R. W. Brockett, eds., D. Riedel, Dordrecht, Holland, 1973.

[7] S. Sternberg, *Lectures on Differential Geometry*, Prentice-Hall, Englewood Cliffs, NJ, 1964.

[8] H. J. Sussmann, *Minimal realizations of nonlinear systems*, Geometric Methods in Systems Theory, D. Q. Mayne and R. W. Brockett, eds., D. Riedel, Dordrecht, Holland, 1973.

[9] ———, *Orbits of families of vector fields and integrability of distributions*, Trans. Amer. Math. Soc., 180 (1973), pp. 171–188.

[10] ———, *Existence and uniqueness of minimal realizations of nonlinear systems*, Math. Systems Theory, 10 (1977), pp. 263–284.

[11] ———, *A generalization of closed subgroup theorem to quotients of arbitrary manifolds*, J. Differential Geometry, 10 (1975), pp. 151–166.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 5, September 1980

CONVERGENCE RATES FOR CONDITIONAL GRADIENT SEQUENCES GENERATED BY IMPLICIT STEP LENGTH RULES*

J. C. DUNN†

Abstract. Conditional gradient algorithms with implicit line minimization and Goldstein–Armijo step length rules are considered for the problem $\min_\Omega F$ with Ω a bounded convex subset of a real Banach space. When the Fréchet derivative F' is uniformly continuous on Ω, the iterates x_n generated by any of the algorithms comprise an "extremizing" sequence in the sense that the quantity, $\langle F'(x_n), x_n\rangle - \inf_{y\in\Omega}\langle F'(x_n), y\rangle$, converges to zero as $n\to\infty$. This ensures that every limit point of $\{x_n\}$ is an extremal, and for compact Ω it then follows that $\{x_n\}$ converges to the set of extremals in Ω. Weak counterparts of these results are also established. Convergence rate estimates are derived for convex F and Lipschitz continuous F'. These estimates are closely related to results obtained in an earlier investigation of two explicit step length formulas for conditional gradient methods. Once again, the growth rate of the function $a(\sigma) = \inf\{\rho = \langle F'(\xi), x-\xi\rangle \mid x\in\Omega, \|x-\xi\|\geqq\sigma\}$ at an extremal ξ, determines how rapidly the functional values $F(x_n)$ converge to $\inf_\Omega F$.

1. Introduction. This paper continues an analysis of conditional gradient algorithms for the problem

$$F(\xi) = \inf_{x\in\Omega} F(x) \tag{1.1}$$

with Ω a bounded convex subset of a Banach space X, and $F\colon X\to\mathbb{R}^1$ a real functional with a continuous Fréchet derivative F'. In [1] it was shown that the behavior of conditional gradient sequences $\{x_n\}$ generated by either of two simple explicit step length formulas is sensitive to the rate at which the function

$$a(\sigma) = \inf_{\substack{x\in\Omega\\ \|x-\xi\|\geqq\sigma}} \langle F'(\xi), x-\xi\rangle \tag{1.2}$$

grows with increasing $\sigma>0$. If F is convex and F' is Lipschitz continuous then for both of the algorithms in question, the functional values $F_n = F(x_n)$ converge to $\inf_\Omega F$ like a term of order $o(1/n)$, geometrically, or in finitely many steps, depending on whether

$$a(\sigma)>0 \quad \text{for } \sigma>0 \tag{1.3}$$

or

$$\exists A>0,\ a(\sigma)\geqq A\sigma^2 \tag{1.4}$$

or

$$\exists A_s>0,\ a(\sigma)\geqq A_s\sigma. \tag{1.5}$$

In the general case where (1.3) is not satisfied, one still has $F_n - \inf_\Omega F = O(1/n)$. An extremal ξ is said to be strongly nonsingular, regular, or strongly regular (relative to F and the norm on X) according to whether (1.3), (1.4) or (1.5) holds. The geometric content of these conditions is developed at length in [1], along with their relationship to classical notions of nonsingularity in optimal control theory.

It is established below that the same growth conditions (1.3)–(1.5) also control the convergence of conditional gradient sequences generated by the implicit line minimization step size rule and two other related implicit schemes originally proposed by

* Received by the editors September 29, 1978, and in revised form June 11, 1979.

† Department of Mathematics, North Carolina State University, Raleigh, North Carolina 27650. This investigation supported by National Science Foundation Grant ENG 78-03385.

Goldstein [2], [3] and Armijo [4]. Moreover, the convergence rate estimates obtained here for the implicit rules are related in an interesting way to estimates derived in [1] for simple *explicit* step length formulas requiring no line search.[1] In particular, when the Lipschitz norm of F' is known, the best a priori convergence rate estimates for the explicit formulas are identical to the sharpest estimates established in § 3 for exact line minimization, and for limiting cases of the Goldstein–Armijo rules.

Less can be said here about the line minimization and Goldstein–Armijo rules when F is not convex or F' is not Lipschitz continuous. In all cases, the descent property holds (i.e., $F_n \geqq F_{n+1}$); consequently, if F is bounded below on Ω, $\{F_n\}$ always converges downward to some limit $l \geqq \inf_\Omega F > -\infty$. Moreover, if F' is uniformly continuous on Ω, then strong limit points (and sometimes weak limit points) of $\{x_n\}$ in Ω are extremals, and this leads to nontrivial convergence theorems for $\{x_n\}$ under compactness assumptions on Ω. The concept of nonsingularity introduced in [1] is also important in this analysis.

2. The algorithms. As in [1], let $T\colon \Omega \to 2^\Omega$ denote the set-valued operator defined by

$$T(x) = \{\bar{x} \in \Omega \mid \langle F'(x), \bar{x}\rangle = \inf_{y\in\Omega} \langle F'(x), y\rangle\} \tag{2.1}$$

as x ranges over $\Omega \subset X$. By definition, $\{x_n\} \subset \Omega$ is called a *conditional gradient sequence* if and only if there exist corresponding sequences $\{\bar{x}_n\} \subset \Omega$ and $\{\omega_n\} \subset [0, 1]$ such that

$$x_{n+1} = x_n + \omega_n(\bar{x}_n - x_n), \qquad \bar{x}_n \in T(x_n). \tag{2.2}$$

Different versions of the conditional gradient method are obtained from each of several different rules which relate the step length parameter ω_n to n, x_n and $\bar{x}_n$ at each stage.

In the classical line minimization step length scheme, the *implicit* constraint,

$$\omega_n = 0 \quad \text{if } \langle F'_n, x_n - \bar{x}_n\rangle = 0, \tag{2.3A}$$

$$F(x_{n+1}) = \min_{0\leqq\omega\leqq1} F(x_n + \omega(\bar{x}_n - x_n)) \quad \text{if } \langle F'_n, x_n - \bar{x}_n\rangle > 0 \tag{2.3B}$$

is imposed along with (2.2) at each n; at the very least, this condition ensures the descent property, $F_n \geqq F_{n+1}$, guarantees that $\{x_n\}$ terminates at x_N if x_N is an extremal (i.e., a fixed point of T), and is known to force convergence of F_n to $\inf_\Omega F$ under suitable convexity and smoothness conditions on Ω and F [5]. Other more tractable implicit step length constraints have been proposed for feasible direction methods by Goldstein [2], [3] and Armijo [4]. In the present context, Goldstein's rule first fixes a δ in $(0, \frac{1}{2}]$ and for $\omega > 0$ puts

$$g(x, \bar{x}; \omega) = \frac{F(x) - F(x + \omega(\bar{x} - x))}{\omega\langle F'(x), x - \bar{x}\rangle},$$

wherever $\langle F'(x), x - \bar{x}\rangle \neq 0$, i.e., wherever x is not an extremal for F in Ω. If F' is continuous, then g is continuous in ω with x and $\bar{x}$ fixed, and a straightforward application of the mean value theorem shows that $\lim_{\omega\to0^+} g(x, \bar{x}; \omega) = 1$. Consequently if $g(x, \bar{x}; 1) < \delta$, it follows at once from the intermediate value theorem that the set

$$W_\delta(x, \bar{x}) = \{\omega \in (0, 1] \mid \delta \leqq g(x, \bar{x}; \omega) \leqq 1 - \delta\}$$

[1] A referee for [1] guessed that Goldstein–Armijo rules would generate geometrically convergent gradient sequences whenever the explicit step length formulas in [1] produce such sequences. The investigation reported here was prompted by that conjecture.

is not empty, in which case the corresponding set

$$\text{(2.4A)} \qquad \bar{W}_\delta(x,\bar{x}) = \begin{cases} \{0\} & \text{if } \langle F'(x), x-\bar{x}\rangle = 0, \\ \{1\} & \text{if } \langle F'(x), x-\bar{x}\rangle > 0 \text{ and } g(x,\bar{x};1) \geqq \delta, \\ W_\delta(x,\bar{x}) & \text{if } \langle F'(x), x-\bar{x}\rangle > 0 \text{ and } g(x,\bar{x};1) < \delta \end{cases}$$

is also not empty. Goldstein's rule now requires that

$$\text{(2.4B)} \qquad \omega_n \in \bar{W}_\delta(x_n, \bar{x}_n)$$

at each $n \geqq 1$. In the closely related Armijo scheme, two parameters δ and β are fixed in $(0, \frac{1}{2}]$ and $(0, 1)$ respectively, and ω_n is determined as follows:

$$\text{(2.5A)} \qquad \omega_n = 0 \quad \text{if } \langle F'_n, x_n - \bar{x}_n\rangle = 0,$$

or

$$\text{(2.5B)} \qquad \omega_n = \beta^{m_n} \quad \text{if } \langle F'_n, x_n - \bar{x}_n\rangle > 0,$$

where

$$\text{(2.5C)} \qquad m_n = \min\{m \geqq 0 \,|\, g(x_n, \bar{x}_n; \beta^m) \geqq \delta\}.$$

The existence of m_n is assured by the fact that $\lim_{\omega\to 0^+} g(x, \bar{x}_n; \omega) = 1$.

Conditions (2.4) and (2.5) impose the descent property $F_n \geqq F_{n+1}$, ensure that $\{x_n\}$ terminates at x_N if x_N is an extremal, and otherwise guarantee that the decrement $F_{n+1} - F_n$ is a "sufficiently large" fraction of the leading linear term in Taylor's formula.

$$\text{(2.6)} \qquad F_{n+1} - F_n = \omega_n \langle F'_n, \bar{x}_n - x_n\rangle + \int_0^{\omega_n} \langle F'(x_n + \sigma(\bar{x}_n - x_n)) - F'_n, \bar{x}_n - x_n\rangle \, d\sigma$$

with $\omega_n \neq 0$. This last observation gains significance when one considers that, for quadratic F, (2.4) and (2.5) "approach" (2.3) in the limit as $\delta \to \frac{1}{2}$ and $\beta \to 1$. Even for nonquadratic F, one still tends to think of the Goldstein–Armijo rules as approximate line minimization schemes. This rough interpretation is at least partially vindicated by the a priori convergence rate estimates of Theorems 3.1 and 3.2 in the next section.

Note 2.1. To conform with Armijo's original scheme in [4], one would put $\delta = \beta = \frac{1}{2}$ in (2.5). The slightly more general treatment given here is similar to Polak's development in [6].

Note 2.2. For $\delta \in (0, \frac{1}{2})$, it is always possible to find an $\omega_n \in \bar{W}_\delta(x_n, \bar{x}_n)$ in (2.4) with simple line search algorithms (e.g., iterated bi-section) requiring only finitely many evaluations of F. Similarly, for $\delta \in (0, \frac{1}{2}]$ and $\beta \in (0, 1)$, the implementation of (2.5) requires just finitely many evaluations of F. However, the *number* of such evaluations can be expected to increase without bound as $\delta \to \frac{1}{2}$ in (2.4), or as $\beta \to 1$ in (2.5). Consequently, these limiting cases, like exact line minimization, are primarily of theoretical interest. For a further discussion of various step size algorithms within the general framework of feasible direction methods, see [7].

3. Convergence rates for convex functionals. The following result is a slightly revised version of Lemma 5.1 in [1].

LEMMA 3.1. *Let β_n and q_n satisfy*

$$\text{(3.1)} \qquad \begin{aligned} &0 < \beta_{n+1} \leqq \beta_n - q_n \beta_n^2; \qquad \beta_1 = 1, \\ &0 < q \leqq q_n \end{aligned}$$

for $1 \leqq n < N$. *Then*

$$0 < \beta_n \leqq \frac{1}{1+q(n-1)}$$

for $1 \leqq n \leqq N$. *Furthermore, if* (3.1) *holds for all* $n \geqq 1$, *and if* $\lim_{n\to\infty} q_n = \infty$, *then* $\beta_n = o(1/n)$.

Proof. In [1].

THEOREM 3.1. *Let* Ω *be a bounded convex subset of a real Banach space* X *and let* $F: X \to \mathbb{R}^1$ *be convex and Fréchet differentiable. Furthermore, suppose that* F' *is Lipschitz continuous on* Ω, *with Lipschitz norm*

$$L_1 = \sup_{\substack{x,y\in\Omega\\ x\neq y}} \frac{\|F'(x)-F'(y)\|}{\|x-y\|} < \infty. \tag{3.2}$$

Finally, suppose that the sequences $\{x_n\} \subset \Omega$, $\{\bar{x}_n\} \subset \Omega$, *and* $\{\omega_n\} \subset [0, 1]$ *satisfy* (2.2) *and the line minimization rule* (2.3) *at each* $n \geqq 1$, *and put*

$$r_n = F_n - \inf_{\Omega} F.$$

Then:

(i) $\{r_n\}$ *decreases monotonically to* 0, *with*

$$0 \leqq r_n \leqq \frac{r_1}{1+q(n-1)} \tag{3.3A}$$

for all $n \geqq 1$, *where*

$$q = \frac{1}{2}\min\left\{1, \frac{r_1}{L_1 D^2}\right\} > 0 \tag{3.3B}$$

and $D = \operatorname{diam} \Omega$.

(ii) *When* F *has a minimizer* ξ *satisfying the strong nonsingularity condition* (1.3), ξ *is necessarily a unique minimizer, the sequences* $\{x_n\}$ *and* $\{\bar{x}_n\}$ *converge strongly to* ξ, *and* $r_n = o(1/n)$.

(iii) *When* ξ *satisfies the regularity condition* (1.4), $\{r_n\}$ *converges geometrically to* 0, *with*

$$0 \leqq r_n \leqq r_1 \lambda^n \tag{3.4A}$$

for all $n \geqq 1$, *where*

$$\lambda = \max\left\{\frac{1}{2}, 1 - \frac{1}{2(L_1/A)(1+L_1/A)^2}\right\} \in \left[\frac{1}{2}, 1\right). \tag{3.4B}$$

Moreover, $\|x_n - \xi\| = O(\lambda^{n/2})$ *and* $\|\bar{x}_n - \xi\| = O(\lambda^{n/2})$.

(iv) *When* ξ *satisfies the strong regularity condition* (1.5), *there is an integer* N_0 *such that*

$$x_n = \bar{x}_n = \xi \tag{3.5A}$$

for all $n > N_0$. *Moreover,*

$$N_0 \leqq 1 - \frac{\log(L_1 r_1/A_s^2)}{\log(1/\lambda_s)}, \tag{3.5B}$$

where

$$\lambda_s = \max\left\{\frac{1}{2},\ 1 - \frac{1}{2(L_1D/A_s)(1+L_1D/A_s)^2}\right\} \in \left[\frac{1}{2},\ 1\right). \tag{3.5C}$$

Proof. The proof closely follows the pattern established in [1]. (See Note 3.1, below.) Under the stated hypotheses, $\inf F > -\infty$ and $r_n \geqq r_{n+1} \geqq 0$ for $n \geqq 1$. In all cases (i)–(iv), $\langle F'_{N+1}, x_{N+1} - \bar{x}_{N+1}\rangle = 0 \Rightarrow x_{N+1}$ is an extremal (and therefore a minimizer of F) $\Rightarrow x_n = x_{N+1}$ and $r_n = 0$ for $n \geqq N+1$. On the other hand, if

$$\langle F'_n, x_n - \bar{x}_n\rangle > 0 \tag{3.6}$$

for $1 \leqq n \leqq N$, then $r_n > 0$ and $\|\bar{x}_n - x_n\| > 0$ for n in this range, in which case (2.6) and (3.2) give

$$\begin{aligned} r_{n+1} - r_n &= \min_{0\leqq\omega\leqq1} \{F(x_n + \omega(\bar{x}_n - x_n)) - F(x_n)\} \\ &= \min_{0\leqq\omega\leqq1} \left\{\omega\langle F'_n, \bar{x}_n - x_n\rangle + \int_0^\omega \langle F'(x_n + \sigma(\bar{x}_n - x_n)) - F'_n, \bar{x}_n - x_n\rangle\, d\sigma\right\} \\ &\leqq \min_{0\leqq\omega\leqq1} \{\omega\langle F'_n, \bar{x}_n - x_n\rangle + \tfrac{1}{2}\omega^2 L_1\|\bar{x}_n - x_n\|^2\} \\ &\leqq -\frac{1}{2}\min\left\{\frac{\langle F'_n, \bar{x}_n - x_n\rangle^2}{L_1\|\bar{x}_n - x_n\|^2},\ \langle F'_n, x_n - \bar{x}_n\rangle\right\} \end{aligned} \tag{3.7}$$

for $1 \leqq n \leqq N$. Observe now that for any $\varepsilon > 0$ there is a $y \in \Omega$ such that $F(y) < \inf_\Omega F + \varepsilon$; hence it follows from (2.1) and the convexity of F that

$$\begin{aligned} \langle F'_n, x_n - \bar{x}_n\rangle &= \langle F'_n, x_n - y\rangle + \langle F'_n, y - \bar{x}_n\rangle \\ &\geqq \langle F'_n, x_n - y\rangle \\ &\geqq F_n - F(y) \\ &\geqq r_n - \varepsilon \end{aligned}$$

for $n \geqq 1$. Since ε can be arbitrarily small here, this means that

$$\langle F'_n, x_n - \bar{x}_n\rangle \geqq r_n \tag{3.8}$$

for $n \geqq 1$. Together, (3.7) and (3.8) yield the fundamental inequality

$$0 < (r_{n+1}/r_1) \leqq (r_n/r_1) - q_n(r_n/r_1)^2 \tag{3.9A}$$

for $1 \leqq n < N$, with

$$q_n = \frac{r_1}{2}\min\left\{\frac{1}{L_1\|\bar{x}_n - x_n\|^2},\ \frac{1}{r_n}\right\}. \tag{3.9B}$$

Since $\operatorname{diam}\Omega = D < \infty$, it follows from (3.9B) that

$$q_n \geqq q = \frac{1}{2}\min\left\{\frac{r_1}{L_1D^2},\ 1\right\} > 0. \tag{3.10}$$

Lemma 3.1 now establishes (i).

If (1.3) holds at a minimizer ξ it follows from (i) and from Lemma 5.2 in [1] that ξ is the only minimizer of F, and x_n converges strongly to ξ. Moreover, the operator T in (2.1) satisfies the continuity condition of Theorem 3.2 in [1] and therefore $\bar{x}_n$ converges strongly to ξ, in which case $\lim_{n\to\infty}\|\bar{x}_n - x_n\| = 0$. If (3.6) holds for all $n \geqq 1$, then (3.9) also holds for $n \geqq 1$, with $\lim_{n\to\infty} q_n = \infty$. Part (ii) now follows from Lemma 3.1.

If (1.4) holds at ξ, then T satisfies the Lipschitz continuity condition of Theorem 3.6 in [1], and therefore

$$\|\bar{x}_n - \xi\| \leqq (L_1/A)\|x_n - \xi\| \tag{3.11}$$

for $n \geqq 1$. The triangle inequality now yields

$$\|\bar{x}_n - x_n\|^2 \leqq (1 + L_1/A)^2\|x_n - \xi\|^2. \tag{3.12}$$

Moreover, since F is convex,

$$r_n = F_n - F(\xi) \geqq \langle F'(\xi), x_n - \xi\rangle,$$

and therefore, by (1.4),

$$r_n \geqq A\|x_n - \xi\|^2. \tag{3.13}$$

Consequently, (3.9) and (3.12) give

$$(r_{n+1}/r_1) \leqq \left(1 - \frac{1}{2}\min\left\{1, \frac{1}{(L_1/A)(1+L_1/A)^2}\right\}\right)(r_n/r_1)$$

for $n \geqq 1$. Finally, it follows from (3.11) and (3.13) that

$$\|x_n - \xi\| \leqq \left(\frac{r_n}{A}\right)^{1/2}$$

and

$$\|\bar{x}_n - \xi\| \leqq \left(\frac{L_1^2}{A^3} r_n\right)^{1/2}.$$

This completes the proof of (iii).

Condition (1.5) implies (1.4) with $A = A_s/D$; consequently (iii) gives

$$0 \leqq r_n \leqq r_1 \lambda_s^n$$

for $n \geqq 1$, with λ_s in (3.5C). It now follows from (3.13) and (1.5) that

$$0 \leqq \|x_n - \xi\| \leqq (r_1/A_s)\lambda_s^n$$

for $n \geqq 1$. Furthermore, one has

$$\begin{aligned} L_1\|x_n - \xi\|\,\|\bar{x}_n - \xi\| &\geqq \langle F'_n - F'(\xi), \xi - \bar{x}_n\rangle \\ &\geqq \langle F'(\xi), \bar{x}_n - \xi\rangle \\ &\geqq A_s\|\bar{x}_n - \xi\|, \end{aligned} \tag{3.14}$$

and therefore

$$(L_1 r_1/A_s^2)\lambda_s^n\|\bar{x}_n - \xi\| \geqq \|\bar{x}_n - \xi\| \tag{3.15}$$

for $n \geqq 1$. If $n > \log(L_1 r_1/A_s^2)/\log(1/\lambda_s)$, then $(L_1 r_1/A_s^2)\lambda_s^n < 1$ and (3.15) implies that $\bar{x}_n = \xi$. Finally, ξ is the only minimizer of F over Ω; hence it follows from (2.3) that $x_k = \bar{x}_k = \xi$ for all $k > n$. This proves (iv). Q.E.D.

THEOREM 3.2. *Let Ω and F satisfy the hypotheses of Theorem* 3.1. *Suppose that the sequences $\{x_n\} \subset \Omega$, $\{\bar{x}_n\} \subset \Omega$ and $\{\omega_n\} \subset [0, 1]$ satisfy* (2.2) *and Goldstein's rule* (2.4) *with δ fixed in* $(0, \frac{1}{2}]$, *or Armijo's rule* (2.5) *with δ and β fixed in* $(0, \frac{1}{2}]$ *and* $(0, 1)$ *respectively. Then conclusions* (i) *through* (iv) *of Theorem* 3.1 *hold once again, except* *that* (3.3B), (3.4B), (3.5B) *and* (3.5C) *are now replaced by*

$$g=\delta\cdot\min\left\{1,\frac{Cr_1}{L_1D^2}\right\}>0, \tag{3.3B$'$}$$

$$\lambda=\max\left\{1-\delta,\,1-\frac{C\delta}{(L_1/A)(1+L_1/A)^2}\right\}, \tag{3.4B$'$}$$

$$N_0\leqq 1+\frac{\log(L_1r_1/CA_s^2)}{\log(1/\lambda_s)}, \tag{3.5B$'$}$$

and

$$\lambda_s=\max\left\{1-\delta,\,1-\frac{C\delta}{(L_1D/A_s)(1+L_1D/A_s)^2}\right\} \tag{3.5C$'$}$$

respectively, where $C=2\delta$ *for Goldstein's rule* (2.4) *and* $C=2\beta(1-\delta)$ *for Armijo's rule* (2.5).

Proof. As in the proof of Theorem 3.1, one has $\inf_\Omega F>-\infty$ and $r_n\geqq r_{n+1}\geqq 0$ for all $n\geqq 1$. Furthermore, it is immediate from (3.8) that

$$0\leqq r_{n+1}\leqq(1-\delta\omega_n)r_n \tag{3.16}$$

for either (2.4) or (2.5). If $\langle F'_{N+1}, x_{N+1}-\bar{x}_{N+1}\rangle=0$, then $x_n=x_{N+1}$ and $r_n=0$ for $n\geqq N+1$, as before. On the other hand, if $\langle F'_n, x_n-\bar{x}_n\rangle>0$ for $1\leqq n\leqq N$, then $r_n>0$ and (2.4), (2.6) and (3.2) yield either

$$\omega_n=1 \tag{3.17A}$$

or

$$1-\delta\geqq\frac{F_n-F_{n+1}}{\omega_n\langle F'_n, x_n-\bar{x}_n\rangle}\geqq 1-\frac{1}{2}\omega_n\frac{L_1\|\bar{x}_n-x_n\|^2}{\langle F'_n, x_n-\bar{x}_n\rangle}.$$

In the latter case, one has

$$1\geqq\omega_n\geqq 2\delta\cdot\frac{\langle F'_n, x_n-\bar{x}_n\rangle}{L_1\|\bar{x}_n-x_n\|^2}\geqq\delta\cdot\frac{r_n}{L_1\|\bar{x}_n-x_n\|^2}, \tag{3.17B}$$

in view of (3.8). It now follows from (3.16) and (3.17) that

$$0<(r_{n+1}/r_1)\leqq(r_n/r_1)-q_n(r_n/r_1)^2 \tag{3.18A}$$

for $1\leqq n\leqq N$, with

$$q_n=r_1\delta\cdot\min\left\{\frac{C}{L_1\|\bar{x}_n-x_n\|^2},\frac{1}{r_n}\right\} \tag{3.18B}$$

and

$$C=2\delta. \tag{3.19}$$

Similarly, if $\langle F'_n, x_n-\bar{x}_n\rangle>0$ for $1\leqq n\leqq N$, then $r_n>0$ and (2.5), (2.6) and (3.2) yield either $m_n=0$ and

$$\omega_n=\beta^0=1 \tag{3.20A}$$

or $m_n>0$ and

$$\delta>\frac{F_n-F(x_n+\beta^{m_n-1}(\bar{x}_n-x_n))}{\beta^{m_n-1}\langle F'_n, x_n-\bar{x}_n\rangle}\geqq 1-\frac{1}{2}\beta^{m_n-1}\frac{L_1\|\bar{x}_n-x_n\|^2}{\langle F'_n, x_n-\bar{x}_n\rangle}.$$

In the latter case, one has

$$
\text{(3.20B)}\qquad
\begin{aligned}
1 \geqq \omega_n = \beta^{m_n} &\geqq 2\beta(1-\delta)\cdot\frac{\langle F'_n, x_n-\bar{x}_n\rangle}{L_1\|\bar{x}_n-x_n\|^2}\\
&\geqq 2\beta(1-\delta)\cdot\frac{r_n}{L_1\|\bar{x}_n-x_n\|^2}
\end{aligned}
$$

because of (3.8). Consequently (3.16) yields (3.18) with

$$\text{(3.21)}\qquad C = 2\beta(1-\delta).$$

From here on, one proceeds exactly as in the proof of Theorem 3.1, with a single exception. In part (iv), the inequality (3.14) still gives $\bar{x}_n = \xi$ when $(L_1/A_s)\|x_n-\xi\| < 1$; however, $\bar{x}_n = \xi$ no longer automatically implies that $x_{n+1} = \xi$. Nevertheless, (1.5), (2.2), (3.17) and (3.20) do yield

$$\text{(3.22)}\qquad (L_1/A_s)\|x_n-\xi\| < C \leqq 1 \Rightarrow \bar{x}_n = \xi \quad\text{and}\quad x_{n+1} = \xi,$$

where C is specified by (3.19) for Goldstein's rule (2.4), or by (3.21) for Armijo's rule (2.5). As in the proof of Theorem 3.1, one also finds that

$$\text{(3.23)}\qquad 0 \leqq \|x_n-\xi\| \leqq (r_1/A_s)\lambda_s^n,$$

where λ_s is given by (3.5C′). The estimate (3.5B′) now follows from (3.22) and (3.23). Q.E.D.

Note 3.1. Apart from conclusion (iv), the proof of Theorem 3.1 after inequality (3.7) is a rewording of the proof of Theorem 5.2 following (5.14) in [1]. Furthermore, in case (iv), estimates comparable with (3.5) and (3.5′) can also be established for the explicit step size formulas of [1]. In particular, the estimate (3.5) holds for the rule of Demyanov and Rubinov in Theorem 5.2 of [1]. Thus, $(L_1/A)\|x_n-\xi\| < 1 \Rightarrow \bar{x}_n = \xi$ once again, and according to (4.2) in [1] one has

$$\omega_n = 1 \quad\text{or}\quad 1 \geqq \omega_n \geqq \frac{\langle F'_n, x_n-\bar{x}_n\rangle}{L_1\|x_n-\bar{x}_n\|^2} \geqq \frac{1}{(L_1/A)\|x_n-\xi\|} \qquad \text{if } x_n \neq \xi \quad\text{and}\quad \bar{x}_n = \xi.$$

Consequently $(L_1/A)\|x_n-\xi\| < 1 \Rightarrow x_k = \bar{x}_k = \xi$ for $k > n$. In all four cases considered, the a priori convergence rate estimates derived here for exact line minimization are therefore identical to corresponding estimates for the explicit rule of Demyanov and Rubinov in the limiting case $L = L_1$, and also for the implicit rules of Goldstein and Armijo in the limit as $\delta \to \frac{1}{2}$ and $\beta \to 1$ (best estimates obtained for the other explicit formula in [1] are similar but not identical).

Note 3.2. It is explained in Note 5.1 of [1] how convergence rate estimates may sometimes be sharpened when Ω satisfies conditions of the uniform convexity type and $\|F'(x)\|$ is bounded away from 0 on Ω. This observation applies to the present analysis as well.

4. Convergence theorems for nonconvex functions. The extremals of F in Ω are the zeros of the functional Φ: $\Omega \to [0, \infty]$ defined by

$$\text{(4.1)}\qquad \Phi(x) = \langle F'(x), x\rangle - \inf_{y\in\Omega}\langle F'(x), y\rangle$$

as x ranges over Ω. Unless F is convex, an extremal need not minimize F even in a local sense, and conditional gradient algorithms need not generate minimizing sequences. On the other hand, extremals for F are always global minimizers of Φ in Ω, and if F' is uniformly continuous the conditional gradient schemes of § 2 always produce "extremizing" sequences $\{x_n\}$, i.e. $\Phi(x_n)\to 0$ as $n\to\infty$. Moreover, it can be shown that continuity of F' implies lower semicontinuity of Φ, and consequently every limit point of $\{x_n\}$ is an extremal if $\Phi(x_n)\to 0$. For compact Ω, this means that $\{x_n\}$ must converge to some subset of the extremal set

$$\Omega_\Phi=\{\xi\in\Omega\,|\,\Phi(\xi)=0\} \tag{4.2}$$

and, under certain circumstances, may converge to some particular element in Ω_Φ. Finally, if F' is weak–strong continuous (i.e., if $\|F'_n-F'(\xi)\|$ converges to 0 whenever x_n converges weakly to ξ), then Φ is weakly lower semicontinuous and therefore $\Phi(x_n)\to 0$ implies that every weak limit point of $\{x_n\}$ is an extremal. For weakly compact Ω, this ensures that $\{x_n\}$ converges at least weakly to the set Ω_Φ and may converge to some specific extremal in Ω_Φ. Theorem 1 of [2] occupies a central position in the following development of these results. However, additional ideas from [1], [5] and [8] are required to complete the analysis; in particular, the concept of nonsingularity introduced in [1] has some importance here, as it does in § 3.

LEMMA 4.1. *Let X be a real Banach space and suppose that $F\colon X\to\mathbb{R}^1$ has a continuous Fréchet derivative on $\Omega\subset X$. Then the associated functional $\Phi\colon\Omega\to[0,\infty]$ defined in* (4.1) *is lower semicontinuous. Moreover, if F' is weak–strong continuous on Ω, then Φ is weakly lower semicontinuous.*

Proof. Suppose that $x_n\to x\in\Omega$. For all $n\geqq 1$ and all $z\in\Omega$, one has

$$\begin{aligned}\Phi(x_n)&=\langle F'_n,x_n\rangle-\inf_{y\in\Omega}\langle F'_n,y\rangle\\ &\geqq\langle F'_n,x_n-z\rangle\\ &=\langle F'_n-F'(x),x_n-z\rangle+\langle F'(x),x_n-z\rangle.\end{aligned} \tag{4.3}$$

Consequently if F' is continuous, it follows that

$$\varliminf_{n\to\infty}\Phi(x_n)\geqq 0+\langle F'(x),x-z\rangle$$

for all $z\in\Omega$, and therefore

$$\varliminf_{n\to\infty}\Phi(x_n)\geqq\langle F'(x),x\rangle-\inf_{z\in\Omega}\langle F'(x),z\rangle=\Phi(x). \tag{4.4}$$

Furthermore, if F' is weak–strong continuous and $x_n\overset{\text{wk.}}{\rightharpoonup}x$, then $F'(x_n)-F'(x)\to 0$, x_n-z is bounded, and therefore (4.4) follows once again from (4.3.) Q.E.D.

Note 4.1. In reflexive spaces, uniform continuity and compactness of F' on the closed ball $B(0;r)$ implies weak–strong continuity of F' on $B(0;r-\varepsilon)$ with $r>\varepsilon>0$. This result and other general sufficient conditions for weak–strong continuity of gradients are established in [8] (where weak–strong continuity is called strong continuity). As a concrete illustration, consider the quadratic functional

$$F(x(\cdot))=\int_0^1 a(t)x(t)\,dt+\frac{1}{2}\int_0^1\int_0^1 K(t,\tau)x(t)x(\tau)\,dt\,d\tau$$

on the Hilbert space $X=\mathscr{L}^2([0,1],\mathbb{R}^1)$, with $a(\cdot)$ fixed in X and $K(\cdot,\cdot)\in\mathscr{L}^2([0,1]\times[0,1],\mathbb{R}^1)$. The derivative of this function is represented by

$$\nabla F(\cdot)=a(\cdot)+\int_0^1 K(\cdot,\tau)x(\tau)\,d\tau$$

at $x(\cdot)\in X$. Since the kernel K is square-integrable, it follows that the associated operator $F'\colon X\to X^*$ is Lipschitz continuous and compact, and therefore weak–strong continuous.

Note 4.2. For deeper and more general results on the continuity properties of minimum sets, see [9].

THEOREM 4.1. *Let Ω be a bounded convex subset of a real Banach space X and let $F\colon X\to\mathbb{R}^1$ have a uniformly continuous Fréchet derivative F' on Ω. Let the sequences $\{x_n\}\subset\Omega$, $\{\bar{x}_n\}\subset\Omega$ and $\{\omega_n\}\subset[0, 1]$ satisfy* (2.2) *and also the line minimization rule* (2.3), *or Goldstein's rule* (2.4) *with δ fixed in* $(0,\frac{1}{2}]$, *or Armijo's rule* (2.5) *with δ and β fixed in* $(0,\frac{1}{2}]$ *and* $(0, 1)$ *respectively. Then the sequence $\{F(x_n)\}$ converges monotonically downward to some limit $l\geqq\inf_\Omega F>-\infty$, and*

$$\lim_{n\to\infty}\Phi(x_n)=0 \tag{4.5}$$

where Φ is defined in (4.1).

Proof. If F' is uniformly continuous on the bounded convex set Ω, then F' is bounded on Ω and it follows from the mean value theorem that $\inf_\Omega F>-\infty$. For any of the rules (2.3), (2.4) or (2.5), one also has $F_n\geqq F_{n+1}$ for all $n\geqq 1$; consequently $\{F_n\}$ converges to some $l\geqq\inf_\Omega F$. Furthermore, in all three cases $\Phi(x_N)=0\Rightarrow\Phi(x_n)=0$ for $n\geqq N$.

Suppose that $\Phi(x_n)>0$ for $n\geqq 1$. Condition (4.5) can be established for the line minimization rule (2.3) by a minor modification of the proof of Theorem 1.1 in [5, p. 118]. Thus, it follows from (2.3) that for all $\omega\in[0, 1]$,

$$\begin{aligned}F_{n+1}-F_n&\leqq\omega\langle F'_n, \bar{x}_n-x_n\rangle+\omega\langle F'(\zeta)-F'_n, \bar{x}_n-x_n\rangle\\&\leqq\omega\langle F'_n, \bar{x}_n-x_n\rangle+\omega\|F'(\zeta)-F'_n\|D,\end{aligned}$$

where ζ is somewhere on the line segment joining x_n to $x_n+\omega(\bar{x}_n-x_n)$, and where $D=\operatorname{diam}\Omega<\infty$. For $\omega\in(0, 1]$, this yields

$$0<\Phi(x_n)\leqq\frac{F_n-F_{n+1}}{\omega}+\|F'(\zeta)-F'_n\|D.$$

Since $\|\zeta-x_n\|\leqq\omega\|\bar{x}_n-x_n\|\leqq\omega D$, and since F' is uniformly continuous, there is an $\omega>0$ so small that $\|F'(\zeta)-F'_n\|\leqq\varepsilon/D$ for all $n\geqq 1$, where ε is any given positive number. For such an ω,

$$0<\Phi(x_n)\leqq\frac{F_n-F_{n+1}}{\omega}+\varepsilon,$$

and therefore

$$0\leqq\varliminf_{n\to\infty}\Phi(x_n)\leqq\varlimsup_{n\to\infty}\Phi(x_n)\leqq\varepsilon,$$

because $F_n\to l$ implies that $F_n-F_{n+1}\to 0$. Since ε can be arbitrarily small here, these inequalities give (4.5) for the line minimization rule (2.3).

Again, suppose that $\Phi(x_n)>0$ for $n\geqq 1$. Condition (4.5) can be established for the Goldstein and Armijo rules (2.4) and (2.5) by a method of proof similar to that devised for Theorem 1 in [2]. Thus, it follows at once from (2.4) or (2.5) that

$$F_n-F_{n+1}\geqq\delta\omega_n\Phi(x_n)\geqq 0$$

for $n \geqq 1$, and since $F_n - F_{n+1} \to 0$, this gives

$$\lim_{n\to\infty} \omega_n \Phi(x_n) = 0.$$

Therefore if (4.5) is false, there is an $\varepsilon > 0$ and a subsequence $\{n_k\}$ such that

$$\lim_{k\to\infty} \omega_{n_k} = 0, \tag{4.6A}$$

while

$$\Phi_{n_k} \geqq \varepsilon. \tag{4.6B}$$

For k sufficiently large, one must then have $\omega_{n_k} < 1$; consequently (2.4) and the mean value theorem give

$$\begin{aligned} 1-\delta \geqq \frac{F_{n_k} - F_{n_k+1}}{\omega_{n_k}\langle F'_{n_k}, x_{n_k} - \bar{x}_{n_k}\rangle} &\geqq 1 + \frac{\langle F'(\zeta_k) - F'_{n_k}, x_{n_k} - \bar{x}_{n_k}\rangle}{\langle F'_{n_k}, x_{n_k} - \bar{x}_{n_k}\rangle} \\ &\geqq 1 - \frac{\|F'(\zeta_k) - F'_{n_k}\| \cdot D}{\Phi(x_{n_k})}, \end{aligned}$$

and therefore

$$\|F'(\zeta_k) - F'_{n_k}\| \geqq \frac{\varepsilon\delta}{D} > 0 \tag{4.7A}$$

for large k, where ζ_k is somewhere on the line segment joining x_{n_k} to $x_{n_k} + \omega_{n_k}(\bar{x}_{n_k} - x_{n_k})$, and so

$$\|\zeta_k - x_{n_k}\| \leqq \omega_{n_k}\|\bar{x}_{n_k} - x_{n_k}\| \leqq \omega_{n_k} D. \tag{4.7B}$$

But, in view of (4.6A) and the uniform continuity of F' it follows from (4.7B) that

$$\lim_{k\to\infty} \|F'(\zeta_k) - F'_{n_k}\| = 0, \tag{4.8}$$

which contradicts (4.7A). This contradiction establishes (4.5) for Goldstein's rule (2.4).

Similarly, if (4.6) holds, then Armijo's rule (2.5) and the mean value theorem produce

$$\delta > \frac{F_{n_k} - F(x_{n_k} + \beta^{-1}\omega_{n_k}(\bar{x}_{n_k} - x_{n_k}))}{\beta^{-1}\omega_{n_k}\langle F'_{n_k}, x_{n_k} - \bar{x}_{n_k}\rangle} \geqq 1 - \frac{\|F'(\zeta_k) - F'_{n_k}\| \cdot D}{\Phi(x_{n_k})},$$

and therefore

$$\|F'(\zeta_k) - F'_{n_k}\| \geqq \frac{\varepsilon(1-\delta)}{D} \tag{4.9A}$$

for large k, where ζ_k is now somewhere on the line segment joining x_{n_k} to $x_{n_k} + \beta^{-1}\omega_{n_k}(\bar{x}_{n_k} - x_{n_k})$, and consequently

$$\|\zeta_k - x_{n_k}\| \leqq \omega_{n_k}\beta^{-1} D. \tag{4.9B}$$

Once again (4.8) follows from (4.6A), (4.9B) and the uniform continuity of F'. This contradicts (4.9A) and establishes (4.5) for Armijo's rule (2.5). Q.E.D.

COROLLARY 1. *Let Ω_Φ denote the extremal set* (4.2) *and put*

$$L_\Phi = \{\xi \in \Omega_\Phi \mid F(\xi) = l\}. \tag{4.10}$$

Then L_Φ is closed and every limit point of $\{x_n\}$ belongs to L_Φ. Furthermore, if every subsequence of $\{x_n\}$ has a limit point in Ω then L_Φ is not empty and $\{x_n\}$ converges strongly to L_Φ in the sense that every open neighborhood $\mathcal{N}$ of L_Φ contains all but finitely many of the x_n's; equivalently,

$$\lim_{n\to\infty}\left\{\min_{\xi\in L_\Phi}\|x_n-\xi\|\right\}=0. \tag{4.11}$$

In particular, if L_Φ consists of a single element ξ, then $\{x_n\}$ converges strongly to ξ.

Proof. According to Lemma 4.1, Φ is lower semicontinuous and therefore Ω_Φ is closed. Since F is continuous it follows that L_Φ is also closed. Furthermore, $x_{n_k}\to\xi\in\Omega$ implies that

$$0=\lim_{k\to\infty}\Phi(x_{n_k})=\varliminf_{k\to\infty}\Phi(x_{n_k})\geqq\Phi(\xi)\geqq 0$$

and

$$l=\lim_{k\to\infty}F(x_{n_k})=F(\xi);$$

consequently, $\xi\in L_\Phi$. By hypothesis, $\{x_n\}$ has a limit point in Ω, therefore L_Φ is not empty. Finally, if $\{x_n\}$ does not converge strongly to L_Φ, there is an open set $\mathcal{N}\supset L_\Phi$ and a subsequence $\{x_{n_k}\}$ with range in $\mathcal{N}'$ = the complement of $\mathcal{N}$. But $\{x_{n_k}\}$ has a limit point $\xi\in\Omega$, and since $\mathcal{N}'$ is closed this gives the contradiction $\xi\in\mathcal{N}'\subset L'_\Phi$. Q.E.D.

COROLLARY 2. *Suppose that F is weakly continuous and F' is weak–strong continuous. Then the set L_Φ in* (4.10) *is weakly closed and every weak limit point of $\{x_n\}$ belongs to L_Φ. Furthermore, if every subsequence of $\{x_n\}$ has a weak limit point in Ω, then L_Φ is not empty and $\{x_n\}$ converges weakly to L_Φ, i.e., every weak open neighborhood $\mathcal{N}$ of L_Φ contains all but finitely many of the x_n's. In particular, if L_Φ consists of a single element ξ, then x_n converges weakly to ξ. Finally, if every extremal in L_Φ satisfies the strong nonsingularity condition* (1.3), *then $\{x_n\}$ converges strongly to L_Φ.*

Proof. The first part of the proof is identical to the proof of Corollary 1, with strong topological concepts replaced by their weak counterparts. If every $\xi\in L_\Phi$ satisfies (1.3), then $x_{n_k}\overset{\text{wk.}}{\rightharpoonup}\xi\Rightarrow a(\|x_{n_k}-\xi\|)\to 0\Rightarrow\|x_{n_k}-\xi\|\to 0$. It follows that every weak limit of $\{x_n\}$ is also a strong limit point, and the rest is now immediate from Corollary 1. Q.E.D.

Note 4.3. In reflexive spaces X, weak–strong continuity of F' on the closed ball $B(0;r)$ implies weak continuity of F on $B(0;r)$. (See [8].)

Note 4.4. At each $x\in\Omega$, the continuous linear functional $\langle F'(x),\cdot\rangle$ is weakly continuous; therefore if Ω is nonempty and weakly compact, the set $T(x)$ is nonempty, and it follows that for each fixed $x_1\in\Omega$, there are corresponding sequences $\{x_n\}$, $\{\bar{x}_n\}$ and $\{\omega_n\}$ which jointly satisfy (2.2), and (2.3) or (2.4).

Note 4.5. If F' is continuous but not uniformly continuous, it is still true that $x_{n_k}\to\xi\Rightarrow\Phi(x_{n_k})\to 0=\Phi(\xi)\Rightarrow\zeta\in L_\Phi$, provided $\inf_\Omega F>-\infty$ and the remaining conditions of Theorem 4.1 are satisfied. This can be shown by a straightforward modification of the proof of Theorem 4.1; however, the additional labor required scarcely seems worthwhile, for the following reason: if Ω is compact, uniform continuity of F' follows automatically from the continuity of F'; on the other hand, if Ω is *not* compact, the inclusion of all limit points of $\{x_n\}$ in L_Φ does not ensure convergence of $\{x_n\}$ to L_Φ.

Note 4.6. For convex F, (4.5) implies that $\lim_{n\to\infty}F_n=l=\inf_\Phi F$, because of (3.8) (cf. part (i) of Theorem 3.1 and 3.2 for Lipschitz continuous F').

Note 4.7. An extremal ξ is said to be *nonsingular* in [1] if and only if ξ is the unique minimizer of the associated linear functional $\langle F'(\xi), \cdot \rangle$ over Ω, i.e., if and only if $T(\xi) = \{\xi\}$. For compact convex Ω, nonsingular extremals are also strongly nonsingular [1, Thm. 3.3]. If Ω is strictly convex (resp., uniformly convex) in the sense of [5], then any extremal ξ is automatically nonsingular (resp., strongly nonsingular) provided $F'(\xi) \neq 0$ (e.g., see [1, Thm. 3.4]).

When Ω is compact, it follows from Corollary 1 of Theorem 4.1 that $\{x_n\}$ can diverge only if F has the same value on two or more extremals. However, even when this happens, $\{x_n\}$ may still converge.

LEMMA 4.2. *Let the hypotheses of Theorem* 4.1. *hold. Furthermore, suppose that every subsequence of* $\{x_n\}$ *has a limit point in* Ω, *and that all extremals in the set* L_Φ *satisfy the strong nonsingularity condition* (1.3). *Then*

$$\lim_{n\to\infty} \|x_n - \bar{x}_n\| = 0, \tag{4.12}$$

and therefore

$$\lim_{n\to\infty} \|x_{n+1} - x_n\| = 0. \tag{4.13}$$

Proof. If (4.12) is false, there is an $\varepsilon > 0$, a $\xi \in \Omega$ and a subsequence $\{x_{n_k}\}$ such that $x_{n_k} \to \xi$ and

$$\|x_{n_k} - \bar{x}_{n_k}\| \geqq \varepsilon \tag{4.14}$$

for $k \geqq 1$. According to Corollary 1 of Theorem 4.1, ξ belongs to L_Φ; consequently it follows from (1.3) and Theorem 3.2 in [1] that $x_{n_k} \to \xi \Rightarrow \bar{x}_{n_k} \to \xi$, and therefore $\|x_{n_k} - \bar{x}_{n_k}\| \to 0$. This contradicts (4.14) and establishes (4.12). Since $\omega_n \in [0, 1]$, (4.13) is immediate from (2.2) and (4.12). Q.E.D.

LEMMA 4.3. *Suppose that* $\{x_n\} \subset \Omega$ *satisfies* (4.13) *and that every subsequence of* $\{x_n\}$ *has a limit point in* Ω. *Let L denote the limit point set for* $\{x_n\}$. *Then* $\{x_n\}$ *either converges to some point of* Ω *or else L has infinitely many elements with*

$$\inf_{\substack{\xi,\eta \in L \\ \xi \neq \eta}} \|\xi - \eta\| = 0. \tag{4.15}$$

Proof. The following argument is an elaboration on part of the proof of Theorem 1 in [2]. Since every subsequence of $\{x_n\}$ has a limit point in Ω it follows that $L \subset \Omega$ and that $\{x_n\}$ converges to ξ if and only if $L = \{\xi\}$. Suppose that L has more than one element and that (4.15) is false. Put

$$\varepsilon = \tfrac{1}{3} \inf_{\substack{\xi,\eta \in L \\ \xi \neq \eta}} \|\xi - \eta\| > 0$$

and

$$B(\xi) = \{x \in \Omega \mid \|x - \xi\| < \varepsilon\}. \tag{4.16}$$

As ξ ranges over L, (4.16) describes a family of pairwise disjoint open balls with

$$\operatorname{dist}\{B(\xi), B(\eta)\} \geqq \varepsilon \tag{4.17}$$

for $\xi \neq \eta$. According to (4.13),

$$\exists N_1 \ni n \geqq N_1 \Rightarrow \|x_{n+1} - x_n\| < \varepsilon. \tag{4.18}$$

Suppose now that

$$\text{(4.19)} \qquad \exists N_2 \ni n \geqq N_2 \Rightarrow x_n \in \bigcup_{\xi \in L} B(\xi).$$

For any fixed $\xi \in L$, there is a $k \geqq \max\{N_1, N_2\}$ such that $x_k \in B(\xi)$. The corresponding set $S_k = \{n > k \mid x_n \notin B(\xi)\}$ is bounded below by k and is also nonempty since x_n must belong to $B(\eta)$ for some $\eta \in L$, $\eta \neq \xi$, and for infinitely many values of $n > k$. Thus S_k has a least element m, and by construction one has $m - 1 \geqq \max\{N_1, N_2\}$, $x_{m-1} \in B(\xi)$, and $x_m \in B(\eta)$ for some $\eta \in L$ with $\eta \neq \xi$. But this is impossible, in view of (4.17) and (4.18), and so (4.19) cannot hold. On the other hand, if (4.19) is false, $\{x_n\}$ has a limit point ζ in the complement of $\bigcup_{\xi \in L} B(\xi)$, which is impossible, since ζ must belong to L by definition. This contradiction proves that if L has more than one element, it must have *infinitely* many elements and condition (4.15) must hold. Q.E.D.

THEOREM 4.2. *Let* Ω, F, $\{x_n\}$, $\{\bar{x}_n\}$ *and* $\{\omega_n\}$ *satisfy the hypotheses of Theorem* 4.1 *and suppose that* Ω *is also compact. Furthermore, suppose that the extremal set* L_Φ *in* (4.10) *has finitely many members*, ξ_i, *and that each* ξ_i *is nonsingular* (*see* Note 4.6). *Then* $\{x_n\}$ *converges strongly to one of the* ξ_i*'s.*

Proof. By Theorem 3.3 of [1] every $\xi_i \in L_\Phi$ satisfies the strong nonsingularity condition (1.3). Consequently, (4.13) follows from Lemma 4.2. According to Corollary 1 of Theorem 4.1, all limit points of $\{x_n\}$ are in the finite set L_Φ; hence $\{x_n\}$ must converge to one of the ξ_i's by Lemma 4.3. Q.E.D.

THEOREM 4.3. *Let* Ω, F, $\{x_n\}$, $\{\bar{x}_n\}$ *and* $\{\omega_n\}$ *satisfy the hypotheses of Theorem* 4.1. *Furthermore, let* F *be weakly continuous, let* F' *be weak–strong continuous, and let* Ω *be weakly compact. Finally, suppose that every member of the extremal set* L_Φ *in* (4.10) *satisfies the strong nonsingularity condition* (1.3), *and that*

$$\text{(4.20)} \qquad \inf_{\substack{\xi, \eta \in L_\Phi \\ \xi \neq \eta}} \|\xi - \eta\| > 0.$$

Then $\{x_n\}$ *converges strongly to some* $\xi \in L_\Phi$.

Proof. Every subsequence of $\{x_n\}$ has a weak limit point in Ω. Moreover, by Corollary 2 of Theorem 4.1, every weak limit point ξ of $\{x_n\}$ belongs to L_Φ and is therefore strongly nonsingular. Consequently $x_{n_k} \overset{\text{wk.}}{\rightharpoonup} \xi \Rightarrow a(\|x_{n_k} - \xi\|) \to 0 \Rightarrow \|x_{n_k} - \xi\| \to 0$, in view of (1.2)–(1.3). Thus, every weak limit point of $\{x_n\}$ is also a strong limit point of $\{x_n\}$ and therefore (4.12) holds. Finally, (4.15) is impossible according to (4.20); hence it follows once again from Lemma 4.3 that $\{x_n\}$ converges strongly to some $\xi \in L_\Phi$. Q.E.D.

Note 4.8. For compact Ω, condition (4.13) holds if and only if L_Φ has finitely many elements. However, for weakly compact Ω (4.20) does not imply that L_Φ is finite.

Note 4.9. When F' is Lipschitz continuous, the conclusions in Theorem 4.1 and its corollaries and Theorems 4.2 and 4.3 also hold for conditional gradient sequences generated by the explicit step length formula of Demyanov and Rubinov (cf. [1] and [5]).

Note 4.10. The quadratic functional in Note 4.1 has a weak–strong continuous derivative on the reflexive space $L^2([0, 1], \mathbb{R}^1)$, and is therefore weakly continuous (see Note 4.3). Furthermore, every closed, bounded convex set Ω in a reflexive space is automatically weakly compact. Finally, if Ω is uniformly convex, then every extremal ξ is automatically strongly nonsingular provided $F'(\xi) \neq 0$ (see Note 4.7). Thus, Theorem 4.3 can help to explain the behavior of projected gradient processes for certain *nonconvex* F (e.g., indefinite or negative semidefinite quadratic functionals with regular kernels in L^2 spaces) on certain *noncompact* sets Ω (e.g., closed balls in L^2 spaces).

REFERENCES

[1] J. C. DUNN, *Rates of convergence for conditional gradient algorithms near singular and nonsingular extremals*, this Journal, 17 (1979), pp. 187–211.

[2] A. A. GOLDSTEIN, *On steepest descent*, this Journal, 3 (1965), pp. 147–151,

[3] A. A. GOLDSTEIN, *On Newton's method*, Numer. Math., 7 (1965), pp. 391–393.

[4] L. ARMIJO, *Minimization of functions having Lipschitz continuous first partial derivatives*, Pacific J. Math., 16 (1966), pp. 1–3.

[5] V. F. DEMYANOV AND A. M. RUBINOV, *Approximate Methods in Optimization Problems*, American Elsevier, New York, 1970.

[6] E. POLAK, *A historical survey of computational methods in optimal control*, SIAM Rev., 15 (1973), pp. 553–584.

[7] J. M. ORTEGA AND W. C. RHEINBOLDT, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[8] M. M. VAINBERG, *Variational Methods for the Study of Nonlinear Operators*, Holden-Day, San Francisco, 1964.

[9] G. B. DANTZIG, J. FOLKMAN AND N. SHAPIRO, *On the continuity of the minimum set of a continuous function*, J. Math. Anal. Appl., 17 (1967), pp. 519–548.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 5, September 1980

0363-0129/80/1805-0002 $01.00/0

CONTROLLABILITY AND STABILIZABILITY IN MULTI-PAIR SYSTEMS*

DAVID P. STANFORD† AND LUTHER T. CONNER, JR.†

Abstract. In this paper discrete-time systems of the form

$$x_{k+1}=C_px_k+D_pU_k,$$

in which the pair (C_p, D_p) is selected from a finite set $\{(C_i, D_i)\}_{i=1}^N$, are studied. Such systems, called "multi-pair systems," arise naturally in the study of multi-rate sampled-data systems. It is shown that the set of points reachable from zero (the controllable set) is a subspace under certain hypotheses, but not always. When this is the case, an extended version of the controllability canonical form is obtained, and it is applied to the study of state deadbeat response and more general forms of stabilizability.

1. Introduction. This paper studies multi-pair systems $L=\{(C_i, D_i)\}_{i=1}^N$ with C_i a real $n\times n$ matrix and D_i a real $n\times m$ matrix. These systems are linear discrete-time systems of the form

$$x_{k+1}=C_px_k+D_pu_k,$$

where, at each stage of the iteration, we select a pair of matrices (C_p, D_p) as well as a control u_k. In contrast to the usual time-varying discrete-time system $x_{k+1}=C_kx_k+D_ku_k$, the pairs (C_p, D_p) are to be selected from a given, fixed, finite set $\{(C_i, D_i)\}_{i=1}^N$, and the selection is, in general, not dependent on the time k or the selected control u_k. Such a system arises from a linear control system of the form

$$\dot{x}=Ax+Bu,$$

where A is a real $n\times n$ constant matrix and B is a real $n\times m$ constant matrix. We apply to this problem the method of multi-rate sampling, in which each sampling interval length is selected from a fixed finite set $\{s_1, s_2, \cdots, s_N\}$ of positive numbers. As explained in [2], we obtain a set of N pairs (C_i, D_i), in which

$$C_i=e^{s_iA}\quad\text{and}\quad D_i=\int_0^{s_i}e^{tA}\,dt\,B.$$

The pairs (C_i, D_i) describe values of x at sampling instants by the formula

$$x_{k+1}=C_px_k+D_pu_k,$$

where x_k and x_{k+1} are the values of x at the beginning and end of a sampling interval of length s_p, and u_k is the constant control applied during that interval.

It also appears that in a sense these multi-pair systems include a discrete version of "variable structure systems" as described in [5]. Given a pair (A, B) with A $n\times n$ and B $n\times m$, we could choose $N\in Z^+$, $C_i=A$ and $D_i=B$ for all i in $\{1, 2, \cdots, N\}$. In this way the choice of N different feedback matrices F_i for the single pair (A, B) is effected. It is probably true that our work on controllability in [3] and in this paper has no applicability to variable structure systems, but the feedback selection in [2] and convergibility properties discussed in the last section of this paper may be applicable.

The stabilizability of multi-pair systems through feedback has been investigated in [2]. Pre-contractiveness and contractiveness of the closed-loop system are introduced,

* Received by the editors August 9, 1979, and in final revised form January 25, 1980.

† College of William and Mary, Williamsburg, Virginia 23185. This research was supported by NASA-Langley Research Center under Grant No. NAS1-14972.

and the selection of feedbacks is discussed. In [3], the property of complete controllability for a multi-pair system is defined, and it is shown that under certain hypotheses, complete controllability implies the capability through feedback of a state deadbeat response. It is conjectured that this implication holds whenever the C_i's are nonsingular and the D_i's are of full rank.

In this paper we show that the set of points reachable from zero under a multi-pair system (the "controllable set") is a subspace of R^n whenever all C_i's are nonsingular. This set is a subspace for more general systems, but not for all systems.

We next introduce the "controllability canonical form" for multi-pair systems, and we apply it to the study of state deadbeat response and more general forms of stabilizability. An interesting corollary to this work is the converse to the conjecture from [3] mentioned above.

2. Controllable sets. Given a sequence $\{C_i\}_{i=1}^N$ of real $n\times n$ matrices and a sequence $\{D_i\}_{i=1}^N$ of real $n\times m$ matrices, we study the discrete-time system

$$L\colon x_{k+1}=C_ix_k+D_iu_k,\qquad i\in\{1,2,\cdots,N\},$$

with $x_k\in R^n$ and $u_k\in R^m$. The system is completely determined by the C_i's and D_i's, and so we define, for n, m, N, $\in Z^+$,

$$\Sigma^0(n,m,N)=\{L=\{(C_i,D_i)\}_{i=1}^N|C_i \text{ real } n\times n, D_i \text{ real } n\times m\}.$$

The multi-pair systems arising from the sampled-data problem in the introduction have the property that the C_i's are nonsingular and the D_i's all have the same rank. Thus we may as well assume the D_i's are of full (column) rank, and we define

$$\Sigma(n,m,N)=\{L=\{(C_i,D_i)\}_{i=1}^N\in\Sigma^0(n,m,N)|C_i \text{ nonsingular}, D_i \text{ full rank}\}.$$

Throughout this paper n, m, and N denote positive integers, $\bar{N}$ denotes $\{1,2,\cdots,N\}$, and, for $k\in Z^+$, Γ_k denotes the set of all k-termed sequences with terms in $\bar{N}$, and U_k denotes the set of all k-termed sequences with terms in R^m. We let Γ denote $\bigcup\{\Gamma_k|k\in Z^+\}$.

Suppose $L=\{(C_i,D_i)\}_{i=1}^N\in\Sigma^0(n,m,N)$, $k\in Z^+$, $u\in U_k$, and $\alpha\in\Gamma_k$. For $x\in R^n$, the *trajectory of x under u and α*, denoted $T(L,x,u,\alpha)$, is the sequence $\{x_i\}_{i=1}^{k+1}$, where $x_1=x$, and $x_{i+1}=C_{\alpha(i)}x_i+D_{\alpha(i)}u_i$ for $i\in\bar{k}$. For $i\in\overline{k+1}$, the ith term of $T(L,x,u,\alpha)$ is denoted by $T_i(L,x,u,\alpha)$. The *terminal point* of $T(L,x,u,\alpha)$ is $T_{k+1}(L,x,u,\alpha)$, denoted by $TP(L,x,u,\alpha)$.

For $L=\{(C_i,D_i)\}_{i=1}^N\in\Sigma^0(n,m,N)$, $k\in Z^+$, and $\gamma\in\Gamma_k$, $i\in\bar{k}$, we define $C(\gamma,i)=C_{\gamma(k)}C_{\gamma(k-1)}\cdots C_{\gamma(i)}$, and $C(\gamma)$ will denote $C(\gamma,1)$. We define the controllability matrix

$$P(L,\gamma)=[D_{\gamma(k)},C(\gamma,k)D_{\gamma(k-1)},C(\gamma,k-1)D_{\gamma(k-2)},\cdots,C(\gamma,2)D_{\gamma(1)}].$$

As observed in [3], the column space of $P(L,\gamma)$ is the set of all $TP(L,0,u,\gamma)$ with $u\in U_k$. We denote this space by $S(L,\gamma)$. The *controllable set* of L is the set $S(L)$ defined by

$$S(L)=\bigcup\{S(L,\gamma)|\gamma\in\Gamma\}.$$

$S(L)$ is thus the set of all points in R^n reachable from 0 in finitely many steps. In [3], we investigated the case $L\in\Sigma(n,m,N)$ and $S(L)=R^n$ (i.e., $L\in\Phi(n,m,N)$). Whenever $S(L)=R^n$, we call L *completely controllable*, and we write $L\in\Phi^0(n,m,N)$.

For $L\in\Sigma^0(n, m, N)$, the controllability set, $S(L)$, is not, in general, a subspace. If, for example,

$$L=\left\{\left(\begin{bmatrix}1&1&0\\0&2&0\\0&0&0\end{bmatrix},\begin{bmatrix}0\\1\\0\end{bmatrix}\right),\left(\begin{bmatrix}0&0&0\\0&0&0\\0&0&1\end{bmatrix},\begin{bmatrix}0\\0\\1\end{bmatrix}\right)\right\},$$

we obtain $S(L)=\operatorname{span}\{e_1, e_2\}\cup\operatorname{span}\{e_3\}$, where the e_i's represent standard basis vectors in R^3.

The basic result of this section is that if $L\in\Sigma(n, m, N)$, then $S(L)$ is a subspace of R^n. We first note that $C_iS(L)\subset S(L)$ for $i\in\bar{N}$, since if x is reachable from 0, then $C_ix=C_ix+D_i0$ is reachable from 0.

For $\alpha\in\Gamma_h$ and $\beta\in\Gamma_k$, $(\beta, \alpha)\in\Gamma_{h+k}$ is defined by

$$(\beta, \alpha)(i)=\begin{cases}\beta(i), & i\leqq i\leqq k,\\ \alpha(i-k), & k+1\leqq i\leqq h+k.\end{cases}$$

THEOREM 1. *If $L=\{(C_i, D_i)\}_{i=1}^N\in\Sigma^0(n, m, N)$, with each C_i nonsingular, then $S(L)$ is a subspace of R^n.*

Proof. Suppose $L=\{(C_i, D_i)\}_{i=1}^N\in\Sigma^0(n, m, N)$, with each C_i nonsingular. Let $r=\max\{\operatorname{rank}(P(L, \gamma))|\gamma\in\Gamma\}$, and let $\alpha\in\Gamma$ such that $\operatorname{rank}(P(L, \alpha))=r$.

For any $\gamma\in\Gamma$,

$$\operatorname{rank}([P(L, \alpha), C(\alpha)P(L, \gamma)])=\operatorname{rank}(P(L, (\gamma, \alpha)))=r,$$

so that $C(\alpha)S(L, \gamma)\subset S(L, \alpha)$, and $S(L, \gamma)\subset C(\alpha)^{-1}S(L, \alpha)$. In particular, $S(L, \alpha)\subset C(\alpha)^{-1}S(L, \alpha)$. Since $C(\alpha)^{-1}$ preserves dimension, $S(L, \alpha)=C(\alpha)^{-1}S(L, \alpha)$. Thus we have $S(L, \gamma)\subset S(L, \alpha)$ for all $\gamma\in\Gamma$. Hence $S(L)=S(L, \alpha)$ is a subspace of R^n. □

If $L\in\Sigma^0(n, m, N)$ and if, for each $i\in\bar{N}$, an $m\times n$ feedback matrix F_i is selected, we may select controls u_k of the form, $u_k=F_ix_k+v_k$, to produce the system

$$x_{k+1}=(C_i+D_iF_i)x_k+D_iv_k.$$

This system corresponds to the sequence $\{(C_i+D_iF_i, D_i)\}_{i=1}^N$, which we will denote by $L(F)$, where $F=\{F_1, F_2, \cdots, F_N\}$. It will be convenient to define

$$\Lambda_N=\{F=\{F_i\}_{i=1}^N|F_i \text{ a real } m\times n \text{ matrix}\}.$$

In [3], we saw that the application of feedback to a system in $\Sigma(n, m, N)$ may produce desirable properties, although the resulting system may not be in $\Sigma(n, m, N)$. Hence the study of systems in the more general $\Sigma^0(n, m, N)$ may be useful in handling the multi-rate sampled-data system. We now show that the use of feedbacks does not affect the controllability set of a system. Our proof, which follows Wonham [6], employs the notation

$$\hat{D}_i=\text{column space of } D_i.$$

THEOREM 2. *If $L\in\Sigma^0(n, m, N)$, $F\in\Lambda_N$, and $\gamma\in\Gamma$, then $S(L(F), \gamma)=S(L, \gamma)$.*

Proof. Let $L=\{(C_i, D_i)\}_{i=1}^N\in\Sigma^0(n, m, N)$, and $F\in\Lambda_N$. For $i\in\bar{N}$, let $H_i=C_i+D_iF_i$. Then for V a subspace of R^n, we have

$$H_iV=C_iV+D_iF_iV\subset C_iV+\hat{D}_i. \tag{1}$$

Thus, for $k\in Z^+$ and $\gamma\in\Gamma_k$,

$$\begin{aligned}S(L(F), \gamma)&=\hat{D}_{\gamma(k)}+H(\gamma, k)\hat{D}_{\gamma(k-1)}+\cdots+H(\gamma, 2)\hat{D}_{\gamma(1)}\\&=\hat{D}_{\gamma(k)}+H_{\gamma(k)}(\hat{D}_{\gamma(k-1)}+H_{\gamma(k-1)}\\&\quad\cdot(\hat{D}_{\gamma(k-2)}+\cdots+H_{\gamma(3)}(\hat{D}_{\gamma(2)}+H_{\gamma(2)}\hat{D}_{\gamma(1)}))\cdots).\end{aligned}$$

Using (1), we obtain

$$(2)\qquad S(L(F),\gamma)\subset \hat{D}_{\gamma(k)}+C(\gamma,k)\hat{D}_{\gamma(k-1)}+\cdots+C(\gamma,2)\hat{D}_{\gamma(1)}=S(L,\gamma).$$

Now (2) holds for any system $L\in\Sigma^0(n,m,N)$, and any $F\in\Lambda_N$. Hence

$$S(L,\gamma)=S(L(F)(-F),\gamma)\subset S(L(F),\gamma).$$

Thus $S(L(F),\gamma)=S(L,\gamma)$. □

Much of our work is valid for all systems L for which $S(L)$ is a subspace of R^n. We define

$$\Sigma^*(n,m,N)=\{L\in\Sigma^0(n,m,N)|S(L)\text{ is a subspace of }R^n\}.$$

Theorem 2 says that if $L\in\Sigma^*(n,m,N)$ (as in the case of multi-rate sampling), and $F\in\Lambda_N$, then $L(F)\in\Sigma^*(n,m,N)$. We have the following corollary.

COROLLARY. *If $L\in\Sigma^0(n,m,N)$, then $S(L(F))=S(L)$ for each $F\in\Lambda_N$. Moreover, if $L\in\Sigma^*(n,m,N)$, then there is $\gamma\in\Gamma$ such that*

$$S(L(F),\gamma)=S(L(F))=S(L)=S(L,\gamma).$$

Proof. Suppose $L\in\Sigma^0(n,m,N)$. Then

$$S(L(F))=\bigcup\{S(L(F),\alpha)|\alpha\in\Gamma\}=\bigcup\{S(L,\alpha)|\alpha\in\Gamma\}=S(L).$$

Now suppose $L\in\Sigma^*(n,m,N)$. If $\dim S(L,\gamma)<\dim S(L)$ for all $\gamma\in\Gamma$, then, since Γ is countable and proper subspaces are nowhere dense in R^n, Baire's category theorem (see [7]) is contradicted. It follows that there exists $\gamma\in\Gamma$ with $S(L(F),\gamma)=S(L,\gamma)=S(L)=S(L(F))$. □

For the system

$$L=\left\{\left(\begin{bmatrix}1&0&0\\0&1&0\\0&0&0\end{bmatrix},\begin{bmatrix}1\\0\\0\end{bmatrix}\right),\left(\begin{bmatrix}0&0&0\\0&1&0\\0&0&1\end{bmatrix},\begin{bmatrix}0\\1\\0\end{bmatrix}\right)\right\},$$

$S(L)=\operatorname{span}\{e_1,e_2\}$, and so $L\in\Sigma^*(3,1,2)$. However, L cannot be obtained by adjoining feedbacks to a member of $\Sigma(3,1,2)$.

For $L\in\Sigma^0(n,m,N)$, trajectories under L are governed by the following theorem.

THEOREM 3. *Let $L\in\Sigma^0(n,m,N)$. Suppose x and y are in R^n. Then there is a trajectory from x to y if and only if there is $k\in Z^+$ and $\alpha\in\Gamma_k$ such that $y-C(\alpha)x\in S(L)$. If so, there is $u\in U_k$ such that $y=TP(L,x,u,\alpha)$.*

Proof. For any $\alpha\in\Gamma_k$, $u\in U_k$, and $x\in R^n$,

$$T_{i+1}(L,x,u,\alpha)=C_{\alpha(i)}C_{\alpha(i-1)}\cdots C_\alpha(1)x+T_{i+1}(L,0,u,\alpha).$$

Thus $TP(L,x,u,\alpha)=C(\alpha)x+TP(L,0,u,\alpha)$, and the theorem follows. □

COROLLARY. *If $L\in\Sigma^*(n,m,N)$ and $x,y\in S(L)=S(L,\gamma)$ with $\gamma\in\Gamma_k$, then there exists $u\in U_k$ such that $y=TP(L,x,u,\gamma)$.*

Proof. The corollary follows since $C(\gamma)S(L)\subset S(L)$. □

If a single pair (C_iD_i) is selected from a system $L\in\Sigma^0(n,m,N)$, it is clear that $S(L)$ contains each member of the controllable space of that pair; i.e., each member of

$$S_i(L)=\text{column space of }[D_i,C_iD_i,\cdots,C_i^{n-1}D_i].$$

Thus if $L \in \Sigma^*(n, m, N)$, $S(L) \supset S_1(L)+S_2(L)+\cdots+S_N(L)$. The following is an example where this inclusion is proper.

$$L=\left\{\left(\begin{bmatrix}1&0&0\\0&0&0\\0&1&0\end{bmatrix},\begin{bmatrix}1\\0\\0\end{bmatrix}\right),\left(\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix},\begin{bmatrix}0\\1\\0\end{bmatrix}\right)\right\},$$

and so

$$S(L)=R^3 \neq S_1(L)+S_2(L)=\text{span}\,\{e_1, e_2\}.$$

For $L \in \Sigma^0(n, m, N)$, $S(L)=\bigcup\{S(L, \gamma)|\gamma \in \Gamma\}$ is the union of countably many subspaces of R^n. In every example we have investigated, $S(L)$ is in fact the union of N or fewer subspaces of R^n. We suspect that this is the case in general, but we have been unable to prove that $S(L)$ is always the union of finitely many subspaces of R^n. We have, however, related this property to that of possessing a bound for the number of steps required to reach members of $S(L)$ from 0.

DEFINITION. For $L \in \Sigma^0(n, m, N)$, the *bound of* L, if it exists, is the number

$$b(L)=\min\,\{k|x \in S(L) \text{ implies there is } \gamma \in \Gamma_i, \text{ for } i \in \bar{k}, \text{ such that } x \in S(L, \gamma)\}.$$

These bounds were used extensively in [3], where we saw, for example, that if $n \leqq 3$ and $L \in \Sigma(n, m, N)$ is completely controllable, then $b(L)$ exists and is not more than 4.

THEOREM 4. *Let $L \in \Sigma^0(n, m, N)$. Then L has a bound $b(L)$ if and only if $S(L)$ is the union of finitely many subspaces of R^n.*

Proof. Suppose $b(L)$ exists. Then $S(L)=\bigcup\{S(L, \gamma)|\gamma \in \Gamma_i \text{ and } i \in \overline{b(L)}\}$ is a finite union of subspaces.

Now suppose $S(L)=\bigcup\{W_j|j=1, 2, \cdots, r\}$ with each W_j a subspace of R^n. Then for each $j \in \bar{r}$,

$$W_j=\bigcup\{W_j \cap S(L, \gamma)|\gamma \in \Gamma\}.$$

Again invoking the Baire category theorem, there is a $k_j \in Z^+$ and a $\gamma^{(j)}$ in Γ_{k_j} such that $W_j=W_j \cap S(L, \gamma^{(j)})$. Hence

$$\bigcup\,\{S(L, \gamma^{(j)})|j \in \bar{r}\} \supset \bigcup\{W_j|j \in \bar{r}\}=S(L),$$

and so $b(L)$ exists and is not larger than $\max\,\{k_j|j \in \bar{r}\}$. □

3. Controllability canonical form. In this section we adapt the controllability canonical form (see [1]) to multi-pair systems. We describe a condition on L which is necessary and sufficient that feedbacks F exist for which $L(F)$ has a controllability canonical form which is block diagonal.

DEFINITION. Suppose $L=\{(C_i, D_i)\}_{i=1}^N \in \Sigma^0(n, m, N)$, G is $n \times n$ and nonsingular, and $J=\{J_1, J_2, \cdots, J_N\}$, where each J_i is $m \times m$ and nonsingular. Then

$$L_{G,J} \text{ denotes the system } \{(GC_iG^{-1}, GD_iJ_i)\}_{i=1}^N.$$

L_G denotes $L_{G,J}$ with each $J_i=I$.

THEOREM 5. *If $L \in \Sigma^0(n, m, N)$, then $S(L_{G,J})=GS(L)$ and, for $L \in \Sigma^*(n, m, N)$, $b(L_{G,J})=b(L)$.*

Proof. For $k \in Z^+$ and $\gamma \in \Gamma_k$, $P(L_{G,J}, \gamma)=GP(L, \gamma)J_\gamma$, where $J_\gamma=\text{diag}\,\{J_{\gamma(k)}, J_{\gamma(k-1)}, \cdots, J_{\gamma(1)}\}$. Thus $S(L_{G,J}, \gamma)=GS(L, \gamma)$, and the theorem follows. □

THEOREM 6. *If $L \in \Sigma^*(n, m, N)$ with* $\dim S(L)=r<n$, *then there is a nonsingular*

G such that

$$L_G=\left\{\left(\begin{bmatrix} C_{i1} & C_{i3} \\ 0 & C_{i2}\end{bmatrix}, \begin{bmatrix} D_{i1} \\ 0\end{bmatrix}\right)\right\}_{i=1}^N,$$

with each C_{i1} $r\times r$ and each D_{i1} $r\times m$. The system $(L_G)_1=\{(C_{i1}, D_{i1})\}_{i=1}^N$ is completely controllable.

Proof. Let $z_1, z_2, \cdots, z_r$ be a basis for $S(L)$ and extend to a basis $Z=\{z_1, z_2, \cdots, z_n\}$ of R^n. If G is the matrix of transition from the standard basis of R^n to Z, then, since $C_iS(L)\subset S(L)$ and $D_i\subset S(L)$ for $i\in\bar{N}$, L_G has the required form. Clearly

$$S(L_G)=GS(L)=\left\{\begin{bmatrix} x \\ 0\end{bmatrix}\middle| x\in R^r\right\},$$

and the complete controllability of $(L_G)_1$ follows. □

Definition. $L=\{(C_i, D_i)\}_{i=1}^N\in\Sigma^*(n, m, N)$, with $\dim S(L)=r$, is in *controllability canonical form* (*CCF*) provided

$$C_i=\begin{bmatrix} C_{i1} & C_{i3} \\ 0 & C_{i2}\end{bmatrix} \quad\text{and}\quad D_i=\begin{bmatrix} D_{i1} \\ 0\end{bmatrix},$$

with C_{i1} $r\times r$ and D_{i1} $r\times m$ for each $i\in\bar{N}$.

To characterize those systems which have a block diagonal CCF, we introduce the concept of L-invariance. The idea is an extension of "(A, B)-invariance" [6].

Definition. Let V be a subspace of R^n and $L=\{(C_i, D_i)\}_{i=1}^N\in\Sigma^*(n, m, N)$. V is *L-invariant* provided $C_iV\subset V+\hat{D}_i$ for each $i\in\bar{N}$.

We note that any subspace invariant under each C_i is L-invariant, and so in particular $S(L)$ is L-invariant. In the following example we exhibit a system L and a subspace V which is L-invariant, but which is not invariant under any of the C_i's.

$$L=\left\{\left(\begin{bmatrix} -9 & 20 & 6 \\ -5 & 11 & 3 \\ 0 & 0 & 1\end{bmatrix}, \begin{bmatrix} 2 \\ 1 \\ 0\end{bmatrix}\right), \left(\begin{bmatrix} -4 & 5 & 1 \\ -1 & -1 & -1 \\ -5 & 15 & 6\end{bmatrix}, \begin{bmatrix} 2 \\ 1 \\ 0\end{bmatrix}\right)\right\}\in\Sigma(3, 1, 2),$$

and

$$V=\operatorname{span}\left\{\begin{bmatrix} 1 \\ 1 \\ -2\end{bmatrix}\right\}.$$

The following theorem is a repeated application of Lemma 4.2 in [6], and we do not prove it.

Theorem 7. *Let $L\in\Sigma^*(n, m, N)$ and V a subspace of R^n. Then V is L-invariant if and only if there is an $F\in\Lambda_N$ such that V is invariant under each $C_i+D_iF_i$.*

Theorem 8. *Let $L\in\Sigma^*(n, m, N)$. The following are equivalent:*

(1) *There is $F\in\Lambda_N$ such that $L(F)$ has a block diagonal* CCF.

(2) *There is an L-invariant subspace V with $R^n=S(L)\oplus V$.*

Proof. Let $L=\{(C_i, D_i)\}_{i=1}^N$ with $\dim S(L)=r$.

Suppose (1). Let $F\in\Lambda_N$ and G nonsingular, such that

$$L(F)_G=\left\{\left(\begin{bmatrix} C_{i1} & 0 \\ 0 & C_{i2}\end{bmatrix}, \begin{bmatrix} D_{i1} \\ 0\end{bmatrix}\right)\right\}_{i=1}^N \quad\text{is in CCF.}$$

Clearly,

$$V=G^{-1}\left\{\begin{bmatrix}0\\y\end{bmatrix}\middle| y\in R^{n-r}\right\}$$

is invariant under each $C_i+D_iF_i$ and $S(L(F))\oplus V=R^n$. Thus by Theorems 2 and 7, $S(L)\oplus V=R^n$ and V is L-invariant.

Now suppose (2). Choose basis $Z=\{z_1,\cdots,z_r,\cdots,z_n\}$ of R^n such that $S(L)=\operatorname{span}\{z_1,\cdots,z_r\}$ and $V=\operatorname{span}\{z_{r+1},\cdots,z_n\}$. Choose $F\in\Lambda_N$ so that V is invariant under each $C_i+D_iF_i$. If G is the matrix of transition from the standard basis of R^n to Z, then $L(F)_G$ is in block diagonal CCF. □

In the example preceding Theorem 7,

$$S(L)=\operatorname{span}\left\{\begin{bmatrix}2\\1\\0\end{bmatrix},\begin{bmatrix}3\\3\\-5\end{bmatrix}\right\},$$

so that $S(L)\oplus V=R^3$. If we choose

$$F_1=F_2=[5\quad -10\quad -3],$$

and

$$G=\begin{bmatrix}1&-1&0\\2&-4&-1\\5&10&3\end{bmatrix},$$

then

$$L(F)_G=\left\{\left(\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix},\begin{bmatrix}1\\0\\0\end{bmatrix}\right),\left(\begin{bmatrix}0&1&0\\1&0&0\\0&0&1\end{bmatrix},\begin{bmatrix}1\\0\\0\end{bmatrix}\right)\right\}$$

is in block diagonal CCF.

4. State deadbeat response. In [3] it was shown that any completely controllable system in $\Sigma(n, m, N)$, with $n\leqq 3$, is capable through feedback of a state deadbeat response, and it is conjectured that the result holds for arbitrary n. In this section, we investigate systems in $\Sigma^*(n, m, N)$ which are capable of a state deadbeat response on subspaces of R^n. Our results include a converse to the conjecture mentioned above.

DEFINITION. Let $L\in\Sigma^0(n, m, N)$ and V be a subspace of R^n. Then L has *state deadbeat response on* V ($L\in$ SDR *on* V) provided there is $F\in\Lambda_N$ such that $L(F)=\{(H_i, D_i)\}_{i=1}^N$ has the property that for each $x\in V$, there is $\gamma\in\Gamma$ with $H(\gamma)x=0$. If $V=R^n$, we say $L\in$ SDR.

THEOREM 9. *If* $L\in$ SDR *on* V, *then there is an* $F\in\Lambda_N$ *and a single* $\gamma\in\Gamma$ *such that* $L(F)=\{(H_i, D_i)\}_{i=1}^N$ *satisfies* $H(\gamma)x=0$ *for all* $x\in V$.

Proof. Let $F\in\Lambda_N$ such that $L(F)=\{(H_i, D_i)\}_{i=1}^N$ has the property that for each $x\in V$ there is a $\gamma\in\Gamma$ with $H(\gamma)x=0$. Then

$$V=\bigcup_{\alpha\in\Gamma}(V\cap \operatorname{NS}(H(\alpha))),$$

where NS(A) denotes the null space of A. Suppose $\dim(V\cap\operatorname{NS}(H(\alpha)))<\dim V$ for each $\alpha\in\Gamma$. Then since Γ is countable, Baire's category theorem is contradicted. It follows that there is $\gamma\in\Gamma$ such that $V=V\cap\operatorname{NS}(H(\gamma))$. □

DEFINITION. If $\gamma\in\Gamma$ satisfies the conditions given in Theorem 9, we say $L\in \mathrm{SDR}_\gamma$ *on* V (or $L\in\mathrm{SDR}_\gamma$, if $V=R^n$).

Clearly, $L\in\mathrm{SDR}_\gamma$ on V if and only if $L(F)\in\mathrm{SDR}_\gamma$ on V for all $F\in\Lambda_N$. Also, by extension of Theorem 9 in [3], $L\in\mathrm{SDR}_\gamma$ on V if and only if $L_{G,J}\in\mathrm{SDR}_\gamma$ on GV.

The following theorem shows that, for $L\in\Sigma^*(n, m, N)$ and $L\in\mathrm{SDR}_\gamma$ on $S(L)$, the terminal point of a trajectory under γ is independent of the initial point in $S(L)$. If the conjecture referred to at the beginning of the section is correct, then it will follow from Theorem 11 that every system in $\Sigma(n, m, N)$ has this property.

THEOREM 10. *Suppose* $L\in\Sigma^*(n, m, N)$ *and* $\gamma\in\Gamma_k$. *Then* $L\in\mathrm{SDR}_\gamma$ *on* $S(L)$ *if and only if there is* $F\in\Lambda_N$ *such that, for each* $u\in U_k$,

$$TP(L(F), x, u, \gamma)=TP(L(F), 0, u, \gamma)\quad \textit{for each } x\in S(L).$$

Proof. The theorem follows, since $TP(L(F), x, u, \gamma)=H(\gamma)x+TP(L(F), 0, u, \gamma)$, as shown in the proof of Theorem 3. □

We observe that if F is selected according to Theorem 10, then, for $z\in R^n$ and $u\in U_k$, $TP(L(F), w, u, \gamma)$ is independent of w in $z+S(L)$. Moreover, if $S(L)=S(L,\gamma)$, then

$$H(\gamma)z+S(L)=\{TP(L(F), z, u, \gamma)|u\in U_k\}.$$

Finally, if in addition $L\in\Sigma(n, m, N)$ and $u\in U_k$, then $TP(L(F), w, u, \gamma)\in H(\gamma)z+S(L)$ if and only if $w\in z+S(L)$.

We may now apply the controllability canonical form for multi-pair systems to obtain the following theorems. Proofs may be found in [4].

THEOREM 11. *Suppose* $L\in\Sigma^*(n, m, N)$ *and* L_G *is in* CCF. *Then for* $\gamma\in\Gamma$, $L\in \mathrm{SDR}_\gamma$ *on* $S(L)$ *if and only if* $(L_G)_1\in\mathrm{SDR}_\gamma$.

THEOREM 12. *Suppose* $L\in\Sigma^*(n, m, N)$ *with* $\dim S(L)<n$, *and*

$$L_G=\left\{\left(\begin{bmatrix} C_{i1} & C_{i3}\\ 0 & C_{i2}\end{bmatrix},\begin{bmatrix} D_{i1}\\ 0\end{bmatrix}\right)\right\}_{i=1}^{N}\quad \textit{is in } \mathrm{CCF}.$$

If $L\in\mathrm{SDR}_\gamma$ *for some* $\gamma\in\Gamma_k$, *then* $\prod_{j=k}^{1} C_{\gamma(j)2}=0$. *Furthermore, if* $\prod_{j=h}^{1} C_{\beta(j)2}=0$ *for some* $\beta\in\Gamma_h$ *and* $L\in\mathrm{SDR}_\alpha$ *on* $S(L)$ *for some* $\alpha\in\Gamma_k$, *then* $L\in\mathrm{SDR}_\gamma$, *where* $\gamma=(\beta,\alpha)$.

THEOREM 13. *Suppose* $L\in\Sigma(n, m, N)$. *Then* $L\in\mathrm{SDR}_\gamma$ *on* V *implies* $V\subset S(L)$.

COROLLARY. *Suppose* $L\in\Sigma(n, m, N)$. *If* $L\in\mathrm{SDR}$, *then* L *is completely controllable.*

This corollary, in conjunction with the result in [3], gives the following:

COROLLARY. *Let* $L\in\Sigma(n, m, N)$, *with* $n\leqq 3$. *Then* L *is completely controllable if and only if* $L\in\mathrm{SDR}$.

5. Stabilizability. The stabilizability of multi-pair systems was investigated in [2]. In this section we extend some of the concepts introduced there in order to describe the action of the system on a subspace of R^n. Using the controllability canonical form, we relate the properties of the system on $S(L)$ to those of the system on all of R^n.

DEFINITION. A set $\{H_1, H_2, \cdots, H_N\}$ of real $n\times n$ matrices is *convergent on the subspace* V *of* R^n provided that, for each $x\in V$, there is a sequence $\{p_i\}_{i=1}^{\infty}$ with each $p_i\in\bar{N}$ such that

$$\lim_{k\to\infty}\left(\prod_{i=k}^{1} H_{p_i}\right)x=0.$$

A system $L\in\Sigma^0(n, m, N)$ is *convergible on* V provided there is an $F\in\Lambda_N$ such that $L(F)=\{(H_i, D_i)\}_{i=1}^{N}$ satisfies the property that $\{H_1, H_2, \cdots, H_N\}$ is convergent on V. When $V=R^n$, we say $\{H_1, H_2, \cdots, H_N\}$ is *convergent*, and L is *convergible*.

Clearly, if L is convergible on V, then $L(F)$ is convergible on V for all $F \in \Lambda_N$. Also the following theorem is easily verified.

THEOREM 14. *Suppose $L \in \Sigma^0(n, m, N)$. For any V, G, and $J = \{J_1, J_2, \cdots, J_N\}$, L is convergible on V if and only if $L_{G,J}$ is convergible on GV.*

Again applying the CCF we obtain the following theorems.

THEOREM 15. *Suppose $L \in \Sigma^*(n, m, N)$ and L_G is in* CCF. *Then L is convergible on $S(L)$ if and only if $(L_G)_1$ is convergible.*

THEOREM 16. *Suppose $L \in \Sigma^*(n, m, N)$ with* $\dim S(L) = r < n$, *and*

$$L_G = \left\{\left(\begin{bmatrix} C_{i1} & C_{i3} \\ 0 & C_{i2} \end{bmatrix}, \begin{bmatrix} D_{i1} \\ 0 \end{bmatrix}\right)\right\}_{i=1}^{N} \quad \textit{is in } \text{CCF}.$$

If L is convergible, then the C_{i2}'s are convergent. Furthermore, if some finite product of the C_{i2}'s is zero and L is convergible on $S(L)$, then L is convergible.

Theorems 14, 15, and 16 remain valid when "convergent" and "convergible" are replaced by "exponentially convergent" and "exponentially convergible." The relevant definitions are these:

DEFINITION. A set $\{H_1, H_2, \cdots, H_N\}$ of real $n \times n$ matrices is *exponentially convergent* on a subspace V of R^n provided there is a $B > 0$ and an $\alpha \in (0, 1)$ such that, for each $x \in V$, there is a sequence $\{p_i\}_{i=1}^{\infty}$ with each $p_i \in \bar{N}$ satisfying

$$\left\|\left(\prod_{i=k}^{1} H_{p_i}\right)x\right\| < B\alpha^k \|x\| \quad \text{for all } k \in Z^+.$$

A system $L \in \Sigma^0(n, m, N)$ is *exponentially convergible* provided there is an $F \in \Lambda_N$ such that $L(F) = \{(H_i, D_i)\}_{i=1}^{N}$ satisfies the property that $\{H_1, H_2, \cdots, H_N\}$ is exponentially convergent.

Of course the equivalence of norms in R^n makes the property of exponential convergence norm-independent.

We remark that the notion of contractiveness of a set of matrices, introduced in [2], is relevant here. $\{H_1, \cdots, H_N\}$ is contractive relative to a norm $\|\cdot\|$ on R^n, provided that each nonzero x in R^n satisfies $\|H_i x\| < \|x\|$ for some $i \in \bar{N}$. Contractiveness is norm-dependent, but contractiveness relative to any norm implies exponential convergence. The following analogues to Theorems 14 and 15 can be easily verified.

THEOREM 17. *Suppose $L \in \Sigma^0(n, m, N)$ and $\|\cdot\|$ is a norm on R^n. For any V, G, and $J = \{J_1, \cdots, J_N\}$, L is contractible on V relative to $\|\cdot\|$ if and only if $L_{G,J}$ is contractible on GV relative to $\|\cdot\|_{G^{-1}}$, where $\|x\|_{G^{-1}} = \|G^{-1}x\|$ for each $x \in R^n$.*

THEOREM 18. *Suppose $L \in \Sigma^*(n, m, N)$ with* $\dim S(L) = r < n$, *and L_G is in* CCF. *Then the following are equivalent*:

(1) *There is a norm $\|\cdot\|$ on R^n such that L is contractible on $S(L)$ relative to $\|\cdot\|$.*

(2) *There is a norm $\|\cdot\|_1$ on R^r such that $(L_G)_1$ is contractible relative to $\|\cdot\|_1$.*

It would be useful to explore further the convergibility of multi-pair systems. It can be shown that every completely controllable system in $\Sigma(2, 1, N)$ is contractible relative to some norm, and thus is exponentially convergible. We suspect this can be made more general. Finally we list some other basic questions concerning convergibility which we have not yet resolved.

Let H denote a set $\{H_1, H_2, \cdots, H_N\}$ of real $n \times n$ matrices.

(1) Does exponential convergence of H imply that H is contractive relative to some norm on R^n?

(2) Does convergence of H imply the existence of a single sequence $\{p_i\}_{i=1}^{\infty}$ from $\bar{N}$

such that

$$\lim_{k\to\infty}\left(\prod_{i=k}^{1} H_{p_i}\right)=0?$$

(3) Does contractiveness of H relative to some norm on R^n imply the existence of a finite product A of the H_i's such that the spectral radius of A is less than 1?

It is not difficult to see that an affirmative answer to (2) implies a like answer to (3). The converse is also true, for suppose (3) has an affirmative answer. Let $H = \{H_1, H_2, \cdots, H_N\}$ be convergent. Then H is pre-contractive (see [2]), and it follows that there is a finite set $\{A_1, \cdots, A_M\}$ of finite products of the H_i's which is contractive relative to the Euclidean norm. Hence by (3), some finite product of the A_i's, and hence of the H_i's, has spectral radius less than 1.

REFERENCES

[1] H. Kwakernaak and R. Sivan, *Linear Optimal Control Systems*, John Wiley and Sons, New York, 1972.

[2] D. P. Stanford, *Stability for a multi-rate sampled-data system*, this Journal, 17 (1979), pp. 390–399.

[3] D. P. Stanford and L. T. Conner, Jr., *Controllability in Multi-Rate Sampled-Data Systems*, Final Report: Task Order NAS1-14972-9, 1978.

[4] ———, *Controllability and Stabilizability in Multi-pair Systems*, Final Report: Task Order NAS1-14972-24, 1979.

[5] V. I. Utkin, *Variable structure systems with sliding modes*, IEEE Trans. Automatic Control, AC-22 (1977), pp. 212–222.

[6] W. M. Wonham, *Linear Multivariable Control*, Springer-Verlag, New York, 1974.

[7] A. C. Zaanen, *Linear Analysis*, North-Holland Publishing Co., Amsterdam, 1960.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 5, September 1980

HARDY SPACES OF ANALYTIC FUNCTIONS AND FEEDBACK STABILITY OF SCALAR CONVOLUTION SYSTEMS*

S. MOSSAHEB†

Abstract. By use of the theory of H^p-spaces of analytic functions, stability of single input single output convolution systems under constant gain feedback is studied. Various L^p-stability results are discussed. These are considerable generalizations of the known results on stable systems. It is shown that corresponding to the classical Nyquist diagram the gain space may be partitioned into mutually disjoint sets, on each of which the stability of the closed-loop system is independent of the gain. It is proved how some seemingly different systems fall into the category of those which are considered in this paper. Finally, the extension of some of the above stability results to more general systems is pointed out.

1. Introduction. A particular graphical test has been developed by Callier and Desoer [3] for the closed-loop stability under constant gain feedback of delayed systems whose transfer functions are the sum of rational functions and the Laplace transforms of a series of delays plus those of integrable functions. The derivation of this test is based on some deep properties of almost periodic functions. As is apparent from the pioneering work of these authors, the behavior of the almost periodic part of the transfer function makes it practically impossible to consider simple encirclement criteria as a means of checking stability and instead they give a remarkable test to this end [4]. This test is expressed in terms of the argument of the return difference operator, and depends on the particular value of the feedback gain. In this paper we give a different test of stability based on the amplitude of the return difference operator, and also complete some of the work of Desoer and Wu on the L^p-stability of the systems under consideration. We shall also show that having considered one value of the feedback gain k one can obtain a maximal interval I containing k such that for any k' in I the stability of the closed-loop system for k' is the same as that for k. We then give a useful result which enables one to reduce some seemingly different systems, e.g. those whose transfer functions contain the ratio of two exponential polynomials, to those which we discuss in this paper. The extension of some of the above work to more general systems is pointed out in an appendix.

The research reported here was carried out under the supervision of Professor A. G. J. MacFarlane at Cambridge. I am greatly indebted to him for his constant help and advice. It is also my pleasure to thank Professor C. A. Desoer for pointing out an error in the original version of this work. Finally, I thank the referee for drawing my attention to some of the work of Professor F. M. Callier which precedes and is related to Theorem 6 below (see the remarks at the end of this theorem), and for his detailed suggestions which have lead to a clearer presentation of this paper.

2. Background materials and notations

2.1. Unless it is clear from the context all functions, measures, etc. are real. The Laplace transform of a distribution T is denoted by $\hat{T}$. For $\sigma \geqq 0$ we write $R(\sigma) = \{s: \operatorname{Re} s > \sigma\}$, $\bar{R}(\sigma) = \{s: \operatorname{Re} s \geqq \sigma\}$, $D = \{z: |z| < 1\}$, $\bar{D} = \{z: |z| \leqq 1\}$, $\bar{D}_1 = \bar{D} \backslash \{1\}$, $C = \{z: |z| = 1\}$, $C_1 = C \backslash \{1\}$. $R(0)$ and $\bar{R}(0)$ are abbreviated to R and $\bar{R}$.

2.2. For $\sigma \geqq 0$ we shall write $A(\sigma)$ for the algebra, under convolution, of locally bounded measures on $[0, \infty)$ which are of the form $dm(t) = f(t) + \sum_{n=0}^{\infty} a_n \delta(t - t_n)$,

* Received by the editors March 26, 1979, and in final revised form January 16, 1980.

† Postgraduate School of Studies in Control Engineering, University of Bradford, Bradford BD7 1DP, West Yorkshire, England.

where $0=t_0<t_1<\cdots$, and $\|dm\|=\int_0^\infty e^{-\sigma t}|f(t)|\,dt+\sum_{n=0}^\infty |a_n|\,e^{-\sigma t_n}<\infty$. Then for Re $s\geqq\sigma$ we have $d\hat{m}(s)=\hat{f}(s)+\sum_{n=0}^\infty a_n\,e^{-st_n}$. $d\hat{m}(s)$ is analytic on $R(\sigma)$, bounded and continuous on $\bar{R}(\sigma)$ and $\lim_{|s|\to\infty}|\hat{f}(s)|=0$, while $\lim_{\mathrm{Re}\, s\to\infty}\sum a_n\,e^{-st_n}=a_0$.

We write A in place of $A(0)$. The following theorem shows when an element of $A(\sigma)$ is invertible [8, pp. 141–150].

THEOREM. *$dm\in A(\sigma)$ is invertible in $A(\sigma)$ if and only if* $\inf_{\bar{R}(\sigma)}|d\hat{m}(s)|>0$.

2.3. We write H^2 for the space of all analytic functions f on R such that $\sup_{x>0}\left(\int_{-\infty}^\infty |f(x+jy)|^2\,dy\right)^{1/2}=\|f\|_{H^2}<\infty$. H^2 has the following properties [9, ch. 8]. $f\in H^2$ if and only if there exists $\phi\in L^2(0,\infty)$ such that $f(s)=\hat{\phi}(s)$ (this fundamental fact is due to Paley and Wiener). If $f\in H^2$ then the boundary values $f(jy)$ exist almost everywhere, $f(jy)\in L^2(-\infty,\infty)$, $\|f\|_{H^2}=\|f(jy)\|_{L^2}$, and for $x>0$ we have

$$f(x+jy)=\frac{1}{\pi}\int_{-\infty}^{\infty} f(jw)\frac{x}{x^2+(y-w)^2}\,dw.$$

If $f(jy)\in L^2(-\infty,\infty)$ is real then the above integral defines $f(x+jy)$ as the real part of an element of H^2.

2.4. We write $H^\infty(D)$ for the space of all analytic functions f on D such that $\sup_D|f(z)|=\|f\|_{H^\infty}<\infty$. $H^\infty(D)$ has the following properties [9, Ch. 5]. For $f\in H^\infty(D)$, the boundary values $f(e^{j\phi})$ exist almost everywhere and $\log|f(e^{j\phi})|$ is integrable on $[-\pi,\pi]$. Let $\{\xi_n\}$ be the set of zeros of f in D, $\xi_n\neq 0$, each of order m_n and let m_0 be the order of the zero of f at the origin. Then

$$f(z)=\rho B(z)F(z)S(z),\quad \text{where } \rho=\exp(ja)\quad \text{with } a=\arg(f/B)(0),$$

$$B(z)=z^{m_0}\prod_{n=1}^{\infty}\left[\frac{\bar{\xi}_n}{|\xi_n|}\cdot\frac{\xi_n-z}{1-\bar{\xi}_n z}\right]^{m_n},$$

$$F(z)=\exp\left[\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{e^{j\phi}+z}{e^{j\phi}-z}\log|f(e^{j\phi})|\,d\phi\right],$$

$$S(z)=\exp\left[-\int_{-\pi}^{\pi}\frac{e^{j\phi}+z}{e^{j\phi}-z}\,dm(\phi)\right],$$

where $dm(\phi)$ is a positive singular measure on $[-\pi,\pi]$. The infinite product defining $B(z)$ converges uniformly and absolutely on any compact subset of the complex plane which does not contain any accumulation point of the sequence $\{\xi_n\}$ nor any point of the closure of $\{1/\xi_n\}$. In particular, B is analytic on D and continuous at every point of C which is not a limit point of $\{\xi_n\}$. Also $|B(z)|=1$ a.e. on C and $|B|<1$ on D. The function f cannot be continuously extended from the interior of D to any point of C which is in the support of dm, nor to any point of C which is a limit point of $\{\xi_n\}$. Thus, if f has no zeros on C_1 and is continuous there, then $dm(\phi)=\lambda\delta(\phi)\,d\phi$ for some $\lambda>0$. Finally if $f\in H^\infty(D)$, then for any r, $0\leqq r<1$, and $\theta\in[-\pi,\pi]$ we have

$$f(r\,e^{j\theta})=\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{1-r^2}{1+r^2-2r\cos(\theta-\phi)}f(e^{j\phi})\,d\phi$$

$$=\frac{1}{2\pi}\int_{-\pi}^{\pi}f(e^{j\phi})\,\mathrm{Re}\,\frac{e^{j\phi}+r\,e^{j\theta}}{e^{j\phi}-r\,e^{j\theta}}\,d\phi.$$

3. System description. We consider a scalar feedback system with input u, error ε and output y. These are real-valued functions on $[0, \infty)$ and satisfy

$$y = g * \varepsilon, \qquad \varepsilon = u - ky, \tag{3.1}$$

where $*$ denotes convolution, g is a real distribution on $[0, \infty)$ and k is a real scalar gain. We assume that g has Laplace transform $\hat{g}(s)$ and

$$\hat{g}(s) = \hat{g}_1(s) + \frac{p(s)}{q(s)}, \tag{3.2}$$

where $p(s)$ and $q(s)$ are polynomials in s such that $\deg p < \deg q$, p and q have no common zeros and are real on the real line, and $\hat{g}_1(s)$ is the Laplace transform of an element of A as in § 2.2 with $\sigma = 0$. Note that if q has any zeros in the open left-halfplane, then $p(s)/q(s) = h(s) + p_1(s)/q_1(s)$, where $h(s)$ is the Laplace transform of a real integrable function, $p_1(s)$ and $q_1(s)$ satisfy the above conditions on p and q, and q_1 has all its zeros in $\bar{R}$. So without loss of generality we may assume that q has no zeros in the open left halfplane. Note also that the zeros of p and q appear in complex conjugates. Denoting the zeros of q by $s_1, \cdots, s_n$, each counted according to its multiplicity and putting $\sigma_1 = \max \{\operatorname{Re} s_i : 1 \leqq i \leqq n\}$ we see that $g \in A(\sigma)$ for all $\sigma > \sigma_1$.

The following result is due to Desoer and Wu [6]. Their proof is based on Gronwall's inequality and is rather complicated. We give a short proof.

PROPOSITION 1. *If* $1 + ka_0 \neq 0$, *then for every locally integrable function* u *on* $(0, \infty)$ (3.1) *has a unique solution in* y *and* ε. *If* u *is Laplace transformable then so are* y *and* ε.

Proof. Any solution y must satisfy $(\delta(t) + kg) * y = g * u$. From $1 + ka_0 \neq 0$ we have $\inf_{\bar{R}(\sigma)} |1 + k\hat{g}(s)| > 0$ for some $\sigma > 0$. Hence by the theorem in § 2.2, there exists $dm \in A(\sigma)$ such that $dm * (\delta(t) + kg) = \delta(t)$, and thus $y = dm * g * u$ and $\varepsilon = u - k\, dm * g * u$. These satisfy (3.1), so existence and uniqueness follow. If u is Laplace transformable so is $dm * g * u$ and thus so are y and ε.

From now on we assume that $1 + ka_0 \neq 0$ and $k \neq 0$. Let us denote by T the above input-output map; thus, $y = Tu$. By Proposition 1, T is defined by an element $dh \in A(\sigma)$ for some $\sigma > 0$ and $y = u * dh$. Note that $d\hat{h}(s) = \hat{g}(s)/(1 + k\hat{g}(s))$ for $\operatorname{Re} s \geqq \sigma$, while the same equation gives a continuation of $d\hat{h}$ to $\bar{R}$. We shall denote this continuation by $d\hat{h}$ as well. From the proof of Proposition 1 it follows that if u is Laplace transformable on R then $\hat{y}(s) = d\hat{h}(s)\hat{u}(s)$. Note also that $1 + k\hat{g}(s)$ is analytic on R except at the zeros of $q(s)$ where it has poles, and in the complement in $\bar{R}$ of the union of any neighborhoods of these points it is bounded. It will be of interest to know when $dh \in A$.

PROPOSITION 2. *The following are equivalent.*

(i) $dh \in A$,

(ii) $d\hat{h}(s)$ *is bounded on* $\bar{R}$,

(iii) $\inf_{s \in \bar{R}} |1 + k\hat{g}(s)| > 0$.

Proof.

(i) $\Rightarrow$ (ii). This is obvious, since $|d\hat{h}(s)| \leqq \|dh\|_A$, $s \in \bar{R}$.

(ii) $\Rightarrow$ (iii). This follows from $|1/(1 + k\hat{g}(s))| = |d\hat{h}(s) - 1/k|$ and the fact that $k \neq 0$.

(iii) $\Rightarrow$ (i). Let n be the number of the zeros of q. Then $d\hat{h}(s) = h_1(s)/h_2(s)$, where $h_1(s) = \{q(s)\hat{g}_1(s) + p(s)\}/(s+1)^n$, $h_2(s) = q(s)(1 + k\hat{g}(s))/(s+1)^n$ and h_1 and h_2 are the Laplace transforms of elements of A. In view of the theorem in § 2.2 it suffices to show that $\inf_{\bar{R}} |h_2(s)| > 0$. First we observe that by (iii) there exists some $r > 0$ such that $\inf \{|h_2(s)| : s \in \bar{R}, |s| \geqq r\} > 0$. We show that for any $r > 0$ we also have $\inf \{|h_2(s)| : s \in \bar{R}, |s| \leqq r\} > 0$. If not, by continuity of h_2 on $\bar{R}$ we would have $h_2(s_0) = 0$ for some

s_0 in $\bar{R}$. Since $1+k\hat{g}(s)$ does not vanish on R we must have $q(s_0)=0$. But then $p(s_0)=0$. The contradiction establishes the proposition.

4. L^p-stability. In discussing stability of delayed systems of the above type, Callier and Desoer defined the system to be stable if $dh \in A$. Desoer and Wu have proved that if this requirement is satisfied the operator T defined by $y = Tu$ is bounded on all L^p-spaces, $1 \leqq p \leqq \infty$. A much stronger result holds. From now on, L^p, $1 \leqq p \leqq \infty$ shall mean $L^p(0, \infty)$.

THEOREM 1. *The following are equivalent.*

(i) *T is bounded on L^p for all $1 \leqq p \leqq \infty$;*

(ii) *T is bounded on L^p for some p, $1 \leqq p \leqq \infty$;*

(iii) *T is bounded on L^2;*

(iv) *$dh \in A$.*

Remark. (iv) $\Rightarrow$ (i) is already in [6]. Another proof of it depends on the following fact which is an easy case of Marcinkiewicz interpolation theorem [12, pp. 111–112].

Fact. Let S be a linear map defined and bounded on L^p and $L^{p'}$ for some p and p', $1 \leqq p < p'$. Then S can be continuously extended to L^r for all r, $p < r < p'$ so as to be bounded there.

Proof of Theorem 1.

(i) $\Rightarrow$ (ii). This is obvious.

(ii) $\Rightarrow$ (iii). Suppose T is bounded on L^p for some p, $1 \leqq p \leqq \infty$. Let q be the conjugate of p, i.e. $1/p + 1/q = 1$ [if $p = 1$ then $q = \infty$, and if $p = \infty$ then $q = 1$]. We show that T is bounded on L^q. Since 2 is either equal to p or lies between p and q, (iii) follows from the above fact.

First suppose that $1 < p < \infty$. Let $f \in L^q$, $g \in L^p$ be continuous and of compact support in $[0, m]$, where m is an arbitrary positive number. Define $\phi(t) = f(m-t)$ and $\psi(t) = g(m-t)$. Then ϕ and ψ are in L^q and L^p respectively, have compact support in $[0, m]$ and $\|\Phi\|_{L^q} = \|f\|_{L^q}$, $\|\psi\|_{L^p} = \|g\|_{L^p}$. It follows from the definition of T and Fubini's theorem that

$$\int_0^\infty f(t)(Tg)(t)\,dt = \int_0^\infty \psi(t)(T\phi)(t)\,dt.$$

Therefore, by Holder's inequality, we have

$$\left|\int_0^\infty \psi T\phi\right| = \left|\int_0^\infty fTg\right| \leqq \|f\|_{L^q}\|Tg\|_{L^p} \leqq K_p\|f\|_{L^q}\|g\|_{L^p} = k_p\|\phi\|_{L^q}\|\psi\|_{L^p}$$

where K_p is the norm of T on L^p. Now since every continuous compactly supported function can be written as a ψ above, and the set of all such functions is dense in L^p [10, p. 68], $\phi \in L^q$ and $\|T\phi\|_q \leqq K_p\|\phi\|_q$. Again, since the set of ϕ is dense in L^q it follows that T is bounded on L^q.

If $p = 1$ then $q = \infty$. Defining ϕ and ψ above and applying the same argument we obtain that $T\phi \in L^\infty$ and $\|T\phi\|_\infty \leqq K_1\|\phi\|_\infty$. However, in this case it is no longer true that the set of ϕ's is dense in L^∞. To show that T has a bounded extension to L^∞ we proceed as follows. Fix $\theta \in L^\infty$. For each $t > 0$ let $\theta_t(x) = \theta(x)$ if $x < t$ and 0 otherwise. Since dh is supported by $[0, \infty)$, it is easily seen that $(T\theta_t)(x) = (T\theta_{t'})(x)$ for almost all x such that $x < t$ and $x < t'$. Putting $(T\theta)(x) = (T\theta_t)(x)$, $x < t$, we obtain a well defined extension of T to L^∞. Moreover, for each $X > 1$ we have

$$\operatorname*{ess\,sup}_{0 \leqq x \leqq X-1} |(T\theta)(x)| \leqq \operatorname*{ess\,sup}_{0 \leqq x \leqq X} |T\theta_X(x)| \leqq K_1\|\theta_X\|_{L^\infty} \leqq K_1\|\theta\|_{L^\infty}.$$

Thus, $\|T\theta\|_{L^\infty} \leqq K_1\|\theta\|_{L^\infty}$ and T is bounded on L^∞.

If $p=\infty$ then $q=1$. With ϕ and ψ as above we have

$$\left|\int \psi T\phi\right| \leqq K_\infty \|\psi\|_\infty \|\phi\|_{L^1}.$$

Thus, choosing a sequence g_n suitably we may assume $\psi_n(x)=\operatorname{sign}\,(T\phi)(x)$ if $x<n$, and 0 otherwise. Therefore,

$$\int_0^n |T\phi| \leqq K_\infty \|\phi\|_{L^1} \quad \text{for every } n>0,$$

so that

$$T\phi \in L^1 \text{ and } \|T\phi\|_{L^1} \leqq K_\infty \|\phi\|_{L^1}.$$

Since the set of ϕ's is dense in L^1 the boundedness of T follows.

(iii) $\Rightarrow$ (iv). Here we use the background material on H^2. Let M be the norm of T on L^2. We show that $|d\hat{h}(s)| \leqq M$ on $\bar{R}$ so that by Proposition 2 the result follows. First we prove that $|d\hat{h}(jw)| \leqq M$ a.e. on $(-\infty, +\infty)$. Suppose the contrary. Then there exists a bounded set S on the imaginary axis of positive measure m, on which $|d\hat{h}(jw)| > M+a$ for some $a>0$. Let $f(jw) \in L^2(-\infty, \infty)$ be such that $|f(jw)| \geqq 1$ on S. Let $u \in L^2(0, \infty)$ be such that $f(jw) = \operatorname{Re} \int_0^\infty u(t) \exp(-jwt)\, dt$ (see § 2.3). Apply the operator T/M to u successively. Since T/M has norm 1, we have $\|(1/M^n)(T^n u)\|_{L^2} \leqq \|u\|_{L^2}$. By Parseval's theorem there is an absolute constant K such that

$$\begin{aligned}\left\|\frac{1}{M^n} T^n u\right\|_{L^2} &= K \left\|\frac{1}{M^n}(T^n u)\hat{}\,(jw)\right\|_{L^2} \\ &\geqq K\left(\frac{M+a}{M}\right)^n \left(\int_S |\hat{u}(jw)|^2\, dw\right)^{1/2} \\ &\geqq K\left(\frac{M+a}{M}\right)^n \sqrt{m},\end{aligned}$$

so $\|u\|_{L^2} \geqq K\sqrt{m}((M+a)/M)^n$ for every $n \geqq 1$, which is impossible. Now $d\hat{h}(jw)$ is continuous on $(-\infty, +\infty)$ except at the zeros of $1+k\hat{g}(jw)$, and tends to infinity as jw tends to any of these zeros. It follows that no such zeros exist and $|d\hat{h}(jw)| \leqq M$ on $(-\infty, +\infty)$.

Next we note that for any s_0 with $\operatorname{Re} s_0 > 0$ there exists $f \in H^2$ such that $f(s_0)=1$. To see this, choose any nonzero $f_1 \in H^2$. Then for some s_1 with $\operatorname{Re} s_1 > \operatorname{Re} s_0$ we have $f_1(s_1) \neq 0$. Choose λ such that $\lambda f_1(s_1)=1$. Then $f(s)=\lambda f_1(s+s_1-s_0)$ is in H^2 and $f(s_0)=1$. Thus, for each s_0 in R there exists $u \in L^2(0,\infty)$ such that $\hat{u}(s_0)=1$. Now since T is bounded on L^2, for any $u \in L^2$ and any $s=x+jy$, $x>0$ we have

$$\begin{aligned}|d\hat{h}(s)\hat{u}(s)| &= \frac{1}{\pi}\left|\int_{-\infty}^{\infty} d\hat{h}(jw)\hat{u}(jw)\frac{x}{x^2+(y-w)^2}\, dw\right| \\ &\leqq \frac{M}{\pi}\int_{-\infty}^{\infty} |\hat{u}(jw)| \frac{x}{x^2+(y-w)^2}\, dw.\end{aligned}$$

By iteration we have

$$|(d\hat{h}(s))^n \hat{u}(s)| \leqq \frac{M^n}{\pi}\int_{-\infty}^{\infty} |\hat{u}(jw)| \frac{x}{x^2+(y-w)^2}\, dw.$$

Note that for any $\hat{u}(jw)$ in L^2, any $x>0$ and any y by Hölder's inequality the above integral is finite. For any s in R choose u as above such that $\hat{u}(s)=1$. Then

$$\left|\frac{1}{M^n}(d\hat{h}(s))^n\right| \leqq \frac{1}{\pi}\int_{-\pi}^{\pi} |\hat{u}(jw)|\frac{x}{x^2+(y-w)^2}\,dw$$

for every $n \geqq 1$. Thus $|d\hat{h}(s)| \leqq M$.

(iv) $\Rightarrow$ (i). It is well known [7] that if $dh \in A$ then T is bounded on L^1 and L^∞, so that by the above fact T is bounded on all L^p, $1 \leqq p \leqq \infty$.

The reader will have observed that in Theorem 1 we implicitly assumed that $1+ka_0 \neq 0$ so as to have $dh \in A(\sigma)$ for some $\sigma > 0$. If this is not the case, (3.1) need not have Laplace transformable solutions for some u. Even if for all u in L^p the solution is in L^p, the following theorem holds.

THEOREM 2. *Let $p \in [1, \infty]$. If $1+ka_0 = 0$, and for every u in L^p* (3.1) *has a solution in L^p, then the map $u \to y$ is not bounded.*

Proof. The relation

$$\hat{y}(s) = \frac{\hat{g}(s)}{1+k\hat{g}(s)}\hat{u}(s)$$

still holds. Since $1+ka_0=0$ then $a_0 \neq 0$, and so by the theorem in § 2.2 there exists some $\sigma > 0$ and some $dm \in A(\sigma)$, such that $dm * g = \delta(t)$. Then $d\hat{m}(s)(1+k\hat{g}(s))\hat{y}(s) = \hat{u}(s)$. Observe that for all $p \in [1, \infty]$, q defined by $p^{-1}+q^{-1}=1$ satisfies $q \in [1, \infty]$, and for all $x > 0$, $\|e^{-x\cdot}\|_{L^q} = (1/qx)^{1/q}$, where it is understood that for $q=\infty$, $\lim_{q\to\infty}(1/qx)^{1/q} = 1 = \|e^{-x\cdot}\|_{L^\infty}$. Moreover, given any $x > 0$, by duality ((a) by [10, Thm. 6.16, p. 128] for $p \in [1, \infty)$ and (b) by [10, Thm. 6.19, p. 131] for $p=\infty$) there exists $u \in L^p$ with $\|u\|_{L^p}=1$, such that $|\hat{u}(x+j0)| > \frac{1}{2}(1/qx)^{1/q}$, $x > 0$. Observe that for $p=\infty$ we pick $u \in C_0 \subset L^\infty$, (see [10, Th. 6.19, p. 131]). Since $1+ka_0=0$, then $d\hat{m}(s)(1+k\hat{g}(s)) \to 0$ as $\operatorname{Re} s \to \infty$. Let x_n be a sequence tending to ∞ and let $\varepsilon_n = d\hat{m}(x_n)(1+k\hat{g}(x_n))$ so that $\varepsilon_n \to 0$. Choosing $u_n \in L^p$ with $\|u_n\|_{L^p}=1$ such that $|\hat{u}_n(x_n)| > \frac{1}{2}(1/qx_n)^{1/q}$, and denoting by y_n the image of u_n under the map (3.1), we have $|\hat{y}(x_n)| > (1/2\varepsilon_n)(1/qx_n)^{1/q}$. By Hölder's inequality the assumption $y_n \in L^p$ gives $|\hat{y}(x_n)| \leqq (1/qx_n)^{1/q}\|y_n\|_{L^p}$, so that $\|y_n\|_{L^p} \geqq 1/2\varepsilon_n$. Since $\|u_n\|_{L^p}=1$ and $\varepsilon_n \to 0$ the result follows.

Stability. In view of Propositions 1 and 2 and Theorems 1 and 2 we say that the system is stable if $dh \in A$. A necessary and sufficient condition for stability is that $\inf_{\bar{R}} |1+k\hat{g}(s)| > 0$.

5. Test of stability. We now give a test of stability based on the values of $1+k\hat{g}(s)$ on the imaginary axis. It is thus necessary to relate the values of $1+k\hat{g}(s)$ in R to its boundary values in a suitable way.

It will appear shortly that it is more convenient to map the right halfplane to the unit disk and carry out the analysis there. First we remove the poles of $1+k\hat{g}(s)$ as follows. Let $s_1, \cdots, s_m$ be the zeros of q on the imaginary axis and $s_{m+1}, \cdots, s_n$ the zeros of q in R. Let

$$A_1(s) = \prod_{i=1}^{m}\left(\frac{s-1}{s+1} - \frac{s_i-1}{s_i+1}\right) \tag{5.1}$$

and

$$A_2(s) = \left(\frac{s-1}{s+1}\right)^{m_0} \prod_{\substack{i=m+1 \\ s_i \neq 1}} \frac{s-s_i}{s+\bar{s}_i} \tag{5.2}$$

where m_0 is the order of the pole of $\hat{g}(s)$ at 1. Let $G(s) = A_1(s)A_2(s)(1+k\hat{g}(s))$ and

$\alpha_i=(s_i-1)/(s_i+1)$, $1\leqq i\leqq n$. Note that each factor in A_2 is a conformal map of R to D which takes s_i to the origin. Moreover under the map $z=(s-1)/(s+1)$, $\alpha_i=(s_i-1)/(s_i+1)$, each factor in A_2 is transformed to $(1-\bar\alpha_i)/(1-\alpha_i)\cdot(z-\alpha_i)/(1-\bar\alpha_i z)$, while each factor in A_1 is transformed to $z-\alpha_i$. Let $\tilde A_r(z)=A_r((1+z)/(1-z))$, $r=1,2$, and

$$(5.3)\qquad f(z)=\tilde A_1(z)\tilde A_2(z)\left(1+k\hat g\left(\frac{1+z}{1-z}\right)\right),\qquad z\in\bar D_1.$$

The reader will easily verify that $f(z)$ is analytic on D, continuous and bounded on $\bar D_1$, and $\inf_{\bar R}|1+k\hat g(s)|>0$ if and only if $\inf_{\bar D_1}|f(z)|>0$; so that the latter is a necessary and sufficient condition for stability. Since $f(z)\in H^\infty(D)$ it has a factorization of the same form as in § 2.4, $f(z)=\rho B(z)F(z)S(z)$. Since f is real on $(-1,+1)$, $\rho=1$ or -1, and since f may be extended continuously from D to $\bar D_1$ according to [9, pp. 68–69], the positive singular measure $dm(\phi)$, defining $S(z)$, has its closed support contained in the singleton $\{0\}$; hence $dm(\phi)=\lambda\delta(\phi)\,d\phi$ with $\lambda\geqq 0$ and $\delta(\phi)\,d\phi$ the Dirac measure at $\phi=0$. Hence $S(z)=\exp(-\lambda(1+z)/(1-z))$ with $\lambda=0$ if f can be continuously extended to $\bar D$, (which is the case if $1+k\hat g(s)$ is singlevalued at ∞). Also $G(s)=A_1(s)A_2(s)(1+k\hat g(s))$ tends to a non-zero multiple of $1+ka_0$ as $\operatorname{Re} s\to\infty$, so that as $z\to 1^-$ on the real line $f(z)$ tends to a non-zero multiple of $1+ka_0$. The reader is asked to remember this when he reads the remark following Theorem 3.

The following result is a modification of a well-known theorem in the theory of H^p-spaces [9, p.62].

THEOREM 3. *With the above notation the following are equivalent.*

(i) $\inf_{\bar D_1}|f(z)|>0$.

(ii) $\inf_{\theta\neq 0}|f(e^{i\theta})|>0$ *and* $f(z)=\rho F(z)$.

(iii) $\inf_{\theta\neq 0}|f(e^{i\theta})|>0$ *and* $\log|f(0)|=\dfrac{1}{2\pi}\displaystyle\int_{-\pi}^{\pi}\log|f(e^{i\theta})|\,d\phi$.

Proof.

(i) ⇒ (ii). Clearly (i) implies $\inf_{\theta\neq 0}|f(e^{i\theta})|>0$ and $B(z)=1$. Since $|f(e^{i\theta})|$ is bounded above on C_1 we have $|\log|f(e^{i\theta})||\leqq M<\infty$, for some M and all $\theta\neq 0$. Now let $0\leqq r<1$. Then by the definition of F in § 2.4

$$\begin{aligned}|F(r)|&\leqq\exp\left\{\frac{M}{2\pi}\int_{-\pi}^{\pi}\left|\operatorname{Re}\frac{e^{i\phi}+r}{e^{i\phi}-r}\right|d\phi\right\}\\&=\exp\left\{\frac{M}{2\pi}\int_{-\pi}^{\pi}\frac{1-r^2}{1+r^2-2r\cos\phi}\,d\phi\right\}\\&=\exp M.\end{aligned}$$

Thus, $|F(r)|\leqq\exp M$ uniformly in $(0,1)$. On the other hand $\exp(-\lambda(1+r)/(1-r))$ tends to zero as r tends to 1 from below unless $\lambda=0$. Hence, $\lambda=0$ and $f(z)=\rho F(z)$.

(ii) ⇒ (i). Let M be as above. Given $z=re^{i\theta}\in D$ we have

$$\left|\frac{1}{2\pi}\int_{-\pi}^{\pi}\log|f(e^{i\phi})|\operatorname{Re}\frac{e^{i\phi}+z}{e^{i\phi}-z}\,d\phi\right|\leqq\frac{M}{2\pi}\int_{-\pi}^{\pi}\frac{1-r^2}{1+r^2-2r\cos(\theta-\phi)}\,d\phi=M,$$

so that $|F(z)|\geqq\exp(-M)$ uniformly in D. Hence, $|f(z)|=|\rho F(z)|=|F(z)|\geqq\exp(-M)>0$ for all z in D. Since by assumption we also have $\inf_{\theta\neq 0}|f(e^{i\theta})|>0$ (i) follows.

(ii) ⇒ (iii). This follows by direct substitution.

(iii)$\Rightarrow$(ii). Consider $f(z)/F(z)$. As before, $|F(z)| \geqq \exp(-M)$ in D, so f/F is analytic on D. Note that $f/F = \rho BS$ on D, and $|\rho BS| = 1$ a.e. on C_1. The hypothesis implies that the modulus of the analytic function f/F attains its maximum at 0 which is inside D. Hence, f/F is constant and so $BS \equiv 1$. From the $H^\infty(D)$ factorization of f it follows that $f = \rho F$.

Remark. From the proof of (i) $\Rightarrow$ (iii) above and as mentioned earlier we see that when $\inf_{\theta \neq 0} |f(e^{i\theta})| > 0$ then $S(z) = 1$, or equivalently $\lambda = 0$ if and only if $f(r) \nrightarrow 0$ as $r \to 1^-$ and this is equivalent to $1 + ka_0 \neq 0$.

With m_0 as the order of the pole of $\hat{g}(s)$ at $s = 1$ let

$$L = \lim_{s \to 1} \left(\frac{s-1}{s+1}\right)^{m_0} (1 + k\hat{g}(s)).$$

Then we have

THEOREM 4. *With the above notation,* $\inf_{\bar{R}} |1 + k\hat{g}(s)| > 0$ *if and only if* $\inf\{|1 + k\hat{g}(jw)| : -\infty < w < \infty\} > 0$ *and*

$$\log|L| = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\log|1 + k\hat{g}(jt)|}{1 + t^2} dt - \sum \log \left|\frac{s_i - 1}{s_i + 1}\right|,$$

where the sum is over all the poles of $\hat{g}(s)$ *in* R *different from* 1.

Proof. By the observation after the definition of $f(z)$ in (5.3) $\inf_{\bar{R}} |1 + k\hat{g}(s)| > 0$ if and only if $\inf_{\bar{D}_1} |f(z)| > 0$, and the latter is equivalent to condition (iii) of Theorem 3. We now show that these are equivalent to the conditions of our theorem. From (5.3) we have $A_1(jw)(1 + k\hat{g}(jw)) = f(e^{i\theta})(A_2(jw))^{-1}$, where w and θ are related by the bijection $e^{i\theta} = (jw - 1)/(jw + 1)$, $-\infty < w < \infty, 0 < \theta < 2\pi$. From its definition A_2 is the product of a finite number of conformal maps of R to the unit disk, and so $|A_2(jw)| = 1$ for all w. Also $A_1(jw)$ is bounded on $(-\infty, \infty)$, has a finite number of zeros there which are the poles of $(1 + k\hat{g}(jw))$ with the same multiplicities, and tends to a nonzero constant as $|w| \to \infty$; $1 + k\hat{g}(jw)$ is everywhere continuous except at its poles. Hence $\inf_{\theta \neq 0} |f(e^{i\theta})| > 0$ implies $\inf\{|1 + k\hat{g}(jw)| : -\infty < w < \infty\} > 0$. Conversely suppose that the latter holds. Choose w_0 such that the zeros of $A_1(jw)$ are in $(-w_0, w_0)$ and outside this interval $|A_1(jw)| > c > 0$ for some c. Then with θ_0 the image of w_0 under the above bijection $\inf\{|f(e^{i\theta})| : \theta \notin [\pi - \theta_0, \pi + \theta_0]\} > 0$. Since the zeros of $A_1(jw)$ are the poles of $1 + k\hat{g}(jw)$ with the same multiplicities, $A_1(jw)(1 + k\hat{g}(jw))$ is continuous and by assumption has no zeros in $[-w_0, w_0]$; so it is bounded away from zero there. Thus, $\inf\{|f(e^{i\theta})| : \theta \in [\pi - \theta_0, \pi + \theta_0]\} > 0$. In short $\inf\{|1 + k\hat{g}(jw)| : -\infty < w < \infty\} > 0$ is equivalent to $\inf_{\theta \neq 0} |f(e^{i\theta})| > 0$. Now with L defined as above, it follows immediately that $\log|f(0)| - \log|L|$ is the sum in the above formula. Let us consider $\int_{-\pi}^{\pi} \log|f(e^{i\phi})| \, d\phi$. From (5.1)–(5.3) we have $\log|f(z)| = \log|\tilde{A}_1(z)| + \log|\tilde{A}_2(z)| + \log|1 + k\hat{g}((1+z)/(1-z))|$. Each factor in $\tilde{A}_1(z)$ is of the form $z - \alpha$ for some α with $|\alpha| = 1$. Since $\log|z - \alpha|$ is harmonic in D, it follows from the mean value property of harmonic functions [10, p. 230] that $\int_{-\pi}^{\pi} \log|e^{i\phi} - \alpha| \, d\phi = 0$, while each factor in $\tilde{A}_2(z)$ is a conformal map of D to itself and so $|A_2(e^{i\phi})| = 1$. Thus, the contribution of $\log|\tilde{A}_1(e^{i\phi})|$ and of $\log|A_2(e^{i\phi})|$ is nil. Now the change of variable $e^{i\phi} \to (jt + 1)/(jt - 1)$ changes integration with respect to $d\phi$ from $-\pi$ to π, to integration with respect to $(2dt)/(1 + t^2)$ from $-\infty$ to ∞, so that the integral in the above formula is in fact equal to

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \log|f(e^{i\phi})| \, d\phi.$$

Remark. Theorem 4 gives a test of stability for systems defined by (3.1). This test is based on the amplitude of $1+k\hat{g}(jw)$ and the exact position of the poles of $\hat{g}(s)$ in R. In contrast, the stability results of Callier and Desoer, [3, Thm. 2] and [4], are based on the argument of $1+k\hat{g}(jw)$ and the number of the poles of $\hat{g}(s)$ in R (except when $\hat{g}(s)$ has poles on the imaginary axis, so that their positions are required to define the argument of $1+k\hat{g}(jw)$ by the convention in [3]). In view of the relationship between the real and imaginary parts of meromorphic functions it is not surprising that different tests of stability involving the amplitude and the argument of $1+k\hat{g}(jw)$ exist side by side. However, it has to be pointed out that a test, such as our Theorem 4, which is based on the exact position of the poles of $\hat{g}(s)$, is insensitive to errors in the location of these poles, while a test based on the number of poles does not require a knowledge of the position of these poles. Hence, in practice the stability result of [4] is much more robust than ours.

6. Maximal intervals of stability. The above test and that of [4] depend on the specific value of the feedback gain k. It would indeed be troublesome if a separate test had to be carried out for each value of k. Fortunately there is a simple way to get round this difficulty. Here, we are guided by the way that any test would be applied. First the graph Γ of $\hat{g}(jw)$, $-\infty<w<\infty$, is drawn and k is chosen so that the point $-1/k$ is a positive distance away from Γ. In general the shape of Γ could be very complicated. In practice the segments of the real line which contain candidates for $-1/k$ will be easily observed and will be disjoint intervals. It is reasonable to expect that if any one such interval is considered then the outcome of the test would be the same for all points of this interval,c.f. the classical Nyquist diagram.

More precisely, let T_0 be the intersection of the closure of $\{\hat{g}(jw): -\infty<w<\infty\}$ and the real line. T_0 contains the set of real values that $\hat{g}(jw)$ takes, and due to the presence of almost periodic elements in $\hat{g}$ it may contain other points as well. Let $T=T_0\cup\{0\}$. Then T is a closed subset of the real line; its complement, being an open set, is a disjoint union of a countable number of open intervals [1, p. 46], none of which contains the point zero. A necessary condition for the gain k to give stability is that $-1/k$ is in one of these intervals. Naturally, this is not, in general, sufficient. We have

THEOREM 5. *Let $L=(l_1, l_2)$ be an arbitrary interval in the complement of T. Then either* $\inf_{\bar{R}}|1+k\hat{g}(s)|>0$ *for every k such that $-1/k\in L$, or* $\inf_{\bar{R}}|1+k\hat{g}(s)|=0$ for every *k such that $-1/k\in L$.*

The theorem shows that either all the gains k such that $-1/k\in L$ give rise to closed-loop stability, or else none of them does. In the proof of Theorem 5 we need the following simple lemma whose proof is included for the sake of completeness.

LEMMA 1. *For $k\neq 0$ let $d(k)=\inf_{\bar{R}}|1+k\hat{g}(s)|$; then $d(k)$ is continuous for $k\neq 0$.*

Proof. Let $k_0\neq 0$ be given. Choose $\delta>0$ such that $0\notin[k_0-\delta, k_0+\delta]=I$. Let $m=\min\{|k|: k\in I\}$ and $M=\max\{|k|: k\in I\}$. Let s_0 be any point of $\bar{R}$ which is not a pole of $\hat{g}(s)$. Then $|1+k\hat{g}(s_0)|\leqq 1+M|\hat{g}(s_0)|$, whenever $k\in I$. Also $|1+k\hat{g}(s)|\geqq|k|\,|\hat{g}(s)|-1$. Choose an open set O containing the poles of $\hat{g}(s)$ in $\bar{R}$, such that $m|\hat{g}(s)|-1>1+M|\hat{g}(s_0)|$ whenever $s\in\bar{R}\cap O$. It follows that $|1+k\hat{g}(s)|>1+M|\hat{g}(s_0)|$ whenever $k\in I$ and $s\in\bar{R}\cap O$ so that for $k\in I$ we have $d(k)=\inf\{|1+k\hat{g}(s)|: s\in\bar{R}\backslash O\}$. Now, $|\hat{g}(s)|$ is bounded on $\bar{R}\backslash O$ by, say, N. For $\varepsilon>0$ arbitrary choose $\lambda>0$ such that $\lambda<\delta$ and $\lambda N<\varepsilon/2$. Choose $\sigma\in\bar{R}\backslash O$ such that $|1+k_0\hat{g}(\sigma)|<d(k_0)+\varepsilon/2$. For $|k-k_0|<\lambda$, we have $k\in I$ and $|1+k\hat{g}(\sigma)|\leqq|1+k_0\hat{g}(\sigma)|+|k-k_0|\,|\hat{g}(\sigma)|<d(k_0)+\varepsilon$. Hence, $d(k)<d(k_0)+\varepsilon$. In exactly the same way we have $d(k_0)<d(k)+\varepsilon$, and so $|d(k)-d(k_0)|<\varepsilon$ whenever $|k-k_0|<\lambda$. So $d(k)$ is continuous.

Proof of Theorem 5. It suffices to show that if $-1/k_0 \in L$ is such that $\inf_{\bar{R}} |1+k_0\hat{g}(s)|>0$, then for any other k such that $-1/k \in L$ we have $\inf_{\bar{R}} |1+k\hat{g}(s)|>0$. Let $M(z)=(1+k_0\hat{g}((1+z)/(1-z)))^{-1}$. Then $M(z)$ is a bounded analytic function on D, and so by § 2.4 for every r, $0\leqq r<1$ and every θ, $-\pi\leqq\theta\leqq\pi$.

$$M(r\,e^{j\theta})=\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{1-r^2}{1+r^2-2r\cos(\phi-\theta)}M(e^{j\phi})\,d\phi.$$

Thus,

$$|M(r\,e^{j\theta})|\leqq \operatorname*{ess\,sup}_{\phi\in[-\pi,\,\pi]}|M(e^{j\phi})|,\qquad 0\leqq r<1,\ \theta\in[-\pi,\pi].$$

Hence, from the definition of $M(z)$ we have,

$$|1+k_0\hat{g}(s)|\geqq \inf_{-\infty<w<\infty}|1+k_0\hat{g}(jw)|>0.$$

Now the function $d(k)=\inf_{s\in\bar{R}}|1+k\hat{g}(s)|$ is continuous on $(-\infty,0)\cup(0,+\infty)$ and is strictly positive at k_0. Therefore, there exists $k_1>k_0$ such that $(-1/k_1,-1/k_0)\subset L$ and $\inf_{\bar{R}}|1+k_1\hat{g}(s)|>0$. Repeat the argument with k_0 replaced by k_1 to obtain a sequence $k_n>k_{n-1}$ such that $(-1/k_n,-1/k_0)\subset L$ and $\inf_{\bar{R}}|1+k_n\hat{g}(s)|>0$ for each $n>0$. We prove that k_n may be chosen so that $-1/k_n$ tends to l_1. The argument shows that we need only consider the case $l_1>-\infty$. Suppose the contrary. Thus, if K is the supremum of the k_n's that can be obtained by the above procedure then $-1/K>l_1$. Note that K is finite and $K\neq 0$. The above argument shows that for any k, $k_0<k<K$ we have $\inf_{\bar{R}}|1+k\hat{g}(s)|\geqq\inf_{-\infty<w<\infty}|1+k\hat{g}(jw)|$. Since $K\neq 0$ by the continuity of $d(k)$ we have $\inf_{-\infty<w<\infty}|1+k\hat{g}(jw)|=0$ so that $-1/K\in T$, which is contrary to $-1/K\in L$. A similar argument applies to l_2, and the result follows.

7. A reduction theorem. So far we have considered systems of the form (3.1). It is not obvious that our results apply to systems whose transfer functions contain such terms as the ratio of the Laplace transform of an element of A and a polynomial in s. Such terms occur frequently, e.g. in differential models of the form $\dot{x}(t)=ax(t)+bu(t)+cu(t-1)$. Let us take a very simple example. Let $\hat{g}(s)=e^{-s}/(s-1)$. Then $\hat{g}(s)=(e^{-s}-e^{-1})/(s-1)+e^{-1}/(s-1)$. Note that $(e^{-s}-e^{-1})/(s-1)$ is the Laplace transform of the function which is zero everywhere except on $[0,1]$ and is defined by $f(t)=-e^{t-1}$ on $[0,1]$. So $f\in L^1$ and g is of the form (3.1). What in effect we have done is to subtract the principal part of the Laurent expansion of $\hat{g}(s)$ from $\hat{g}(s)$ and prove that what is left is the Lapace transform of an L^1 function. Precisely the same idea works for more complicated forms of $\hat{g}(s)$, but the proof depends on Paley–Wiener theorem on H^2 functions (see § 2.3). Thus, consider a transfer function of the form $\hat{g}(s)/q(s)$ where $\hat{g}(s)$ is the Laplace transform of any bounded measure dm on $[0,\infty)$ and q is a polynomial in s. Noting that if $\operatorname{Re} a<0$ then $\hat{g}(s)/(s-a)$ is the Laplace transform $dm * e^{at}\in L^1$, we may assume that all the zeros of q are in $\bar{R}$.

THEOREM 6. *Let dm be a bounded measure on $[0,\infty)$ and $q(s)=\prod_{i=1}^{N}(s-s_i)^{r_i}$ be a polynomial in s with zeros s_i in $\bar{R}$. Let $r=\max_{1\leqq i\leqq N}\{r_i\}$, and suppose that $\int_0^\infty t^i d|m|(t)\leqq M<\infty$ for $0\leqq i\leqq r+1$. Then*

$$\frac{d\hat{m}(s)}{q(s)}=\hat{f}(s)+\frac{p(s)}{q(s)},$$

where $f\in L^1(0,\infty)$ and $p(s)$ is a polynomial in s with $0\leqq\deg p(s)<\deg q(s)$. If in addition dm is real and $q(s)$ is real on the line, then f is real and $p(s)$ is real on the line.

Proof. By expansion into partial fractions we can write $d\hat{m}(s)/q(s)$ as a linear combination of terms of the form $d\hat{m}(s)/(s-s_i)^n$, for some $n \leqq r$. So for the first part of the proof it suffices to show that each such term can be written as in the statement of the theorem. Let $h(s) = d\hat{m}(s)$. For each k, $0 \leqq k \leqq n+1$, the hypothesis on dm implies that the Laplace transform $(d/ds)^k h(s)$ of the bounded measure $(-1)^k t^k dm(t)$ is bounded and continuous on $\bar{R}$. Now let

$$p(s) = \sum_{k=0}^{n-1} \frac{1}{k!}(s-s_i)^k \left(\frac{d}{ds}\right)^k h(s_i).$$

Then $0 \leqq \deg p < \deg q$ and $h(s)/q(s) = (h(s)-p(s))/q(s) + p(s)/q(s)$. We show that the analytic function $l(s) = (h(s)-p(s))/q(s)$ is the Laplace transform of some $f \in L^1(0, \infty)$. This is done as follows. We first show that $l(s) \in H^2$ so that by § 2.3 it is the Laplace transform of some $f \in L^2(0, \infty)$. We then show that $l'(s) \in H^2$, so that $l'(s)$ is the Laplace transform of some $f_1 \in L^2(0, \infty)$. But then $f_1(t) = -tf(t)$, and thus $(1+t)f(t) \in L^2(0, \infty)$. Therefore, by Hölder's inequality, $f \in L^1(0, \infty)$ for

$$\int_0^\infty |f(t)|\, dt \leqq \left(\int_0^\infty \left(\frac{1}{1+t}\right)^2 dt\right)^{1/2} \left(\int_0^\infty (1+t)^2 |f(t)|^2\, dt\right)^{1/2} < \infty.$$

To see that $l(s) \in H^2$ let Q be a square centered at s_i whose sides have unit length. Let $\bar{Q} = Q \cap \bar{R}$. For each $x > 0$, let $I_1(x) = Q \cap \{x+jy: -\infty < y < \infty\}$, and $I_2(x) = \{x+jy: -\infty < y < \infty\} \backslash I_1(x)$. Then

$$\int_{-\infty}^\infty |l(x+jy)|^2\, dy = \int_{I_1(x)} |l(x+jy)|^2\, dy + \int_{I_2(x)} |l(x+jy)|^2\, dy.$$

Note that for any $x > 0$,

$$\sup_{I_1(x)} |l(x+jy)| \leq \sup_{\bar{R}} \left|\left(\frac{d}{ds}\right)^n h(s)\right| \leqq M < \infty,$$

so that the first integral is bounded by M^2 uniformly in x. In the second integral we note that the distance between $x+jy$ and s_i is at least $1/2$. Moreover,

$$\left|\frac{h(s)-p(s)}{(s-s_i)^n}\right| \leqq \sum_{k=0}^{n-1} \frac{1}{K!} \frac{2M}{|s-s_i|^{n-k}}.$$

Thus,

$$\int_{I_2(x)} \frac{ds}{|s-s_i|^{2(n-k)}} \leqq 2 \int_{1/2}^\infty \frac{dy}{y^{2(n-k)}} \leqq 4^{2(n-k)},$$

and $\int_{I_2(x)} |l(x+jy)|^2\, dy$ is uniformly bounded in x.

So $l(s) \in H^2$. The proof that $l'(s) \in H^2$ is similar. We have

$$\begin{aligned} l'(s) &= \frac{\{(h'(s)-p'(s))(s-s_i) - (h(s)-p(s))n\}}{(s-s_i)^{n+1}} \\ &= l_1(s) + l_2(s), \end{aligned}$$

where

$$l_1(s) = \frac{-n\left\{h(s)-p(s)-\dfrac{1}{n!}(s-s_i)\left(\dfrac{d}{ds}\right)^n h(s_i)\right\}}{(s-s_i)^{n+1}}$$

and

$$l_2(s)=\frac{\left\{h'(s)-p'(s)-\frac{1}{(n-1)!}\left(\frac{d}{ds}\right)^n h(s)\right\}\cdot}{(s-s_i)^n}$$

Now the previous argument applies word for word to each of $l_1(s)$ and $l_2(s)$, so that $l'(s)\in H^2$. Finally, the last assertion of the theorem is easily verified by examining the above proof.

Remarks. (i) As the reader will have noticed, the only use made of the hypothesis about dm is to obtain existence and boundedness of $(d/ds)^k d\hat{m}(s)$, $1\leqq k\leqq r+1$ in $\bar{R}$. As far as the validity of the theorem is concerned, the assumption on dm can be replaced by this requirement; e.g., when dm is compactly supported, in particular, when it is a finite sum of impulses so that $d\hat{m}(s)=\sum_{i=1}^n a_i e^{-st_i}$ for some integer n and real numbers $a_1,\cdots,a_n,t_1,\cdots,t_n$.

(ii) The above theorem was discovered in ignorance of the work of Callier along the same lines [2, Lemma 1]. A slightly more general form of Callier's result is given in [5, Thm. 2.2] where it is shown that if dm has no nonatomic singular part and the zeros of $q(s)$ are all in the open right halfplane, then $d\hat{m}(s)/q(s)$ may be written as in our Theorem 6. Here there is no other requirement such as our $\int_0^\infty t^i\, d|m|(t)<\infty$, unless q has zeros on the imaginary axis, in which case such requirements will become necessary. In fact with trivial modifications the proof given in [2], [5] works even when dm has nonzero nonatomic singular parts. On the other hand, our proof holds in a much more general setting. Indeed in place of $d\hat{m}(s)/q(s)$ one may consider transfer functions of the form

$$h(s)=\frac{b_0(s)s^n+b_1(s)s^{n-1}+\cdots+b_n(s)}{s^m+d_1(s)s^{m-1}+\cdots+d_m(s)},$$

where $0\leqq n<m$, and each b_i and d_i is analytic and together with its derivative is bounded in some open set S containing $\bar{R}$. For under these assumptions the denominator of $h(s)$, being analytic and of order of $|s|^m$ for $|s|$ large and s in S, has at most a finite number of zeros in S. Thus, $h(s)$ has a finite number of poles in S, and with trivial modifications the proof of Theorem 6 applies to $h(s)$. It should be noted that this proof requires splitting off the principal part of the Laurent expansion of $h(s)$ at its poles in $\bar{R}$; for this to be possible $h(s)$ must be meromorphic in an open set containing $\bar{R}$. Thus, in general, boundedness of b_i and d_i and their derivatives on $\bar{R}$ is not sufficient for this splitting off procedure to work. In this connection see the remarks following [5, Thm. 2.2].

Appendix. Recall that in (3.1) the system's transfer function was of the form $\hat{g}_1(s)+p(s)/q(s)$, with $g_1\in A$. It may be of interest to know whether g_1 may be taken to be any bounded measure on $[0,\infty)$. The difficulty here is that if g_1 has a nonatomic singular part then the inversion theorem in § 2.2 will no longer hold, and so the proof of Proposition 1 breaks down. In the case of L^2-stability this difficulty can be overcome as follows. First observe that y satisfies (3.1) if and only if

$$(\delta(t)+kg)*y=g*u,$$

so for the uniqueness of the solution it is necessary and sufficient that $(\delta(t)+kg)*y=0$ has no nontrivial solution in y. Note that by convolving this equation with a suitable function in $L^1(0,\infty)$, we may replace $\delta(t)+kg$ by an L^1 function; for if $\hat{g}(s)=p(s)/q(s)+\hat{g}_1(s)$ and q has poles $s_1,\cdots,s_n$ in the closed right halfplane then $\prod_1^n(s-$

$s_i)/(s+1)^{n+1}$ is the Laplace transform of some $f\in L^1$ and $h=f*g\in L^1$. Now by Titchmarsh's convolution theorem [11] it follows that the only locally integrable y for which $h*y=0$ is 0, and uniqueness follows. For existence and L^2-boundedness we consider under what conditions the map given by $\hat{y}(s)=\hat{g}(s)\hat{u}(s)/(1+k\hat{g}(s))$ is bounded on L^2. First if the map is bounded then a word for word repetition of the implication (iii) $\Rightarrow$ (iv) of Theorem 1 gives the existence of a bound $M>0$ such that $|\hat{g}(s)/(1+k\hat{g}(s))|<M$ for Re $s\geqq 0$. Hence

$$\left|\frac{1}{1+k\hat{g}(s)}\right|=\left|1-\frac{k\hat{g}(s)}{1+k\hat{g}(s)}\right|<1+|k|M,$$

and so $\inf_{\bar{R}}|1+k\hat{g}(s)|>0$. Conversely if $N=\inf\{|1+k\hat{g}(s)|: s\in\bar{R}\}>0$ then from $\hat{y}(s)=k^{-1}\hat{u}(s)\{1-1/(1+k\hat{g}(s))\}$, we have $|\hat{y}(s)|\leqq|k|^{-1}(1+N)|\hat{u}(s)|$, $s\in\bar{R}$, so that $\|\hat{y}\|_{H^2}\leqq|k|^{-1}(1+N)\|u\|_{H^2}$. Therefore, by § 2.3, $\hat{y}$ is the Laplace transform of a function $y\in L^2$ with support in $(0,\infty)$ such that $\|y\|_{L^2}\leqq|k|^{-1}(1+N)\|u\|_{L^2}$. Hence the map $u\in L^2\to y\in L^2$ is well defined and bounded.

REFERENCES

[1] J. C. Burkill and H. Burkill, *A Second Course in Mathematical Analysis*, Cambridge University Press, London 1970.

[2] F. M. Callier, *On the stability of convolution feedback systems with dynamical feedback*, Automatica, 11 (1975), pp. 85–91.

[3] F. M. Callier and C. A. Desoer, *A graphical test for checking the stability of linear time-invariant feedback systems*, I.E.E.E. Trans. Auto. Control, AC-17 (1972), pp. 773–780.

[4] ———, *On simplifying a graphical stability criterion for linear distributed feedback systems*, I.E.E.E. Trans. Auto. Control, AC-21 (1976), pp. 128–129.

[5] ———, *An algebra of Transfer functions for distributed linear time-invariant systems*, I.E.E.E. Trans. Circuits and Systems, CAS-25 (1978), pp. 651–662, and its Corrections in CAS-26 (1979), p. 360.

[6] C. A. Desoer and M. Y. Wu, *Stability of linear time-invariant systems*, I.E.E.E. Trans. Circuit Theory, CT-15 (1968), pp. 245–250.

[7] J. Dieudonne, *Treatise on Analysis*, Vol. II, Academic Press, New York, 1970.

[8] E. Hille and R. S. Phillips, *Functional Analysis and Semi-Groups*, Revised Edition, A.M.S., Providence, Rhode Island, 1957.

[9] K. Hoffman, *Banach Spaces of Analytic Functions*, Prentice-Hall, Englewood Cliffs NJ, 1962.

[10] W. Rudin, *Real and Complex Analysis*, International Student Edition, McGraw-Hill Inc. New York, 1970.

[11] K. Yosida, *Functional Analysis*, Academic Press, New York, 1965.

[12] A. Zygmund, *Trigonometric Series*, Vol. 2 second edition, Cambridge University Press, London, 1968.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 5, September 1980

SEMIMARTINGALE MODELS OF STOCHASTIC OPTIMAL CONTROL, WITH APPLICATIONS TO DOUBLE MARTINGALES*

RENÉ BOEL† AND MICHAEL KOHLMANN‡

Abstract. The paper gives a fairly general approach to the stochastic optimal control problem, using some recent results on semimartingales. After a short review on these results, an abstract model of stochastic optimal control is described, where the influence of the control is modeled by exponentials of martingales. This model gives a synthesis of the martingale approaches of Davis and Varaiya on the one hand, and Striebel on the other hand. Necessary and sufficient conditions for optimality from these papers are adapted to the abstract model. Topological existence results, which apply to the complete observation case as well as to the partial observation case, are described when the likelihood ratios describing the dynamics of the system are subsets respectively of an L_2- and an L_1-space.

These abstract results are specialized to the problem where the martingales are represented as stochastic integrals. The results can then be written in a more explicit form; in particular it is possible to formulate the optimality conditions as a maximum principle.

All these results are then applied to the double martingale control problem. Combining results of Elliott and Jacod and Mémin, a representation property for double martingales is given. Some L_2 and L_1 existence results are derived from the abstract existence theorems, and a maximum principle for the partially observable martingale control problem is given.

1. Introduction. This paper gives a fairly general approach to the stochastic optimal control problem, using recent results on semimartingales, especially on their exponentials and transformations. We attempt to illustrate where, in the development both of existence results and optimality criteria, the different sets of usual assumptions and specializations become necessary. Therefore, we proceed from the most general situation we could deal with to the more specialized examples and their more easily applicable theorems.

After reviewing, in § 2, the pertinent results of abstract martingale theory, we describe in § 3 an abstract model, which is then interpreted as a stochastic optimal control problem. The influence of the control is modeled by exponentials of martingales, while the cost is described by an integrable semimartingale. This gives a synthesis of the martingale approaches of Davis and Varaiya [6] on the one hand, and Striebel [25] on the other hand. In § 3.3 known necessary and sufficient conditions for optimality are adapted to our model. In § 3.2 we give some abstract existence results for the case where the likelihood ratios, describing the dynamics, are subsets of an L_2- or of an L_1-space. These existence results are applicable even in the partial observation case.

In § 4 we investigate the simplification obtained by assuming that the control martingales are represented as stochastic integrals over a fixed family of basic (locally square integrable) martingales. The results can then be given in a more explicit form. More easily verified, topological, conditions on the class of integrands describing the admissible control martingales, guarantee existence of an optimal control. Also, for the complete information case, in § 4.3 we derive an optimality criterion in the form of a maximum principle.

* Received by the editors February 10, 1978, and in final revised form November 15, 1979.

† Engineering Division, Rijksunversiteit te Gent, St. Pietersnieuwstraat 25, Ghent, Belgium, as research associate of NFWO (Belgian National Foundation of Scientific Research). During part of his work the author was a guest of the SFB72, B1, Deutsche Forschungsgemeinschaft, Bonn, Germany.

‡ Institut fur Angewandte Mathematik der Universitat Bonn, Wegelerstrasse 6, 5300 Bonn, Germany (BRD). The work of this author was partly supported by the SFB 72 of the Deutsche Forschungsgemeinschaft.

All these results are applied to the double martingale control problem, in § 5. Double martingales are sums of integrals over Brownian motion and a compensated jump process. Combining results of Elliott [13] and Jacod and Mémin [17], a representation property for double martingales is proved in § 2. This is combined with the results of § 4 to obtain some L_2 and L_1 existence theorems, and a maximum principle for the partially observable control problem. In [27] we treat this same problem, which was also recently studied by Gertner and Rapaport [14] who used martingales in a slightly different way. For a more traditional control engineering discussion of systems with both Wiener noise and jump process noise, and for applications, see Sworder [23].

Summarizing the role played by martingales in the present approach to stochastic control, the following main properties can be distinguished:

(*a*) The "preservation of mean" property of martingales simplifies the statement of the principle of optimality.

(*b*) The integral representation property of martingales allows identification of the value function with a stochastic integral (and interpretation of the integrand as a dual variable).

(*c*) The semimartingale decomposition allows representation of the state vector and the cost.

(*d*) The explicit form of the martingale exponential allows us to use the Girsanov theorem in describing models. Notice that (*a*) and (*b*) deal with the abstract model, while (*c*) and (*d*) only serve for the interpretation of the abstract model in a particular application.

2. Mathematical preliminary: review of martingale theory. Some recent results on martingale theory, useful for our abstract model, will be reviewed. For all details the reader is referred to Meyer [21] and Jacod and Yor [16]. The results on double martingales at the end of this section are new and will be proved.

We start with a probability space $(\Omega, \mathcal{F}, \mathcal{P})$, and an increasing family of σ-algebras $(\mathcal{F}_t)_{t\in[0,1]}$, $\mathcal{F} = \bigvee_{t\in[0,1]} \mathcal{F}_t = \mathcal{F}_1$. The family $(\mathcal{F}_t)$ always satisfies the usual conditions of completeness and right continuity, and $\mathcal{F}_0$ contains all $\mathcal{P}$ null sets in $\mathcal{F}$. All stochastic processes (x_t) are assumed $\mathcal{F}_t$-adapted, i.e., x_t is $\mathcal{F}_t$-measurable. A stochastic process is called predictable, if it is measurable, as a mapping on $(\Omega\times R_+, \mathcal{F}\otimes\mathcal{B}_t)$, with respect to the σ-algebra generated by all adapted processes which are left-continuous on $(0, 1]$.

The family of all uniformly integrable $(\mathcal{F}_t, \mathcal{P})$-martingales, m_t, with $m_0 = 0$, is denoted by $\mathcal{M}^1$ (or $\mathcal{M}^1(\mathcal{F}_t, \mathcal{P})$ if necessary to avoid ambiguities). Similarly $\mathcal{M}^2 := \{m_t \in \mathcal{M}^1, Em_1^2 = \sup_{t\in[0,1]} Em_t^2 < \infty\}$ is the family of square integrable martingales. The class of local martingales, $\mathcal{M}^1_{\text{loc}}$, is defined by

$$\begin{aligned}\mathcal{M}^1_{\text{loc}} &:= \mathcal{M}^1_{\text{loc}}(\mathcal{F}_t, \mathcal{P})\\ &= \{m_t \mid \exists T_n,\ T_n \text{ an } \mathcal{F}_t\text{-stopping time, } T_n \uparrow 1 \text{ w.p.1, } m_{t\wedge T_n} \in \mathcal{M}^1\}\end{aligned}$$

The family of all adapted processes a_t with integrable variation and a_0 integrable, is denoted by $\mathcal{A}$. $\mathcal{M}^2_{\text{loc}}$ and $\mathcal{A}_{\text{loc}}$ are defined in the same way as $\mathcal{M}^1_{\text{loc}}$. The class $\mathcal{S}_p$ of "special semimartingales" (further on called semimartingales for brevity) contains all processes x_t of the form

$$x_t = m_t + a_t, \quad m_t \in \mathcal{M}^1_{\text{loc}}, \quad a_t \in \mathcal{A}_{\text{loc}}, \quad a_t \text{ predictable.}$$

This decomposition is then automatically unique. In particular, if $m_t \in \mathcal{M}^2_{\text{loc}}$, then $m_t^2 \in \mathcal{S}_p$ is a submartingale; its increasing predictable part is denoted by $\langle m\rangle_t$, (i.e., $m_t^2 - \langle m\rangle_t \in \mathcal{M}^1_{\text{loc}}$), and is called the predictable variation of the local martingale m_t.

When $m_t, n_t \in \mathcal{M}^1_{\text{loc}}$, the product $m_t \cdot n_t$ is not automatically a special semimartingale; if it is—e.g., if one is locally bounded or if both are in $\mathcal{M}^2_{\text{loc}}$—its predictable part $\langle m, n\rangle_t$ is defined by the property $m_t n_t - \langle m, n\rangle_t \in \mathcal{M}^1_{\text{loc}}$. It is called the predictable covariation.

Every local martingale m_t can be uniquely decomposed into a continuous and a purely discontinuous part; the latter will also be referred to as the compensated jump part of m_t:

$$m_t = m^c_t + m^d_t.$$

Since $m^c_t \in \mathcal{M}^2_{\text{loc}}$, the adapted covariation

$$[m, n]_t = \langle m^c, n^c\rangle_t + \sum_{s \leqq t} \Delta m_s \, \Delta n_s \in \mathcal{A}_{\text{loc}}$$

exists for all local martingales.

This fact allows us to define stochastic integrals for the class LB of locally bounded integrands, where

$LB := \{h_t | h_t$ is $\mathcal{F}_t$-predictable and there exists a sequence of $\mathcal{F}_t$-stopping times T_n, $T_n \uparrow 1$ with probability 1, $|h_{t\wedge T_n}| \leqq K_n\}$. Then if $m_t \in \mathcal{M}^1_{\text{loc}}$, $h_t \in LB$, there exists a unique local martingale $l = h \circ m$ such that for all $n \in \mathcal{M}^1_{\text{loc}}$,

$$[l, n]_t = (h \circ [m, n])_t = \int_0^t h_s \, d[m, n]_s.$$

Since $l_t(\omega)$ coincides with $\int_0^t h_s(\omega)\, dm_s(\omega)$ for all ω such that the Stieltjes integral is defined (i.e., either $h_t(\omega)$ or $m_t(\omega)$ of integrable variation), we write,

$$(h \circ m)_t = \int_0^t h_s \, dm_s.$$

If m_t is the Brownian motion process w_t, the above defined stochastic integral is indistinguishable from the Ito integral. If m_t is the compensated jump martingale,

$$m_t = I_{t \geqq T} - a_{t\wedge T}$$

where a_t is increasing, then

$$(h \circ m)_t = h_T \cdot I_{t\geqq T} - \int_0^{t\wedge T} h_s \, da_s.$$

We shall very often use the Ito–Meyer differentiation rule: Let $x_t = m_t + a_t \in \mathcal{S}_p$, f a twice continuously differentiable function; then,

$$f(x_t) = f(x_0) + \int_0^t f'(x_{s-})\, d(m_s + a_s) + \tfrac{1}{2}\int_0^t f''(x_s)\, d\langle m^c\rangle_s$$
$$+ \sum_{0 \leqq s \leqq t} [f(x_s) - f(x_{s-}) - f'(x_{s-})\, \Delta x_s] \in \mathcal{S}_p.$$

A first example of its use is the proof by simple arithmetic that the stochastic differential equation, for $x_t \in \mathcal{S}_p$,

$$l_t = 1 + \int_0^t l_{s-}\, dx_s$$

has as its (unique) solution ($l_0 = 1$),

$$\varepsilon(x)_t = l_t = \exp\left(x_t - \tfrac{1}{2}\langle x^c\rangle_t\right) \cdot \prod_{s \leqq t} (l + \Delta x_s)\, e^{-\Delta x_s}.$$

If $x_t \in \mathcal{M}^1$ and bounded, then $\varepsilon(x)_t \in \mathcal{M}^1$. For a positive local martingale $x_t \in \mathcal{M}^1_{\text{loc}}$, we have $\varepsilon(x)_t \in \mathcal{M}^1_{\text{loc}}$, and $\varepsilon(x)_t$ is also a positive supermartingale if moreover $1+\Delta x_s \geqq 0$.

Whenever we use the notation $\varepsilon(n)_t$ in this paper, we assume that $\varepsilon(n)_t \geqq 0$ and $E\varepsilon(n)_1 = 1$ (equivalently $\varepsilon(n)_t$ is uniformly integrable). For local martingales n_t used in $\varepsilon(n)_t$, we also assume that for $x_t \in \mathcal{M}^1_{\text{loc}}$ (x_t will become a standard notation for the state process in our model in § 3), the predictable covariation $\langle x, n\rangle_t$ (with respect to $\mathcal{F}_t$ and $\mathcal{P}$) exists. These assumptions allow us to apply Girsanov's theorem (as generalized by van Schuppen and Wong [24]): For x_t and n_t as above, define the new probability measure $\mathcal{P}_n$ on $(\Omega, \mathcal{F})$ by $d\mathcal{P}_n/d\mathcal{P} \triangleq \varepsilon(n)_1$. Then the $(\mathcal{F}_t, \mathcal{P})$-local martingale x_t is transformed into an $(\mathcal{F}_t, \mathcal{P}_n)$-semimartingale such that

$$x_t - \langle x, n\rangle_t \in \mathcal{M}^1_{\text{loc}}(\mathcal{F}_t, \mathcal{P}_n).$$

Note that Yoeurp [26] has shown that any $\mathcal{P}$-semimartingale x_t is also a $\mathcal{P}_n$-semimartingale if and only if $\langle x, n\rangle_t$ exists (with respect to $\mathcal{P}$).

Now let $(Z, \mathcal{Z})$ be a measurable space. Define on $(\Omega, \mathcal{F}, \mathcal{F}_t, \mathcal{P})$ the random measure $q(A, s, \omega): Z \times R_+ \times \Omega \to R_+$, with the following properties:

(i) $q(A, t) \in \mathcal{M}^2_{\text{loc}}(\mathcal{F}_t, \mathcal{P})$ for each $A \in \mathcal{Z}$.

(ii) $\langle q(A, \cdot), q(A, \cdot)\rangle_t$ is a random measure on $(Z, \mathcal{Z})$, also denoted by $\langle q\rangle(A, t)$.

(iii) $\langle q(A, \cdot), q(B, \cdot)\rangle_t = 0$ for $A, B \in \mathcal{Z}$, $A \cap B = \emptyset$, $t \in [0, 1]$. Define

$$L^1(q) = L^1(\langle q\rangle) := \left\{ h(z, t) | h \quad \text{is } \mathcal{F}_t\text{-predictable [2], [17]}, \right.$$
$$\left. E\int_Z \int_0^1 |h(z, s)| \langle q\rangle(dz, ds) < \infty \right\}.$$

$L^1_{\text{loc}}(q)$, $L^2(q)$ and $L^2_{\text{loc}}(q)$ are defined in an obvious way (which coincides with the definition in [17]). We shall say that $(\Omega, \mathcal{F}, \mathcal{F}_t, \mathcal{P})$ has the (martingale) representation property with respect to the random measure $q(A, t)$ if for all $n \in \mathcal{M}^1_{\text{loc}}(\mathcal{F}_t, \mathcal{P})$, n_t a local martingale, there exists a predictable process $h \in L^1_{\text{loc}}(q)$ such that

$$n_t = \int_Z \int_0^t h(z, s) q(dz, ds).$$

Existence of the stochastic integral is part of this assumption.

Examples.

(1) Let $Z = \{1, \cdots, n\}$ and $\langle n(i), n(j)\rangle_s = \delta_{ij}s$, $n_t(i)$ continuous, then the above property describes the well known result of Kunita and Watanabe on the representation of Wiener martingales.

(2) Let $(Z, \mathcal{Z})$ be a Blackwell space, x_t a fundamental jump process in the sense of [2], [7], with

$$q(A, s) = p(A, s) - \tilde{p}(A, s), \quad A \in Z,$$

where $p(A, s)$ is the semimartingale associated with (x_t), i.e.,

$$p(A, t) = \sum_{s \leqq t} I_{\{x_s \neq x_{s^-}\}} I_{\{x_s \in A\}}.$$

Then the above property describes the result on the representation of fundamental jump processes with

$$L^1_{\text{loc}}(q) = L^1_{\text{loc}}(p) = L^1_{\text{loc}}(\tilde{p}).$$

Remark. We shall need later on that, assuming $E\tilde{p}(Z, 1) < \infty$, then for $m \in \mathcal{M}^1$ there exists $h \in L^1(q)$ such that

$$m \in \mathcal{M}^1 \quad \Leftrightarrow \quad m_t = \int_Z \int_0^t h(z, s) q(dz, ds).$$

The proof of Elliott for the single jump case [11] can be adapted.

(3) As a third example of the representation property we now discuss double martingales. This is done in more detail since all results of the paper are applied to this case. Suppose $(\Omega, \mathcal{F}, \mathcal{P})$ is the product space obtained by joining a Wiener space $(\mathcal{C}[0, 1], \mathcal{W}, \mathcal{P}^1)$ and a space of jump processes $(\mathcal{D}[0, 1], \mathcal{F}, \mathcal{P}^2)$ (where $\mathcal{D}[0, 1]$ contains all sample paths which are constant except for a finite number of jumps):

$$\Omega = \Omega' \times \Omega'' = \mathcal{C}[0, 1] \times \mathcal{D}[0, 1].$$

This means that $\mathcal{F}_t = \mathcal{W}_t \otimes \mathcal{Z}_t$ is generated by a Brownian motion $w_t(\omega')$ and an independent jump process $z_t(\omega'')$ with $(\mathcal{Z}_t, \mathcal{P})$-Lévy system $(n(A, t), \Lambda(t))$. Let $p(A, t)$ be the counting measure associated with z_t, $\tilde{p}(A, t)$ its predictable projection, and $q(A, t) = p(A, t) - \tilde{p}(A, t)$ the basic martingales, $A \in \mathcal{Z}$. Then for $\phi_s(\omega', \omega'') \in L^2_{\text{loc}}(w)$ and $\psi(z, s)(\omega', \omega'') \in L^1_{\text{loc}}(q)$, the following stochastic integral is well defined:

$$x_t = \int_0^t \phi_s \, dw_s + \int_Z \int_0^t \psi(z, s) q(dz, ds) \in \mathcal{M}^1_{\text{loc}},$$

(by the definition of $L^1(q)$ above, ϕ_s and $\psi(z, s)$ are $\mathcal{F}_t$-predictable). Note that the first integral is continuous and a.s. of unbounded variation, while the second is purely discontinuous (i.e., a compensated sum of jumps).

The translation theorem allows us to construct other probability measures on the above defined space, such that the Brownian motion and the jump part are no longer independent. Let $\phi \in L^2_{\text{loc}}(w)$, $\psi \in L^1_{\text{loc}}(q)$; then the exponential of

$$m_t = m_0 + \int_0^t \phi_s \, dw_s + \int_Z \int_0^t \psi(z, s) q(dz, ds)$$

can explicitly be written as

$$\varepsilon(m)_1 = \exp\left[\int_0^1 \phi_s \, dw_s - \tfrac{1}{2} \int_0^1 \phi_s^2 \, ds - \int_Z \int_0^1 \psi(z, s) n(dz, s) \Lambda(ds)\right] \cdot \prod_{s \leqq 1} [1 + \psi(z_s, s) I_{\{z_{s-} \neq z_s\}}].$$

Since we assume that $E\varepsilon(m)_1 = 1$ (and since $\langle m, w\rangle_t$, $\langle m, q(A, \cdot)\rangle_t$ exist), we can define the new probability space $(\Omega, \mathcal{F}, \mathcal{F}_t, \mathcal{P}_m)$ by $d\mathcal{P}_m / d\mathcal{P} = \varepsilon(m)_1$. Then $w_t - \int_0^t \phi \, ds$ is a $\mathcal{P}_m$-Brownian motion, now dependent on the jump process $z_t(\omega)$, which has the $(\mathcal{F}_t, \mathcal{P}_m)$-Lévy system:

$$\left(\int_A (1 + \psi(z, t)) n(dz, t), \Lambda(t)\right)^{1} .$$

For the space $(\Omega, \mathcal{F}, \mathcal{P})$ where the Brownian motion and the jump process are independent, Elliott [13] has shown that any local martingale is a stochastic integral over w_t and $q(A, t)$. This shows the existence of integrands which are integrable and

[1] As in [4], and unlike [2], we do not insist on $n(Z, t) = 1$. This is for notational convenience only.

measurable in the following restricted sense,

$$\phi_t(\omega', \omega'') \in L^2_{\mathrm{loc}}(\Omega', w) \quad \text{for each fixed } \omega'',$$

$$\psi(z, t, \omega', \omega'') \in L^1_{\mathrm{loc}}(\Omega'', q) \quad \text{for each fixed } \omega'.$$

However, Davis [29] has shown $\phi \in L^2_{\mathrm{loc}}(\omega)$ and $\psi \in L^1_{\mathrm{loc}}(q)$ when the jump process is a Poisson process. Recently, J. Jacod and M. Yor proved the following result: For any $K_t \in \mathcal{M}^2(\mathcal{W}_t \otimes \mathcal{Z}_t, \mathcal{P})$, $K_0 = 0$, there exist unique, $(\mathcal{W}_t \otimes \mathcal{Z}_t)$-predictable processes $\phi \in L^2_{\mathrm{loc}}(w)$ and $\psi \in L^2_{\mathrm{loc}}(q)$, such that

$$K_t = \int_0^t \phi_s \, dw_s + \int_Z \int_0^t \psi(z, s) q(dz, ds).$$

Proof. (published with the permission of J. Jacod and M. Yor): It suffices to prove the assertion for the dense subset of $\mathcal{M}^2$ consisting of $K_t = E(u \cdot v | \mathcal{W}_t \otimes \mathcal{Z}_t)$ where u and v are bounded, respectively $\mathcal{W}_\infty$, and $\mathcal{Z}_\infty$ measurable random variables. Define u_t to be the right continuous version of

$$E(u|\mathcal{W}_t) = E(u|\mathcal{W}_t \otimes \mathcal{Z}_t),$$

and similarly v_t of

$$E(v|\mathcal{F}_t) = E(v|\mathcal{W}_t \otimes \mathcal{Z}_t).$$

Then from Ito's formula,

$$K_\infty = u_0 v_0 + \int_0^\infty v_{s-} \cdot du_s + \int_0^\infty u_{s-} \, dv_s + [u, v]_\infty$$

$$= u_0 v_0 + \int_0^\infty v_{s-} U_s \, dw_s + \int_0^\infty \int_Z u_{s-} V_s(z) q(dz, ds),$$

where U_s and V_s exist and satisfy the required measurability and integrability conditions by the $\mathcal{M}^2$-representation theorems for Wiener martingales and jump process martingales. Then

$$\phi_s = v_{s-} \cdot U_s \quad \text{and} \quad \psi(z, s) = u_{s-} \cdot V_s(z),$$

satisfy the required conditions.

Since for any $m_t \in \mathcal{M}^1_{\mathrm{loc}}$, there exists a strongly reducing sequence of $\mathcal{W}_t \otimes \mathcal{Z}_t$ stopping times T_n (see [21]) the above result can be applied to $m_{t \wedge T_n}$.

Theorem 8.3 of Jacod and Mémin [17] shows that the representation property is preserved if $\mathcal{P}$ is replaced by $\mathcal{P}_0 (\ll \mathcal{P})$. Hence we make the following very weak assumption:

(A_0) For a given probability space $(\Omega, \mathcal{F}, \mathcal{F}_t, \mathcal{P}_0)$ on which are defined a Brownian motion w_t and a jump process z_t, there exists a probability measure $\mathcal{P}$, such that $\mathcal{P}_0 \ll \mathcal{P}$ and such that $(\Omega, \mathcal{F}, \mathcal{F}_t, \mathcal{P})$ is isomorphic to the product probability space defined above, where (the images of) w_t and z_t are $\mathcal{P}$-independent.

THEOREM 2.1. *Under assumption* A_0, *for every* $m_t \in \mathcal{M}^1_{\mathrm{loc}}(\mathcal{F}_t, \mathcal{P}_0)$ *there exist unique predictable processes*

$$\phi \in L^2_{\mathrm{loc}}(w)$$

and

$$\psi \in L^1_{\mathrm{loc}}(q_0),$$

such that

$$m_t = \int_0^t \phi_s \, dw_s + \int_Z \int_0^t \psi(z, s) q_0(dz, ds).$$

Under the same assumption A_0, consider any

$$n_t = n_0 + \int_0^t h_s \, dw_s + \int_Z \int_0^t k(z, s) q_0(dz, ds),$$

($h \in L^2_{\text{loc}}(w)$, $k \in L^1_{\text{loc}}(q)$) such that $d\mathcal{P}_n/d\mathcal{P} = \varepsilon(n)_1$ is well defined. Then

$$w^n(t) = w_t - \int_0^t h_s \, ds \in \mathcal{M}^2_{\text{loc}}(\mathcal{F}_t, \mathcal{P}_n),$$

$$q^n(A, t) = p(A, t) - \int_A \int_0^t (1 + k(z, s)) \tilde{p}_0(dz, ds) \in \mathcal{M}^1_{\text{loc}}(\mathcal{F}_t, \mathcal{P}_n),$$

and w_t^n is a $\mathcal{P}_n$-Brownian motion. One easily verifies that the translated martingale (for m_t as in Theorem 2.1)

$$m_t - \langle m, n \rangle_t = \int_0^t \phi_s \, dw_s^n + \int_Z \int_0^t \psi(z, s) q^n(dz, ds)$$

has the same integrands ϕ_s and $\psi(z, s)$ for its integral representation, independent of the probability measure $\mathcal{P}_n$.

Remark. If $\mathcal{D}[0, 1]$ contains sample paths with accumulation points (as in Elliott [11]), the above results can only be stated for ψ locally bounded.

3. Abstract model of stochastic optimal control.

3.1 The model. In a remarkable 1971 paper [1], Beneš reformulated a problem of dynamic control with Wiener noise, into an optimization problem in Wiener space using the Girsanov transformation [15]. Thus he bypassed the problem of existence of solutions to a stochastic differential equation. Since then this approach has been used both to prove existence of optimal controls (Beneš [1], Duncan and Varaiya [10] with Wiener noise, Kohlmann [18] with jump process disturbances), and for deriving optimality conditions (Davis and Varaiya [6], Elliott [12] for Wiener noise, Boel and Varaiya [4], Davis and Elliott [8] and Kohlmann [18] with jump process disturbances). These successes suggest the use of this method for the abstract model below, which synthesizes methods and results of Striebel [25] and Boel and Varaiya [4].

To define our *abstract model* we assume as given the basic probability space $(\Omega, \mathcal{F}, \mathcal{F}_t, \mathcal{P})$ and an increasing family of sub-σ-algebras $\mathcal{G}_t \subset \mathcal{F}_t$, satisfying the usual conditions. The control decisions are expressed as follows. We are given a set $\mathcal{N}(\subset \mathcal{M}^1_{\text{loc}}(\mathcal{F}_t, \mathcal{P}))$ of control martingales n_t, and a set $\mathcal{U}$ of U-valued, $\mathcal{G}_t$-adapted decision rules u (i.e., $\mathcal{U} \subset \{u_t(\omega): [0, 1] \times \Omega \to U, u_t(\omega) \text{ is } \mathcal{G}_t\text{-adapted}\}$). With each choice $u \in \mathcal{U}$ there corresponds an $n^u \in \mathcal{N}$ such that (consistency condition):

(i) $u_s(\omega) = \tilde{u}_s(\omega), s \leqq t \Rightarrow n_s^u(\omega) = n_s^{\tilde{u}}(\omega), s \leqq t;$
(ii) $u_s(\omega') = u_s(\omega''), 0 \leqq t_1 \leqq s \leqq t_2 \leqq 1 \Rightarrow n_s^u(\omega') - n_{t_1}^u(\omega') = n_s^u(\omega'') - n_{t_1}^u(\omega'').$

In agreement with the assumptions in § 2 about exponential formulas, the set $\mathcal{N}$ of control martingales is further restricted such that for all $n \in \mathcal{N}$:

(i) $E\varepsilon(n)_1 = 1$ (or equivalently $\varepsilon(n)_t \in \mathcal{M}^1)'$,
(ii) $\varepsilon(n)_t \geqq 0;$

and we assume that the mapping from $\mathscr{U}$ to $\mathcal{N}$ is onto. The family of all admissible likelihood ratios is denoted $\mathscr{D}(\mathcal{N}):=\{\varepsilon(n)_1 | n \in \mathcal{N}\} \subset L^1(\Omega, \mathscr{F}, \mathscr{P})$. Hence $d\mathscr{P}_n/d\mathscr{P} = \varepsilon(n)_1$ defines, for each $n \in \mathcal{N}$, a probability measure $\mathscr{P}_n$. We will also use the notation $\mathscr{P}_u = \mathscr{P}_{n^u}$; for any $u \in \mathscr{U}$, $E_n(E_u)$ will denote $\mathscr{P}_n(\mathscr{P}_u)$-expectation.

To complete the description of the abstract model, we assume given a family of real-valued, integrable, $(\mathscr{F}_t, \mathscr{P}_n)$-semimartingales,

$$Y_t^n = m_t^{y,n} + a_t^{y,n},$$

where $m_t^{y,n} \in \mathscr{M}_{\text{loc}}^1(\mathscr{F}_t, \mathscr{P}_n)$. The purpose of the control problem is to find a $u^* \in \mathscr{U}$ (or $n^* \in \mathcal{N}$) such that

$$J(u) \triangleq E_u(Y_1^{n^u}) = E(Y_1^{n^u} \cdot \varepsilon(n^u)_1),$$

is minimized, i.e.,

$$J^* \triangleq J(u^*) \triangleq E(Y_1^{n^*} \cdot \varepsilon(n^*)_1) = \inf_{u \in \mathscr{U}} J(u).$$

We now *interpret* this model as a *stochastic optimal control problem*. Suppose given on $(\Omega, \mathscr{F}, \mathscr{F}_t, \mathscr{P})$ a real valued $(\mathscr{F}_t, \mathscr{P})$-semimartingale x_t, called the state process,

$$x_t = x_0 + m_t + a_t,$$

such that:

I$_1$. $\langle x, n\rangle_t$ (the $(\mathscr{F}_t, \mathscr{P})$-predictable covariation) exists for all $n \in \mathcal{N}$, $t \in [0, 1]$.

Then the state process x_t is also an $(\mathscr{F}_t, \mathscr{P}_n)$-semimartingale

$$x_t = x_0 + (m_t - \langle x, n\rangle_t) + (a_t + \langle x, n\rangle_t);$$

i.e., the predictable bounded variation part is changed. This is completely analogous to what happens to a controlled diffusion equation.

$$x_t = x_0 + \int_0^t \sigma_s \, dw_s + \int_0^t f(s, x_s, u_s) \, ds,$$

where the predictable bounded variation part is changed by

$$\int_0^t f(s, x_s, u_s) \, ds.$$

Taking

$$n_t^u = \int_0^t f(s, x_s, u_s) \, dw_s,$$

would transcribe the controlling diffusion problem into the form of our abstract model.

Note that here as in most applications there is a stochastic process $w_t(\omega)$, a fixed function of ω, underlying the control martingales $n_t^u(\omega)$ and the state $x_t^u(\omega)$. The dynamics of the problem are expressed by the change of probability measure on $(\Omega, \mathscr{F})$. Note also that if the "state of the world", modeled by x_t, is vector valued, the above assumptions and equations are to be considered for each component. Notice also that the semimartingale representation results of § 3 are only applied to the cost, which we assume real valued.

The increasing σ-algebras $\mathscr{G}_t$ describe the information available for the control decisions. The larger $\mathscr{G}_t$, the more information available for influencing n_t. The special case $\mathscr{G}_t = \mathscr{F}_t$, complete information about the past state of the system, will lead to considerable simplifications.

Remarks.

(i) The assumption that $u_t(\omega)$ is $\mathcal{G}_t$-adapted does not imply that n_t^u is $\mathcal{G}_t$-adapted, as the above example of a diffusion control problem illustrates.

(ii) $(\mathcal{F}_t, \mathcal{P})$ and $(\mathcal{G}_t, \mathcal{P})$ are assumed to satisfy the usual assumptions of completeness and right continuity. However there is no guarantee that all $\mathcal{P}_n$-null sets will be in $\mathcal{F}_t$ or $\mathcal{G}_t$. Thus we will have to be careful with statements such as "a.s. $\mathcal{P}_n$", or when using theorems involving the choice of appropriate versions. For a detailed study of this difficulty see [5].

To complete the interpretation of the abstract model as a control problem, it is necessary to make the following assumption (similar to I_1):

I_2. $\langle Y^n, n\rangle_t$ (the $(\mathcal{F}_t, \mathcal{P}_n)$-predictable covariation) exists for all $n \in \mathcal{N}$, $t \in [0, 1]$.

Often the cost in an optimal control problem is of the form

$$J(u) = E_{n^u} \int_0^1 c^u(t, x_t)\, dt,$$

$c^u(\cdot\,, \cdot)$: $[0, 1] \times \mathbf{R}^n \to \mathbf{R}_+$, jointly measurable. Choosing $Y_t^u = \int_0^t c^u(s, x_s)\, ds$ clearly defines an $(\mathcal{F}_t, \mathcal{P}_n)$-semimartingale, and

$$J(u) = E_{n^u} Y_1^u$$

completes the description of an abstract control problem. For problems with a terminal cost $J = E_n(Y_1)$ we obtain a well defined abstract problem by taking as cost $Y_t^n = E_n(Y_1 | \mathcal{F}_t)$.

It is often useful to have Y_t independent of n. This can be achieved by proceeding as in Beneš [1], extending the probability space Ω by adjoining to it a Brownian motion (ω_t^0), independent of x_t under $\tilde{\mathcal{P}}$. The new probability measure $\tilde{\mathcal{P}}_n$ (on the extended space $(\tilde{\Omega}, \tilde{\mathcal{F}})$) is chosen such that w_t^0 has $\tilde{\mathcal{P}}_n$-drift $c^n(t, x_t)$, i.e.,

$$\tilde{\varepsilon}(n)_1 = \frac{d\tilde{\mathcal{P}}_n}{d\tilde{\mathcal{P}}} = \exp\left[\int_0^1 c^n(t, x_t)\, dw_t^0 - \int_0^1 \tfrac{1}{2}[c^n(t, x_t)]^2\, dt\right] \cdot \varepsilon(n)_1.$$

(Here $\tilde{\varepsilon}$, $\tilde{E}$ refer to the extended space; for details see [1]). Then

$$J(n) = E_n \int_0^1 c^n(t, x_t)\, dt = \tilde{E} w_1^0 \cdot \tilde{\varepsilon}(n)_1 = \tilde{E}_n w_1^0,$$

and choosing $Y_t^n = w_t^0$ again defines the abstract model. Instead of extending Ω by adjoining a Wiener process, one could also have added a counting process (p_t) with rate $c^n(t, x_t)$ under $\mathcal{P}_n$, and identify $Y_t^n = p_t$ (see [18]).

To simplify the statement of theorems later on, we formulate some frequently used assumptions.

N1. (Closure under concatenation for $\mathcal{G}_t$). Let $n^1, n^2 \in \mathcal{N}$. Then for all $A \in \mathcal{G}_t$ the process n, defined by

$$n_s = n_s^1, \qquad s \leqq t,$$
$$n_s = n_s^1 I_A + n_s^2 I_{A^c}, \qquad s > t,$$

is an element of $\mathcal{N}$.

C1. $E_n|Y_t^n| < \infty$ for all $n \in \mathcal{N}$, and $\int_{t_1}^{t_2} (da_s^{Y,n}) \geqq -\int_{t_1}^{t_2} f_s\, ds$ with f_t a deterministic, integrable function $(0 \leqq \int_0^1 f_t\, dt < \infty)$.

C2. (causality of cost): For any $A \in \mathcal{G}_t$, if $n_s^1(\omega) = n_s^2(\omega)$, $\omega \in A$, $0 \leqq t_1 \leqq s \leqq t_2 \leqq 1$, for some $n^1, n^2 \in \mathcal{N}$, then

$$(Y_{t_2}^{n^1} - Y_{t_1}^{n^1})I_A = (Y_{t_2}^{n^2} - Y_{t_1}^{n^2})I_A.$$

C3. For all $n \in \mathcal{N}$: $Y_t^n = Y_t$.

N1 and C2 will be used to define a value function, as a first step in the dynamic programming approach; C1 is necessary to define conditional expectations of the cost. Finally C3 insures that the cost $J(n)$ is a linear functional of $\varepsilon(n)_1$ which will be used for existence results.

3.2 Abstract existence theorem. In this section we give two abstract existence results, the first under the assumption that $\mathcal{D}(\mathcal{N}) \subset L_2(\Omega, \mathcal{F}, \mathcal{P})$, the second for $\mathcal{D}(\mathcal{N}) \subset L_1(\Omega, \mathcal{F}, \mathcal{P})$.

THEOREM 3.2.1. *For the abstract model of* § 3.1, *assume* C3 *holds, let* $|Y_1| \leqq K$, *and let* $\mathcal{D}(\mathcal{N})$ *be weakly closed,* $L_2(\Omega, \mathcal{F}, \mathcal{P})$. *Then there exists an optimal control, i.e. there is an* $n^* \in \mathcal{N}$ *such that* $J(n^*) \leqq J(n)$ *for all* $n \in \mathcal{N}$.

Proof. Since Y_1 is bounded, $J(n) = E(Y_1 \cdot \varepsilon(n)_1)$ is a continuous linear functional on $\mathcal{D}(\mathcal{N})$. As a weakly closed subset of the weakly compact unit ball in L_2, $\mathcal{D}(\mathcal{N})$ is weakly compact in L_2. A continuous linear functional attains its minimum over a weakly compact set.

The proof of the L^1 existence result is a bit more laborious. A subset $\mathcal{D} \subset L_1(\Omega, \mathcal{F}, \mathcal{P})$ is called strongly uniformly integrable if (cf. [20, *II-T* 22]):

$$\sup \{E(|d|^\gamma)|\, d \in \mathcal{D}\} < \infty, \quad \text{for some } \gamma > 1.$$

We will repeatedly require the following assumption, ($\chi_n = I_{\{|Y_1|^q > N\}}$).

E1. For some $\gamma > 1$ (depending on the context) let q be such that $1/\gamma + 1/q = 1$, then assume

$$E(|Y_1|^q \cdot I_{\{|Y_1|^q > N\}}) \underset{n\to\infty}{\longrightarrow} 0.$$

THEOREM 3.2.2. *Let* $\mathcal{D}(\mathcal{N})$ *be strongly uniformly integrable and weakly closed* (*in* L_1), *and assume* E1 *to be fulfilled*; *then there exists an optimal control.*

Proof. $\mathcal{D}(\mathcal{N})$ is bounded and weakly closed in $L_1(\Omega, \mathcal{F}, \mathcal{P})$, hence weakly sequentially compact. There exists a minimizing sequence (n_i) such that $\varepsilon(n_i)_1$ converges weakly to an element $\varepsilon(n)_1 \in \mathcal{D}(\mathcal{N})$. In the inequality

$$|E(Y_1 \chi_N \varepsilon(n)_1)| \leqq E^{1/q}(\chi_N |Y_1|^q) E^{1/\gamma}(\varepsilon(n)_1^\gamma),$$

the second factor is uniformly bounded, and the first goes to zero when N goes to infinity, by E1. Now we choose N so large that for given $\varepsilon > 0$,

$$|E(Y_1 \chi_N \varepsilon(n)_1)| < \varepsilon/3$$

for all $\varepsilon(n)_1 \in \mathcal{D}(\mathcal{N})$. Then the weak L_1-convergence of $\varepsilon(n_i)$ allows us to choose i_0 so that for $i \geqq i_0$ the following equality holds,

$$|E(Y_1(1-\chi_N)(\varepsilon(n)_1 - \varepsilon(n_i)_1))| < \varepsilon/3$$

This finally implies that

$$|J(n_i) - J(n)| < \varepsilon, \qquad i > i_0.$$

Since ε was arbitrary, $n = n^*$ is an optimal control martingale.

Remark. The above theorems hold for partial as well as complete information. This is related to the fact that we have made strong assumptions involving only the random variable Y_1 and the topological properties of the set $\mathcal{D}(\mathcal{N})$.

3.3 Optimality conditions. Consider the abstract problem of § 3.1 and assume C1 and C2 (integrability and causality of the cost) are also satisfied. Using

$$\begin{aligned}E_n(Y_1^n|\mathcal{G}_t) &= E_n(Y_t^n|\mathcal{G}_t)+E_n(Y_1^n - Y_t^n|\mathcal{G}_t)\\ &= E_n(Y_t^n|\mathcal{G}_t)+E_n(a_1^{y,n}-a_t^{y,n}|\mathcal{G}_t)\\ &> E_n(Y_t^n|\mathcal{G}_t)-\int_t^1 f_s\,ds,\end{aligned}$$

the value function $U_t(n^t)$ can be defined by,

$$U_t(n^t) \triangleq E_n(Y_t^n|\mathcal{G}_t)+\mathcal{P}_n-\operatorname*{ess\,inf}_{\tilde{n}\in\mathcal{N}_{n,t}} E_{\tilde{n}}(Y_1^{\tilde{n}}-Y_t^{\tilde{n}}|\mathcal{G}_t).$$

Here n^t denotes $(n_s, s \leqq t)$ and

$$\mathcal{N}_{n,t} := \{\tilde{n}\in\mathcal{N}|\tilde{n}_s = n_s,\ s\leqq t\}.$$

To avoid trivialities, we assume that for each $n\in\mathcal{N}$ and each $t\in[0,1]$ there exists at least one $\tilde{n}\in\mathcal{N}_{n,t}$ such that $E_{\tilde{n}}\ (Y_1^{\tilde{n}}-Y_t^{\tilde{n}}|\mathcal{G}_t)$ is integrable. Then $E_n|U_t(n^t)|<\infty$ and $U_t(n^t)\geqq -\int_0^1 f_s\,ds$ for all t. Note that by C2, $U_t(n^t)$ only depends on the control used up to time t. Indeed for any $A\in\mathcal{F}_t$, $\tilde{n}\in\mathcal{N}_{n,t}$,

$$\begin{aligned}\mathcal{P}_n(A) &= \int_\Omega I_A\cdot\varepsilon(n)_1\cdot d\mathcal{P} = \int_\Omega I_A\cdot\varepsilon(n)_t\,d\mathcal{P}\\ &= \int_\Omega I_A\cdot\varepsilon(\tilde{n})_t\cdot d\mathcal{P} = \tilde{\mathcal{P}}_n(A),\end{aligned}$$

and hence $E_n(Z|\mathcal{G}_t) = E_{\tilde{n}}(Z|\mathcal{G}_t)$ for any $\mathcal{F}_t$-measurable random variable Z. Since $Y_t^n = Y_t^{\tilde{n}}$, when $\tilde{n}\in\mathcal{N}_{n,t}$, we can also interpret $U_t(n^t)$ as

$$U_t(n^t) = \mathcal{P}_n-\operatorname*{ess\,inf}_{\tilde{n}\in\mathcal{N}_{n,t}} E_{\tilde{n}}(Y_1^{\tilde{n}}|\mathcal{G}_t),$$

the smallest achievable total cost at time t. The minimal future expected cost will be denoted,

$$W_t(n^t) = \mathcal{P}_n-\operatorname*{ess\,inf}_{\tilde{n}\in\mathcal{N}_{n,t}} (Y_1^{\tilde{n}}-Y_t^{\tilde{n}}|\mathcal{G}_t).$$

It is now trivial to adapt the proof of Theorem 1 of Striebel [23] to our model.

THEOREM 3.3.1. *For the abstract model, assume* C1 *and* C2 *and assume there exists for each $n\in\mathcal{N}$ an integrable, $\mathcal{G}_t$-adapted process $\tilde{U}_t(n^t)$, bounded below by $-\int_0^1 f_s\,ds$, such that*

(i) $\forall\tilde{n}\in\mathcal{N}_{n,t}$: $\tilde{U}_t(\tilde{n}^t) = \tilde{U}_t(n^t)$,

(ii) $E_n(\tilde{U}_1(n^1)) = E_n(Y_1^n|\mathcal{G}_1)$,

(iii) $\tilde{U}_t(n^t)$ *is a* $(\mathcal{P}_n, \mathcal{G}_t)$*-submartingale.*

Further assume there exists $\hat{n}\in\mathcal{N}$ such that

(iv) $\tilde{U}_t(\hat{n}^t)$ *is a martingale.*

Then necessarily

$$\tilde{U}_t(\hat{n}^t) = U_t(\hat{n}^t) = E_{\hat{n}}(Y_1^{\hat{n}}|\mathcal{G}_t)$$

and $\hat{n}$ is an optimal control law with minimal cost

$$E_{\hat{n}}(Y_1) = \tilde{U}_0.$$

Moreover for all $n\in\mathcal{N}$

$$\tilde{U}_t(n^t)\leqq U_t(n^t).$$

In order to improve this result, and obtain a necessary and sufficient condition for optimality, one has to assume the ε-lattice property [25]:

Let $n^1, n^2\in\mathcal{N}$, $n_s^1=n_s^2$, $s\leqq t$; then there exists an $n\in\mathcal{N}$ such that

$$\begin{aligned} n_s&=n_s^1=n_s^2, && s\leqq t,\\ J(n)&\leqq J(n^i)+\varepsilon, && i=1,2.\end{aligned}$$

Intuitively this guarantees not only optimal control, but an optimal continuation of any control law (cf. [22]). It turns out that assumption N1, of closure under concatenation, guarantees this, independently of the cost structure.

LEMMA 3.3.2. *Assuming* N1, C1, C2, *then the abstract control problem has the* ε*-lattice property for all* $\varepsilon\geqq 0$.

Proof. Let

$$A=\{E_{n^1}(Y_1^{n^1}|\mathcal{G}_t)\leqq E_{n^2}(Y_1^{n^2}|\mathcal{G}_t)\}.$$

Then $A\in\mathcal{G}_t$, and

$$\begin{aligned} \tilde{n}_s&=n_s^1=n_s^2, && s\leqq t,\\ &=n_s^1, && s>t \text{ and } \omega\in A,\\ &=n_s^2, && s>t \text{ and } \omega\in A^c,\end{aligned}$$

defines $\tilde{n}\in\mathcal{N}$, and

$$\varepsilon(\tilde{n})_1=\varepsilon(n^1)_1I_A+\varepsilon(n^2)_1I_{A^c}.$$

Then

$$\begin{aligned} E_{\tilde{n}}(Y_1^{\tilde{n}})&=E[I_A\varepsilon(n^1)_1Y_1^{n^1}]+E[I_{A^c}\varepsilon(n^2)_1Y_1^{n^2}]\\ &\leqq E(\varepsilon(n^i)_1Y_1^{n^i}), \qquad i=1,2.\end{aligned}$$

Remark. Davis and Varaiya [6, Lemma 3.1] have shown that the property holds for any number of continuations of the control law, if $\varepsilon>0$ is taken, i.e., there is $n^\varepsilon\in\mathcal{N}_{n,t}$ such that

$$E_{n^\varepsilon}(Y_1^{n^\varepsilon}|_t)\leqq U_t(n^t)+\varepsilon.$$

Striebel [25] has shown that under the ε-lattice property, the value function $U_t(n^t)$ is a $(\mathcal{P}_n, \mathcal{G}_t)$-submartingale, which immediately leads to the same necessary and sufficient optimality condition as in [6], [4].

THEOREM 3.3.3. *For the above abstract model of* § 3.1, *assume* N1, C1, C2. *Then* $U_t(n^t)$ *is a* $(\mathcal{P}_n, \mathcal{G}_t)$ *integrable submartingale, such that* $U_1(n)=E_n(Y_1^n)=J(n)$. *It is a* (*uniformly integrable*) *martingale if and only if* n *is an optimal law.*

This is equivalent to the principle of optimality: For all $n\in\mathcal{N}$ and for all pairs of $\mathcal{G}_t$-stopping times $0\leqq\tau_1\leqq\tau_2\leqq 1$,

$$W_{\tau_1}(n^{\tau_1})\leqq E_n(a_{\tau_2}^{Y^n}-a_{\tau_1}^{Y^n}|\mathcal{G}_{\tau_1})+E(W_{\tau_2}(n^{\tau_2})|\mathcal{G}_{\tau_1}),$$

with end condition

$$W_1(n)=E(Y_1^n|\mathcal{G}_1),$$

and equality if and only if n is optimal.

The above theorems are simplified in an obvious way in the complete information case, $\mathcal{G}_t = \mathcal{F}_t$. The minimal future cost is then independent of the past control law: If $n_s^1 = n_s^2$, $s \geqq t$, then by C2,

$$Y_1^{n^1} - Y_t^{n^1} = Y_1^{n^2} - Y_t^{n^2},$$

and by definition

$$\frac{\varepsilon(n^1)_1}{\varepsilon(n^1)_t} = \frac{\varepsilon(n^2)_1}{\varepsilon(n^2)_t}.$$

Hence

$$E_{n^1}(Y_1^{n^1} - Y_t^{n^1} | \mathcal{F}_t) = E_{n^2}(Y_1^{n^2} - Y_t^{n^2} | \mathcal{F}_t),$$

and by taking the essential infimum,

$$W_t((n^1)^t) = W_t((n^2)^t).$$

However,

$$U_t(n^t) = Y_t^n + W_t$$

still depends on the control law n up to time t, through the past cost Y_t^n (except when C3 is satisfied). When $U_t(n^t)$ (or $W_t(n^t)$) is known to have a right continuous version, the Doob–Meyer decomposition theorem could be applied to obtain a result as in § 4.2 of [4]. Since this is much more easily stated with the additional martingale representation assumption, we will postpone it until § 4.3 and § 5.3.

4. Control martingales represented as stochastic integrals.

4.1. The representation property model. We now consider a special case of the abstract model. We assume that $(\Omega, \mathcal{F}, \mathcal{F}_t, \mathcal{P})$ has the martingale representation property defined in § 2, i.e., we are given a random measure $q(A, t) \in \mathcal{M}_{\text{loc}}^2(\mathcal{F}_t, \mathcal{P})$, with $\langle q\rangle(A, t)$ a random measure as explained in § 2, such that any $(\mathcal{F}_t, \mathcal{P})$-martingale is a stochastic integral over $q(A, t)$. The set of control martingales can hence be described by

$$\Phi := \left\{\phi(z, t, u_t) \in L_{\text{loc}}^1(q) \,\middle|\, \int_Z \int_0^t \phi(z, s, u_s) q(dz, ds) = n^\phi \in \mathcal{N}, 1 + \Delta n_t^\phi \geqq 0\right\}.$$

We assume from now on that u_t is $\mathcal{G}_t$-predictable ($\mathcal{F}_t$-predictable is actually sufficient), and that $\phi(z, t, u)$ is continuous in u for fixed z, t, to insure $\mathcal{F}_t$-predictability of ϕ. We denote $\mathcal{D}(\Phi) := \mathcal{D}(\mathcal{N})$. All other assumptions of the abstract model of § 3.1 remain unchanged.

4.2. Existence results. For the model of § 4.1, we prove the existence results corresponding to Theorems 3.2.1 and 3.2.2. We first consider the case $\Phi \subset L_{\text{loc}}^1(q)$. By a sequence of lemmas as in [18] it can be seen that the following holds:

LEMMA 4.2.1. *For the model of* § 4.1, *let* Φ *be closed and convex in* $L^2(q)$ *and assume* $\langle q\rangle(z, t) \leqq \mu(t)$, μ *a deterministic, increasing function of time, and finally assume full information* $\mathcal{F}_t = \mathcal{G}_t$. *Then* $\mathcal{D}(\Phi)$ *is closed and convex in* $L_2(\Omega, \mathcal{F}, \mathcal{P})$.

Now $\mathcal{D}(\Phi)$ is a weakly closed subset of $L_2(\Omega, \mathcal{F}, \mathcal{P})$, so that Theorem 3.2.1 can be applied.

THEOREM 4.2.2. *Under the assumptions of Lemma* 4.2.1, *suppose* C3 *is satisfied and* $|Y_1|$ *is bounded. Then there exists an optimal control.*

It would have been possible to give an L_1 existence result analogous to that in [18, T 2.2.3] by imposing some general and hence very restrictive assumptions on Φ, so that $\mathscr{D}(\Phi)$ becomes strongly uniformly integrable. Here we want to assume a special representation property for $(\Omega, \mathscr{F}, \mathscr{F}_t, \mathscr{P})$ yielding results that are more interesting and can be applied more directly to the double martingale control problem.

We assume that the continuous part of each $(\mathscr{F}_t, \mathscr{P})$-local martingale has the representation property with respect to a finite family of Gaussian processes y_t^i, i.e.,

$$m_t^c = \sum_{i=1}^{n} \int_0^t \phi_i(s)\, dy_s^i, \qquad \phi_i \in L^1_{\mathrm{loc}}(y^i),$$

where the y^i are continuous, and

$$\langle y^i, y^j \rangle_t = \delta_{ij} \int_0^t \beta^i(s)\, ds, \qquad 1 \leqq i, j \leqq k,$$

for some deterministic function $\beta^i(t) \geqq 0$.

The discontinuous part is assumed to be of the form

$$m_t^d = \int_Z \int_0^t \psi(z, s) r(dz, ds),$$

with $r(A, t)$ purely discontinuous. Summarizing:

$$m \in \mathscr{M}^1_{\mathrm{loc}}(\mathscr{F}_t, \mathscr{P}) \quad \Leftrightarrow \quad m_t = \sum_{i=l}^{k} \int_0^t \phi_i(s)\, dy_s^i + \int_Z \int_0^t \psi(z, s) r(dz, ds),$$

$$\phi = (\phi_1, \cdots, \phi_n, \psi) \in \Phi \subset \prod_{i=1}^{k} L^1_{\mathrm{loc}}(y^i) \times L^1_{\mathrm{loc}}(r),$$

$(y_t^i, i = 1, \cdots, k; r(A, t))$ can be interpreted as a random measure $q(A, t)$ replacing Z by $\{1, 2, \cdots, k\} \cup Z$. Under the above assumptions, for a control martingale n_t, $\varepsilon(n)$ decomposes multiplicatively

$$\varepsilon(n)_1 = \varepsilon(n^c)_1 \varepsilon(n^d)_1 = \varepsilon(\phi_1, \cdots, \phi_k)_1 \cdot \varepsilon(\psi)_1$$

$$= \prod_{i=1}^{k} \varepsilon(\phi_i) \cdot \varepsilon(\psi)_1 .$$

In order to apply Theorem 3.2.2 we need strong uniform integrability of $\mathscr{D}(\Phi)$:

$$\sup_{\phi \in \Phi} E[\varepsilon(\phi_1, \cdots, \phi_k)_1^\gamma \cdot \varepsilon(\psi)_1^\gamma], < \infty,$$

for some $\gamma > 1$. By Hölder's inequality it suffices to prove that there exist $\gamma_1 > 1$ and γ_2 sufficiently large, such that,

$$\sup_{\phi \in \Phi} E(\varepsilon(\phi_1, \cdots, \phi_k)_1^{\gamma_1}) E(\varepsilon(\psi)_1^{\gamma_2}) < \infty;$$

or such that

(i) $\sup_{\phi \in \Phi} E[\varepsilon(\phi_1, \cdots, \phi_k)_1^\alpha] < \infty$ for some $\alpha > 1$,

(ii) $\sup_{\phi \in \Phi} E[\varepsilon(\psi)_1^\alpha] < \infty$ for sufficiently large α.

LEMMA 4.2.3. (i) *is satisfied under the following assumptions*:

A^{c}1. *For all* $\phi \in \Phi$: $|\phi_i(t)|^2 \leqq K(1 + |\sum_{i=1}^k y_i(1)|^2)$ *for some finite* K.

A^{c}2. $0 \leqq \beta^i(t) \leqq \beta$, $i = 1, \cdots, k$, *i.e.*, $\beta^i(t)$ *bounded.*

Proof. *Step* 1. We now prove that (A^{c}1) and (A^{c}2) imply

$$E(\varepsilon(\lambda\phi_i, i = 1, \cdots, k)_1) = 1, \quad \text{for all } \lambda \geqq 1.$$

The development of $\varepsilon(\lambda\phi_i,\ i=1,\cdots,k)_1$ into its Hermite polynomials [21, IV-40] assures that $\varepsilon(\lambda\phi_i,\ i=1,\cdots,k)_1=\sum_{i=0}^{\infty}\lambda^i P_1^i$, where the Hermite polynomials are defined recurrently by:

$$P_1^0=1,$$

$$P_1^1=\sum_{i=1}^{k}\int_0^1 \phi_i(s)dy_s^i,$$

$$P_1^2=\frac{1}{2}\left[\left(\sum_{i=1}^{k}\int_0^1 \phi_i(s)\,dy_s^i\right)^2-\sum_{i=1}^{k}\int_0^1 \phi_i^2(s)\beta_s^i\,ds\right],$$

$$P_1^l=\frac{1}{l}\left[\left(\sum_{i=1}^{k}\int_0^1 \phi_i(s)\,dy_s^i\right)P_1^{l-1}-\left\langle\sum_{i=1}^{k}\int_0^1 \phi_i^{\varepsilon}(s)\,dy_s^i\right\rangle_1\cdot P_1^{l-2}\right]$$

$$=\int_0^1 P_s^{l-1}\cdot dP_s^1.$$

Let $C_N(t)$ be a convex cone in $\mathbb{R}^N$:

$$C_N(t):=\{0<u_1<u_2<\cdots<u_N<t\},$$

$$C_N(1)=C_N.$$

It easily follows, from the recurrence formula for Hermite polynomials and from Meyers [21, IV-50] results on multiple integrals, that

$$P_t^N=\sum_{i=1}^{k}\int_{C_N(t)}\phi_i(u_1)\cdots\phi_i(u_N)\cdot dy_{u_1}^i\cdots dy_{u_N}^i.$$

We find the following chain of inequalities

$$E\left|\sum_{i=1}^{k}\int_{C_N(t)}\phi_i(u_1)\cdots\phi_i(u_N)\,dy_{u_1}^1\cdots dy_{u_N}^1\right|^2$$

$$\leqq\beta^N E\left|\sum_{i=1}^{k}\int_{C_N(t)}\phi_i^2(u_1)\cdots\phi_i^2(u_N)\,du_1\cdots du_N\right|\quad\text{(by A}^c\text{2)}$$

$$\leqq K^N\beta^N kE\left[1+\sup_{t\in[0,1]}\left|\sum_{i=1}^{k}y^i(t)\right|^2\right]^N$$

$$\leqq K^N\beta^N k\left[1+\sum_{j=1}^{N}\binom{N}{j}\left(\frac{2j}{2j-1}\right)^{2j}E\left[\left|\sum_{i=1}^{k}y_1^i\right|^{2j}\right]\right]\quad\text{(by [9, T-VII-3-4]).}$$

This is finite since the y^i are Gaussian. Now [21, IV-48.3] can be applied since the assumptions of [21, IV-45] are fulfilled by A^{c}2. Hence

$$E\sum_{i=1}^{k}\int_{C_N}\phi_i(u_1)\cdots\phi_i(u_N)\,dy_{u_1}^i\cdots dy_{u_N}^i=0\quad\text{for each }N.$$

By Fatou's Lemma $\varepsilon(\lambda\phi_i,\ i=1,\cdots k)_t$ is a supermartingale with expectation between 0 and 1. Fubini's Theorem finally yields

$$E[\varepsilon(\lambda\phi_i,i=1,\cdots,k)_1]=1,\qquad\lambda\geqq 1.$$

The above result can also be derived from [15, Lemma 7] by slightly generalizing the methods in [1, Lemma 0].

Step 2. Under the assumptions A^{c}1 and A^{c}2, there is an $\alpha>1$ such that

$$E[\varepsilon(\phi_i,i=1,\cdots,k)_1^{\alpha}]<\infty$$

uniformly in $\phi \in \Phi$. To show this, consider the $\varepsilon(\alpha\phi_i, i=1, \cdots, k)_1\, d\mathscr{P}$-local martingale,

$$x_t = \sum_{i=1}^{k} y_t^i - \alpha \sum_{i=1}^{k} \int_0^t \phi_s^i \beta_s^i \, ds.$$

Since

$$\left|\sum_{i=1}^{k} y_t^i\right|^2 \leqq 2|x_t|^2 + 2\alpha^2 \left|\sum_{i=1}^{k} \int_0^t \phi_s^i \beta_s^i \, ds\right|^2$$
$$\leqq 2|x_t|^2 + 2\alpha^2\beta^2 \sum_{i=1}^{k} \int_0^t |\phi_s^i|^2 \, ds,$$

applying A^c2 yields

$$\sup_{t\in[0,1]} \left|\sum_{i=1}^{k} y_t^i\right|^2 \leqq 2 \sup_{t\in[0,1]} |x_t|^2 + 2kK\alpha^2\beta^2\left(1 + \int_0^1 \sup_{0\leqq s\leqq t} \left|\sum_{i=1}^{k} y_s^i\right|^2 \cdot dt\right).$$

Now the Gronwall inequality implies

$$\sup_{t\in[0,1]} \left|\sum_{i=1}^{k} y_t^i\right|^2 \leqq (2 \sup_{t\in[0,1]} |x_t|^2 + 2kK\alpha^2\beta^2)\, e^{2kK\alpha^2\beta^2},$$

and we can write the following chain of inequalities:

$$E(\varepsilon(\phi_1, \cdots, \phi_k)_1^\alpha) \leqq E\left(\varepsilon(\alpha\phi_i, i=1, \cdots, k)_1 \cdot \exp\frac{\alpha^2-\alpha}{2} \sum_{i=1}^{k} \int_0^1 \phi_s^{i^2} \beta_s^i \, ds\right)$$
$$\leqq E\left(\varepsilon(\alpha\phi_i, i=1, \cdots, k)_1 \cdot \exp\frac{\alpha^2-\alpha}{2}\beta \sum_{i=1}^{k} \int_0^1 (\phi_s^i)^2 \, ds\right)$$
$$\leqq E\left[\varepsilon(\alpha\phi_i, i=1, \cdots, k)_1 \cdot \exp\frac{\alpha^2-\alpha}{2}\beta k K\left(1 + \int_0^1 \sup_{s\in[0,t]} |\sum y_s^i|^2 \, dt\right)\right]$$
$$\leqq h(\alpha) E\left[\varepsilon(\alpha\phi_i, i=1, \cdots, k)_1 \cdot \exp\left(k \cdot K \cdot \beta\frac{\alpha^2-\alpha}{2} \cdot e^{2kK\alpha^2\beta^2} \cdot \sup_{t\in[0,1]} |x_t|^2\right)\right].$$

As $h(\alpha)$ is bounded for α near 1, and x_t is a Gaussian process under $\varepsilon(\alpha\phi_i, i=1, \cdots, k)_1 \cdot d\mathscr{P}$, the expectation is bounded uniformly in $\phi \in \Phi$ for some $\alpha > 1$.

LEMMA 4.2.4. *Condition* (ii), *i.e.*, $\sup_{\phi\in\Phi} E[\varepsilon(\psi)_1^\alpha] < \infty$ *for some* α, *is satisfied under the assumption*

$$\mathrm{A}^d\!: \qquad \left\{\int_Z \int_0^1 \psi(z, s) q(dz, ds) \middle| (\phi_1, \cdots, \phi_k, \psi) \in \Phi\right\} \subset L_1,$$

is an exponentially uniformly integrable subset; *i.e.*,

$$\sup_{\phi\in\Phi} E\left|\exp \alpha \int_Z \int_0^1 \psi(z, s) q(dz, ds)\right| < \infty.$$

Proof. Since $\Delta \int_Z \int_0^t \phi(z, s) q(dz, ds) \in [-1, \infty]$ by our general assumption, we have

$$0 \leqq \left(1 + \Delta \int_Z \int_0^t \phi(z, s) q(dz, ds)\right)^\alpha \cdot \exp\left(-\alpha\Delta \int_Z \int_0^t \phi(z, s) q(dz, ds)\right) \leqq 1,$$

(remembering that $(1+x)e^{-x} \leqq 1$ for $x \geqq -1$). Then we obtain the following chain of inequalities:

$$|E(\varepsilon(\psi)_1^\alpha)| \leqq E \exp\left(\alpha \int_Z \int_0^1 \psi(z, s)q(dz, ds)\right)$$

$$\prod_{0 \leqq t \leqq 1} \left(1 + \Delta \int_Z \int_0^t \psi(z, s)q(dz, ds)\right)^\alpha$$

$$\cdot \exp\left(-\alpha\Delta \int_Z \int_0^t \psi(z, s)q(dz, ds)\right)$$

$$\leqq E \exp\left(\alpha \int_Z \int_0^1 \psi(z, s)q(dz, ds) + 1\right) < \infty,$$

uniformly in ψ by A^{d}.

Combining these results we have

LEMMA 4.2.5. *Under* $\mathrm{A}^{\mathrm{c}}1$, $\mathrm{A}^{\mathrm{c}}2$ *and* A^{d}, $\mathscr{D}(\Phi)$ *is strongly uniformly integrable.*

We can now apply Theorem 3.2.2.

THEOREM 4.2.6. *Let Φ fulfill the assumptions* $\mathrm{A}^{\mathrm{c}}1$, $\mathrm{A}^{\mathrm{c}}2$ *and* A^{d}, *let the cost criterion satisfy $E1$, and let $\mathscr{D}(\Phi)$ be weakly closed in L_1. Then there exists an optimal control* $\phi^* \in \Phi$.

Remarks.

(i) A sufficient condition for A^{d} has been given in a more easily verifiable form [18]. The form of A^{d} was chosen here because it expresses the essential difficulty that would have been removed by a different assumption.

(ii) As we do not assume any convexity on the set of controls, Theorems 3.2.1, 3.2.2 and 4.2.6 also apply to the partial observation case.

4.3. Optimality conditions: The maximum principle. To illustrate the use of the martingale representation property we will here derive a maximum principle for the model of § 4.1, similar to the result of Elliott [12]. We will do this only under the following strong assumptions:

(i) There exists an optimal control $n^* \in \mathcal{N}$ (i.e., an optimal $\phi^* \in \Phi$).

(ii) All probability measures are equivalent: $\mathscr{P} \approx \mathscr{P}_\phi$, $(E\varepsilon(\phi))_1 = 1$ and $E(1/\varepsilon(\phi)_1) = 1$, in particular $\varepsilon(\phi)_1 > 0$ a.s. $\mathscr{P}$ and $\mathscr{P}_\phi$).

(iii) The cost semimartingale Y_t is independent of the control used (C3).

(iv) Complete information ($\mathscr{G}_t = \mathscr{F}_t$).

The most important condition is (ii), which guarantees that $L^1_{\mathrm{loc}}(q)$ is independent of the probability measures $\mathscr{P}_\phi$ ($T_n \uparrow 1 \mathscr{P}_\phi$ a.s. $\Rightarrow T_n \uparrow 1 \mathscr{P}$ a.s.). This condition (as well as (iii) and (iv)) can be considerably relaxed for double martingales, making use of specific properties of jump processes and Wiener processes. For the model of § 4.1, with the above assumptions (i) to (iv), we have for the value function, $U_t - U_0 = Y_t + W_t - U_0 \in \mathscr{M}^1(\mathscr{F}_t, \mathscr{P}^*)$ where

$$\frac{d\mathscr{P}^*}{d\mathscr{P}} = \varepsilon(n^*)_1 = \varepsilon(\phi^*)_1.$$

Hence there exists a predictable process $g^* \in L^1_{\mathrm{loc}}(q)$ such that

$$U_t = J^* + \int_Z \int_0^t g^*(z, s)(q(dz, ds) - \phi^*(z, s)\langle q\rangle(dz, ds)).$$

Now

$$E\int_0^{T_n}\int_Z |g^*(z,s)|\langle q\rangle(dz,ds)<\infty \quad \text{a.s. } \mathcal{P}_\phi,$$

(by definition of the random measure and A^{c}2), which implies [26, Thm. 1.13] that

$$\left\langle \int_Z\int_0^t g^*(z,s)q(dz,ds),\, q(A,\cdot)\right\rangle_t,$$

exists for all $\mathcal{P}_\phi$. By [26, Prop. 2.2.],

$$\int_Z\int_0^t g^*(z,s)(q(dz,ds)-\phi(z,s)\langle q\rangle(dz,ds))\in\mathcal{M}^1_{\mathrm{loc}}(\mathcal{F}_t,\mathcal{P}_\phi)\quad\text{for all } \phi\in\Phi.$$

The Doob–Meyer decomposition of the $\mathcal{P}_\phi$-submartingale U_t can be written explicitly as:

$$U_t=J^*+\int_0\int_0^t g^*(z,s)(q(dz,ds)-\phi(z,s)\langle q\rangle(dz,ds))$$
$$+\int_Z\int_0^t g^*(z,s)(\phi(z,s)-\phi^*(z,s))\langle q\rangle(dz,ds).$$

The second integral is the unique predictable increasing process associated to U_t. It is zero if and only if U_t is a martingale, i.e., if and only if ϕ is optimal. The following theorem is now obvious:

THEOREM 4.3.1. *Consider the stochastic control model of* § 4.1 (*i.e. the abstract model of* § 3.1 *with the martingale representation property*) *and assume* N1, C1, C2 *and assumptions* (i), (ii), (iii) *and* (iv) *of this section. Moreover let*

$$\langle q\rangle(A,t)=\int_A\int_0^t \mu(dz,ds)\lambda_s\,ds,$$

for an $\mathcal{F}_t$-adapted finite measure $\mu(A,t)$ on $(Z,\mathcal{F})$ and λ_t an $\mathcal{F}_t$-adapted, non-negative process. Then a control is optimal if and only if it achieves at each time t, the $l\otimes\mathcal{P}$-essential infimum of the Hamiltonian

$$\lambda_t\cdot\int_Z g^*(z,t)\phi(z,t)\mu(dz,t).$$

COROLLARY 4.3.2. *Under the assumptions of Theorem* 4.3.1, *if for $\phi\in\Phi$, for some $\varepsilon>0$,*

$$\lambda_t\cdot\int_Z g^*(z,t)\phi(z,t)\mu(dz,t)\leqq l\otimes\mathcal{P}\text{-}\operatorname*{ess\,inf}_{\tilde\phi\in\Phi}\lambda_t\int_Z g^*(z,t)\cdot\tilde\phi(z,t)\mu(dz,t)+\varepsilon$$

for almost all t, $\mathcal{P}$-a.s., then ϕ is ε-optimal, i.e.,

$$J(\phi)\leqq\inf_{\tilde\phi\in\Phi}J(\tilde\phi)+\varepsilon,\quad \mathcal{P}\text{-a.s.}$$

Remark. For the complete observation case, U_t has for any $\mathcal{P}$ a right continuous modification with left-hand limits [20, VI-T 16]. Hence the Doob-Meyer decomposition can be applied, and we obtain (as in Davis–Varaiya, [6], Boel–Varaiya, [4]), that there exists

$$g_s^\phi \in L^1_{\mathrm{loc}}(q, \mathcal{P}_\phi),$$

$$U_t = J^* + \int_Z \int_0^t g^\phi(z,s)(q(dz,ds) - \phi(z,s)\langle q,q\rangle(dz,ds)) + A_t^\phi,$$

where the predictable increasing process,

$$A_t^\phi = w \cdot \lim_{h\to 0} \int_0^t \frac{E_\phi(U_{s+h} - U_s | \mathcal{F}_s)}{h} ds,$$

is 0 if and only if ϕ is optimal. However only under strong additional assumptions can it be shown that A_t^ϕ is absolutely continuous. Moreover the dual variable in the Hamiltonian then depends on ϕ.

5. The double martingale case

5.1. Control problem with double martingale noise. In this last section we consider the abstract control model under the assumption that $(\Omega, \mathcal{F}, \mathcal{F}_t, \mathcal{P})$ is the canonical space used in the definition of double martingales (see § 2). By Theorem 2.1 $(\Omega, \mathcal{F}, \mathcal{F}_t, \mathcal{P})$ has the martingale representation property with basic random measure martingale $\bar{q}(A,t) = (w_t; q(A,t))$, where w_t is an $\mathbb{R}^k$-valued Brownian motion, $q(A,t)$ a compensated jump process. Hence we can apply the results of § 4, with $\Phi = \{\alpha^1, \cdots, \alpha^k, \psi\}$, $\alpha^i \in L^2_{\mathrm{loc}}(w^i)$, $\psi \in L^1_{\mathrm{loc}}(q)$. The cost functional is then given by

$$J(\phi) = E(Y_1^\phi \cdot \varepsilon(\alpha_1, \cdots, \alpha_n)_1 \cdot \varepsilon(\psi)_1).$$

This model can easily be interpreted as a control problem where a complex interconnected system is regulated for small noise influences (Brownian motion), and also as a scheduling problem for repair and maintenance of subsystems, subject to breakdown at random times (jump process). Such problems have been treated by Sworder [23], and recently with martingale methods by Gertner and Rapaport [14].

5.2. Existence results. Theorems 4.2.2 and 4.2.6 will now be applied to the control model of § 5.1.

THEOREM 5.2.1. *For the control problem of* § 5.1, *assume* Y_t *satisfies* C3 *and* $|Y_1|$ *is bounded, assume complete information* $\mathcal{G}_t = \mathcal{F}_t$, *assume* $\langle q\rangle(Z,t) \leqq \mu_t$ *where* $\mu : \mathbb{R}_+ \to \mathbb{R}_+$ *is an increasing deterministic function. Let* Φ *be closed and convex in* $L^2(w) \times L^2(q)$. *Then* $\mathcal{D}(\Phi)$ *is closed and convex in* $L_s(\Omega, \mathcal{F}, \mathcal{P})$ *and there exists an optimal control.*

THEOREM 5.2.2. *For the control model of* § 5.1, *assume* A^{c}1, A^{c}2, A^d *and* E1 *satisfied and let* $\mathcal{D}(\Phi)$ *be weakly closed in* $L_1(\Omega, \mathcal{F}, \mathcal{P})$. *Then there exists an optimal control.*

The following corollaries give examples of applications of Theorem 5.2.2.

COROLLARY 5.2.3. ([1]). *Let* $Y_t = w_0(t)$, *a one-dimensional Brownian motion. Then* $Y_1 = w_0(1)$ *satisfies* A^{c}2 *and* E1. *Further assume* Φ *to satisfy* A^{c}2 *and* A^d *and let* $\mathcal{D}(\Phi)$ *be weakly closed. Then there exists an optimal control* $\phi^* \in \Phi$ *for the problem of* § 5.1.

COROLLARY 5.2.4. ([18]). *Let* $Y_t = P^y(A,t)$ *be the counting process corresponding to a fundamental jump process with values in* $\mathbb{R}$, *let* Φ *satisfy* A^{c}1 *and* A^d, *assume* E1 *and let* $\mathcal{D}(\Phi)$ *be weakly closed. Then there exists an optimal control.*

We combine the two preceding corollaries by assuming

$$Y_t = w_0(t) + P^y(A,t).$$

References [1], [18] show how this is the transformed cost for some general form of cost functional $J(\phi)=E_\alpha(\int_0^1 c_t^\alpha\,dt)+E_\psi(\int_Z\int_0^1 c_t^\psi\,dt)$.

COROLLARY 5.2.5. *Let Φ fulfill* $\mathrm{A}^{\mathrm{c}}1$, A^{d}, *and let $P^y(A, 1)$ satisfy* E1. *Then there exists an optimal control.*

5.3. The partial observation maximum principle. We consider the model of § 5.1 with assumptions N1, C1, C2, and *assume* that an optimal control $(\alpha^*, \phi^*)\in\Phi$ *exists* (e.g., let the assumptions of Theorem 5.2.1 or 5.2.2 be satisfied). In fact any set of assumptions guaranteeing existence of a $\mathscr{P}^*$ such that $E_{n^*}(Y_t^*\,|\mathscr{G}_t)+U_t(n^*)$ is a $(\mathscr{G}_t, \mathscr{P}^*)$-martingale is sufficient, but we have not been able to find verifiable conditions for this. Finally we assume that all likelihood ratios $\varepsilon(n)_t=\varepsilon(\phi)_t$ are $\mathscr{P}$-locally bounded martingales, i.e., there exists a sequence of $\mathscr{F}_t$-stopping times T_n (independent of ϕ) such that $|\varepsilon(\phi)_{t\wedge T_n}|\leqq K_n$ and $T_n\uparrow 1\,\mathscr{P}$-a.s. This will be satisfied if $\psi\in[-1, M_n]$ on $[0, T_n]$ and $\Lambda(t)$ is continuous (i.e., all jumps are totally inaccessible).

From Theorem 3.3.3 we then have that

$$U_t(n^t)=E(Y_t^n\,|\mathscr{G}_t)+\mathscr{P}_n\text{-}\operatorname*{ess\,inf}_{\tilde n\in N_{n,t}} E_{\tilde n}(Y_1^{\tilde n}-Y_t^{\tilde n}\,|\mathscr{G}_t)$$

is a $(\mathscr{G}_t, \mathscr{P}_n)$-submartingale for all n, and for the optimal control n^* (i.e., (α^*, ψ^*)) we have

$$U_t(n^{*t})\in\mathscr{M}^1(\mathscr{G}_t, \mathscr{P}_{n^*}),$$

i.e.,

$$U_t(n^{*t})=E_{n^*}(Y_1^{n^*}\,|\mathscr{G}_t)=E_{n^*}(Y_t^{n^*}+\tilde W_t|\mathscr{G}_t),$$

where

$$\tilde W_t=E_{n^*}(Y_1^{n^*}-Y_t^{n^*}\,|\mathscr{F}_t).$$

Following Elliott [12] it is easy to verify,

$$\begin{aligned}E_{n^*}(Y_t^n+\tilde W_t|\mathscr{G}_t)&=E_{n^*}(Y_t^n-Y_t^{n^*}\,|\mathscr{G}_t)+E_{n^*}(Y_t^{n^*}+\tilde W_t|\mathscr{G}_t)\\&\leqq E_{n^*}(Y_t^n-Y_t^{n^*}\,|\mathscr{G}_t)+E_{n^*}(Y_t^{n^*}+E_n(Y_{t+h}^n-Y_t^n+\tilde W_{t+h}|\mathscr{F}_t)|\mathscr{G}_t).\end{aligned}$$

Intuitively this corresponds to using a non-optimal control n between t and $t+h$. This proves

THEOREM 5.3.1. *For the model of* § 5.1, *under the additional assumptions of the beginning of this section, if $n^*\in\mathscr{N}$ is optimal then $E_{n^*}(Y_t^{n^*}+\tilde W_t|\mathscr{G}_t)\in\mathscr{M}^1(\mathscr{G}_t, \mathscr{P}_{n^*})$. For all $n\in\mathscr{N}$, n^* being an optimal control,*

$$E_{n^*}(Y_t^n+\tilde W_t|\mathscr{G}_t)\leqq E_{n^*}(E_n(Y_{t+h}^n+\tilde W_{t+h}|\mathscr{F}_t)|\mathscr{G}_t).$$

Since $Y_t^{n^*}+\tilde W_t\in\mathscr{M}^1(\mathscr{F}_t, \mathscr{P}_{n^*})$ we have the representation theorem, that, after stopping at T_n, there exist uniquely defined (up to $\mathscr{P}$-equivalence) processes

$$g^*\in L^1(w), \qquad h^*\in L^1(q),$$

such that

$$\begin{aligned}Y_{t\wedge T_n}^{n^*}+\tilde W_{t\wedge T_n}&=J^*+\int_0^{t\wedge T_n} g_s^*\,(dw_s-\phi_s^*\,ds)\\&\quad+\int_Z\int_0^{t\wedge T_n} h^*(z, s)(p(dz, ds)-(1+\psi^*(z, s))n(dz, s)\Lambda(ds)).\end{aligned}$$

Note that because of the boundedness of $\varepsilon(n)_t$ and $1/\varepsilon(n)_t$ on $[0, T_n]$, $g^* I_{[0,T_n]} \in L^1(w)$ and $h^* I_{[0,T_n]} \in L^1(q)$ for all $\mathscr{P}_n$. Since $T_n \uparrow 1$ for all $\mathscr{P}_n$, the following results hold for $t \in [0, 1)$. (The point $t = 1$ is not very interesting for control purposes anyway.)

Suppose now we can write the cost Y_1^n in the following integral form;

$$\begin{aligned} Y_t^n = &\int_0^t k_s^1(n)(dw_s - \phi_s^n\, ds) \\ &+ \int_Z \int_0^t k_s^2(n)[p(dz, ds) - (1+\psi^n) n(dz, s) \Lambda(ds)] \\ &+ \int_0^t c(s, n_s)\, ds, \end{aligned}$$

where $k_s^1 \in L^1(w)$, $k_s^2 \in L^1(q)$ (in particular, $\mathscr{F}_t$-predictable), exist by the double martingale representation theorem, but the differentiability of $\Lambda(t)$ and $a_t^{Y^n}$ are new assumptions. Then we can write

$$\begin{aligned} Y_t^n + \tilde{W}_t = J^* &+ \int_0^t g_s^* (dw_s - \phi_s^*)\, ds \\ &+ \int_Z \int_0^t h^*(z, s)(p(dz, ds) - (1+\psi^*) n(dz, s) \lambda_s\, ds) \\ &+ \int_0^t k_s^1(n)(dw_s - \phi_s^n\, ds) - \int_s^t k_s^1(n^*)(dw_s - \phi_s^*\, ds) \\ &+ \int_Z \int_0^t k_s^2(n)(p(dz, ds) - (1+\psi^n) n(dz, s)\lambda_s\, ds) \\ &- \int_Z \int_0^t k_s^2(n^*)(p(dz, ds) - (1+\psi^*) n(dz, s) \lambda_s\, ds) \\ &+ \int_0^t (c(s, n_s) - c(s, n_s^*))\, ds. \end{aligned}$$

By the integrability assumptions g_s, g_s^*, $k_s^1 \in L^1(w)$ for $\mathscr{P}_n$ (up to T_n first, and let $T_n \uparrow 1$) and similarly h_s, h_s^*, $k_s^2 \in L^1(q)$ for $\mathscr{P}_n$; then stochastic integrals over $(dw_s - \phi_s^n\, ds)$ and $(p(dz, ds) - (1+\psi^n) n(dz, s)\lambda_s\, ds)$ are $\mathscr{P}_n$-local martingales. Then by Lemma 5.2.1,

$$\begin{aligned} 0 &\leqq E_{n^*}[E_n(Y_{t+h}^n - \tilde{W}_{t+h} - Y_t^n - \tilde{W}_t | \mathscr{F}_t) | \mathscr{G}_t] \\ &= E_{n^*}\left(\left[\int_t^{t+h} (g_s^*(\phi_s^n - \phi_s^*) - k_s^1(n_s^*)(\phi_s^n - \phi_s^*) + (c(s, n_s) - c(s, n_s^*))\, ds \right.\right. \\ &\qquad \left.\left. + \int_t^{t+h} \int_Z (h^*(z, s) - k_s^2(n_s^*))(\psi^n(z, s) - \psi^*(z, s)) n(dz, s) \lambda_s\, ds \right] \middle| \mathscr{G}_t \right), \end{aligned}$$

with equality holding if and only if n is optimal. Then the following theorem is obvious.

THEOREM 5.3.2. *Under all the assumptions made earlier in* § 5.3 *with $n^* \in \mathscr{N}$ an optimal control, there exist dual variables*

$$g^* \in L^1(w), \qquad h^* \in L^1(q),$$

such that for all $n \in \mathcal{N}$ the Hamiltonian is positive:

$$E_{n^*}[(g_t^* - k_t^1(n_t^*))(\phi_t^n - \phi_t^*) + c(t, n_t) - c(t, n_t^*)$$
$$- \lambda_t \int_Z (k^2(z, t, n_t^*) - h^*(z, t))(\psi^n(z, t) - \psi^*(z, t))n(dz, t) | \mathcal{G}_t] \geqq 0,$$

for almost all t. For any other optimal control n the minimum (zero) is achieved.

Remarks.

(i) Recently Elliott [28] has proved the sufficient part of the above theorem (= 0 implies n optimal), for a model with Brownian motion noise. His proof apparently carries over to the double martingale model considered here.

(ii) To compare Theorem 5.3.2 with the maximum principle in Elliott, [12, Thms. 9.3, 9.8], one should take the cost structure

$$Y_t^n = Y_t = \int_0^t k_s^1 \, dw_s + \int_0^t k^2(z, s) q(dz, ds).$$

Then the Hamiltonian takes the form

$$E_{n^*}\left[g_t^*(\phi_t^n - \phi_t^*) + \lambda_t \int_Z h^*(z, t)(\psi^n(z, t) - \psi^*(z, t))n(dz, t) \middle| \mathcal{G}_t \right].$$

Our result is weaker, however, since the exceptional sets, where the inequality is not satisfied, can depend on t because we could not prove the generalization of the differentiability results in [12, § 7].

5.4. Final remarks and applications. The results on double martingales in the preceding sections suggest some game theoretical extensions. Let us once again look at the cost criterion in the double martingale case

$$J(\phi) = J(\alpha, \psi) = E(Y_1 \varepsilon(\alpha)_1 \varepsilon(\psi)_1),$$

that was to be minimized in the control problem above. Now assume that there are two players, P_1 and P_2, governing respectively the continuous and discontinuous parts of the dynamics; i.e., with respect to the dynamics

$$dx_t = \alpha_t \, dw_t + \int_Z \psi(z, t) q(dz, dt),$$

we have to solve the problem,

$$\min_{\alpha} \max_{\psi} J(\alpha, \psi).$$

Problems of this kind as well as applications to economics will be described in a forthcoming paper [19].

Acknowledgment. The authors would like to thank J. Jacod and M. Yor for permission to use their elegant proof of Theorem 2.1.

REFERENCES

[1] V. E. BENEŠ, *Existence of optimal stochastic control laws*, Siam J. Control, 9 (1971), pp. 446–472.

[2] R. BOEL, P. VARAIYA AND E. WONG, *Martingales on jump processes I., representation results*, Siam J. Control, 13 (1975), pp. 999–1021.

[3] ———, *Martingales on jump processes II., applications*, Siam J. Control, 13 (1975), pp. 1022–1061.

[4] R. Boel and P. Varaiya, *Optimal control of jump processes*, this Journal, 15 (1977), pp. 92–119.

[5] P. Brémaud and M. Pietri, *On the abstract dynamic programming approach to continuous time stochastic control*, preprint (1976).

[6] M. H. A. Davis and P. Varaiya, *Dynamic programming conditions for partially observable stochastic systems*, Siam J. Control, 11 (1973), pp. 226–261.

[7] M. H. A. Davis, *The representation of martingales of jump processes*, this Journal, 14 (1976), pp. 623–237.

[8] M. H. A. Davis and R. J. Elliott, *Optimal control of a jump process*, Z. Wahrscheinlichkeitstheorie und verw. Gebeite, 40 (1977), pp. 183–202.

[9] J. L. Doob, *Stochastic Processes*, John Wiley, New York, 1967.

[10] T. Duncan and P. Varaiya, *On the solutions of a stochastic control system*, Siam J. Control, 9 (1971), pp. 354–371.

[11] R. J. Elliott, *Martingales of a Jump Process and Absolutely Continuous Changes of Measure*, Symposium on Stochastic Systems, University of Kentucky, 1975.

[12] ———, *The optimal control of a stochastic system*, this Journal, 16 (1977), pp. 756–778.

[13] ———, *Double martingales*, Z. Wahrscheinlichkeitstheorie und verw. Gebeite, 34 (1976), pp. 17–28.

[14] I. Gertner and D. Rapaport, *Stochastic control of system with unobserved jump parameter process*, Information Sciences, 13 (1977), pp. 269–282.

[15] I. V. Girsanov, *On transforming a certain class of stochastic processes by absolutely continuous substitution of measures*, Theor. Probability Appl., 5 (1960), pp. 285–301.

[16] J. Jacod and M. Yor, *Etude des solutions extrémales et représentation intégrale des solutions pour certains problemès de martingales*, Z. Wahrscheinlichkeitstheorie und verw. Gebeite, 38 (1977), pp. 83–125.

[17] J. Jacod and J. Mémin, *Charactéristiques locales et conditions de continuite absolue pour les semimartingales*, Université de Rennes, preprint (1977).

[18] M. Kohlmann, *On control of jump processes: a martingale approach*, preprint 184 SFB 72 Universität Bonn.

[19] ———, *A game with Wiener noise and jump process disturbances*, preprint 250 SFB 72 Universität Bonn.

[20] P. A. Meyer, *Probabilités et Potentiel*, Hermann, Paris, 1966.

[21] ———, *Un cours sur les intégrales stochastiques*, Séminaire de Probabilité X, Université de Strasbourg, Lecture Notes in Mathematics, Vol. 511, Springer Verlag, Berlin, 1976.

[22] R. Rishel, *Necessary and sufficient dynamic programming conditions for continuous time stochastic optimal control*, Siam J. Control, 8 (1970), pp. 559–571.

[23] D. Sworder, *Feedback control of a class of linear systems with jump parameters*, IEEE Trans. Automatic Control, 14 (1969), pp. 9–14.

[24] J. H. van Schuppen and E. Wong, *Transformation of local martingales under a change of law*, Ann. Probability, 2 (1974), pp. 879–888.

[25] Ch. Striebel, *Martingale conditions for the optimal control of continuous time stochastic systems*, Department of Mathematics, University of Minnesota, preprint.

[26] Ch. Yoeurp, *Décompositions des martingales locales et formules exponentielles*, Séminaire de Probabilité X, Université de Strasbourg, Lecture Notes in Mathematics, Vol. 511, Springer Verlag, Berlin, 1976.

[27] R. Boel and M. Kohlmann, *Stochastic Optimal Control over Double Martingales*, Proceedings of the International Conference on the Analysis and Optimization of Stochastic Systems, Oxford, 1978.

[28] R. J. Elliott, *A Sufficient Condition for the Optimal Control of a Partially Observed System*, Proceedings of the International Conference on the Analysis and Optimization of Stochastic Systems, Oxford, 1978.

[29] M. H. A. Davis, *Martingales of Wiener and Poisson processes*, J. London Math. Soc., 13 (1976), pp. 336–338.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 5, September 1980

0363-0129/80/1805-0005 $01.00/0

CONTROLLABILITY PROPERTIES OF PSEUDO-PARABOLIC BOUNDARY CONTROL PROBLEMS*

L. W. WHITE†

Abstract. We study the boundary control of a pseudo-parabolic equation and compare the results to those of parabolic equations. We find that the pseudo-parabolic equation on a finite interval is controllable by means of boundary controls, but on a semi-infinite interval it is not controllable. These results are a consequence of uniqueness and support properties of solutions of the pseudo-parabolic equation.

1. Introduction and preliminaries. In this note we present controllability results for boundary control problems governed by pseudo-parabolic equations. We show that on a bounded domain a pseudo-parabolic equation is controllable by means of boundary controls alone. In addition, we give an example to show that on an unbounded domain it is not controllable by means of boundary controls. This last result varies from that observed for parabolic problems [4]. This difference is due to the different uniqueness properties of pseudo-parabolic and parabolic equations.

For simplicity we consider the following situation and mention generalizations at the end of each section. We study a problem of the form

$$y_t(x,t)-y_{xtx}(x,t)-y_{xx}(x,t)=0, \quad \text{in } (0,1)\times(0,+\infty), \tag{1}$$

$$y(x,0)=0 \qquad \text{in } (0,1), \tag{2}$$

with boundary conditions

$$y(0,t)=u_0(t) \quad \text{and} \quad y(1,t)=u_1(t), \quad \text{in } (0,+\infty). \tag{3}$$

Equation (1) is to be satisfied for each $t\in(0,+\infty)$, and it suffices to take u_0 and u_1 to be continuous functions with compact support on $(0,+\infty)$, i.e., u_0 and u_1 belong to $C_0(0,\infty)$. We denote the dependence of y upon the controls by $y(x,t;\underset{\sim}{u})$ where $\underset{\sim}{u}=(u_0,u_1)$. We focus our attention on the trace $y(\cdot,T;\underset{\sim}{u})$ for some fixed finite T, $0<T<+\infty$, and define the set

$$Y(T)=\{y(\cdot,T;\underset{\sim}{u}): \underset{\sim}{u}\in C_0(0,\infty)\times C_0(0,\infty)\}. \tag{4}$$

When we say that (1)–(3) is controllable, we mean that $Y(T)$ is a dense subspace of $L^2(0,1)$, cf. [4, 6].

Equation (1) is of pseudo-parabolic type. These equations arise in areas such as fluid flow [13], heat transfer [2], and the diffusion of radiation [7]. Roughly speaking, pseudo-parabolic equations account for higher order correction in the model than do parabolic equations. The study of pseudo-parabolic equations and their relation to parabolic equations was begun in [3], [12], [14]. We refer to [1] for an extensive bibliography concerning equations of this type.

Control problems governed by pseudo-parabolic equations were first studied in [15], [16]. In this work, however, attention was restricted to control distributed over the entire space time cylinder. Boundary control for pseudo-parabolic equations is treated here for the first time. The contribution of this study is that it establishes controllability

* Received by the editors July 10, 1979, and in revised form November 25, 1979. This research was supported in part by National Science Foundation Grant MCS-7902037.

† Department of Mathematics, University of Oklahoma, Norman, Oklahoma 73069.

properties for an important class of equations and that it compares these results to those better known results for parabolic equations.

In § 2 we prove the controllability of (1)–(3). In § 3 we give an example with $(0, +\infty)$ replacing $(0, 1)$ and with (3) replaced by

$$y(0, t) = y_0(t). \tag{3$'$}$$

In this case (1), (2), and (3)′ are shown not to be controllable.

2. The controllability result. We prove that $Y(T)$ is dense in $L^2(0, 1)$ where u_0 and u_1 are allowed to vary in $C_0(0, \infty)$. Although this result is the same as in the parabolic case, cf. [4], [6], it is for different reasons. The difference arises from the uniqueness and support properties of the solutions of these equations [3], [10], and enables us in § 3 to construct an example of a pseudo-parabolic problem that is not controllable.

THEOREM 1. *Problem* (1)–(3) *is controllable.*

Proof. To show that $Y(T)$ is dense in $L^2(G)$, we let $\xi(\cdot)$ be an element in $L^2(G)$ with the property

$$(\xi(\cdot), y(\cdot, T; \underset{\sim}{u}))_0 = 0, \tag{5}$$

for all $\underset{\sim}{u} \in C_0(0, \infty) \times C_0(0, \infty)$. We show that equation (5) implies $\xi = 0$ in $L^2(G)$, cf. [6]. To this end we introduce the following adjoint problem,

$$-q_t(x, t) + q_{xtx}(x, t) - q_{xx}(x, t) = 0, \quad \text{in } (0, 1) \times (-\infty, T], \tag{6}$$

$$q(x, T) = \eta(x), \quad \text{in } (0, 1), \tag{7}$$

$$q(0, t) = q(1, t) = 0, \quad \text{in } (-\infty, T], \tag{8}$$

where η is the solution to the problem

$$\begin{gathered} \eta(x) - \eta_{xx}(x) = \xi(x), \quad \text{in } (0, 1), \\ \eta(0) = \eta(1) = 0. \end{gathered} \tag{9}$$

Consider the integral

$$\int_0^T \int_0^1 q(x, t)[y_t(x, t) - y_{xtx}(x, t) - y_{xx}(x, t)]dxdt = 0, \tag{10}$$

and integrate to obtain

$$\begin{aligned} &[q(x, T)y_x(x, T)]_0^1 - [q_x(x, T)y(x, T)]_0^1 \\ &\quad + \int_0^1 [q(x, T) - q_{xx}(x, T)]y(x, T)\, dx - \int_0^1 q(x, 0)[y(x, 0) - y_{xx}(x, 0)]\, dx \\ &\quad - \int_0^T \{[y_x(x, t)(q(x, t) - q_t(x, t))]_0^1 - [y(x, t)(q_x(x, t) - q_{xt}(x, t))]_0^1\}\, dt \\ &\qquad\qquad + \int_0^T \int_0^1 [-q_t(x, t) + q_{xtx}(x, t) - q_{xx}(x, t)]y(x, t)\, dx\, dt = 0. \end{aligned} \tag{11}$$

By using equations (1)–(3), (5), and (6)–(8) in equation (11), we have

$$
\begin{aligned}
&-\{q_x(1,T)u_1(T)-q_x(0,T)u_0(T)\}\\
&\quad+\int_0^T u_1(t)(q_x(1,t)-q_{xt}(1,t))\,dt\\
&\quad+\int_0^T u_o(t)(q_x(0,t)-q_{xt}(0,t))\,dt=0.
\end{aligned}\tag{12}
$$

Since u_0 and u_1 are arbitrary continuous functions on $[0, T]$, we deduce that

$$
\begin{aligned}
q_x(0,t)-q_{xt}(0,t)=0 \quad \text{a.e. in } [0,T],\\
q_x(0,T)=0,
\end{aligned}\tag{13}
$$

and

$$
\begin{aligned}
q_x(1,t)-q_{xt}(1,t)=0 \quad \text{a.e. in } [0,T],\\
q_x(1,T)=0.
\end{aligned}\tag{14}
$$

In fact, equality holds for all t in $[0, T]$ due to the smoothness properties of solutions of pseudo-parabolic equations [11], [12]. Equations (13) and (14) now imply that

$$
q_x(0,t)=q_x(1,t)=0, \qquad \text{for all } t\in[0,T].\tag{15}
$$

At this point for the parabolic case, we are done by the unique continuation property for the solution of parabolic equations [5], [8]. In this case, however, there is no such property [10]. This is due to the fact that the pseudo-parabolic equation has two families of characteristic curves, x = constant and t = constant, and here the auxiliary data is given along the curve x = constant. Indeed, it is possible for solutions of pseudo-parabolic equations to have support contained in $x > x_0$, [10]. Uniqueness for our problem, however, results from the property that the solution of a pseudo-parabolic equation cannot have compact support in the space variable [10].

We now finish the proof, cf. [10]. For ease we change variables by setting $\tau = T - t$ to obtain

$$q_\tau(x,\tau)-q_{xx\tau}(x,\tau)-q_{xx}(x,\tau)=0, \quad \text{in } (0,1)\times[0,\infty); \tag*{(6)$'$}$$

$$q(x,0)=\eta(x), \quad \text{in } (0,1); \tag*{(7)$'$}$$

$$q(0,\tau)=q(1,\tau)=0, \quad \text{in } [0,\infty); \tag*{(8)$'$}$$

$$q_x(0,\tau)=0, \quad \text{in } [0,T]. \tag*{(15)$'$}$$

Now the operator $-d^2/dx^2$ with domain $H_0^1(0,1)\cap H^2(0,1)$ has eigenvalues $\{n^2\pi^2\}_{n=1}^\infty$ with eigenfunctions $\{\sin n\pi x\}_{n=1}^\infty$. The solution of (6)′–(8)′ is given by

$$
q(x,\tau)=\sum_{n=1}^\infty q_n e^{-\mu_n\tau}\sin \nu\pi x,\tag{16}
$$

where $\mu_n = n^2\pi^2/(1+n^2\pi^2)$ and $\eta(x)=\sum_{n=1}^\infty q_n \sin n\pi x$. Furthermore, since $\eta \in H_0^1(0,1)\cap H^2(0,1)$, it is clear that

$$
q_x(x,\tau)=\sum_{n=1}^\infty n\pi q_n e^{-\mu_n\tau}\cos n\pi x
$$

is in $L^2(0, 1)$ for each $\tau \in [0, \infty)$ and is analytic for $\tau \geqq 0$ and $x \in [0, 1]$. Thus, we see that, in particular,

$$q_x(0, \tau) = \sum_{n=1}^{\infty} n\pi q_n e^{-\mu_n \tau}$$

is an analytic function of τ for $\tau \geqq 0$. Condition (15)′ then gives that

$$\sum_{n=1}^{\infty} n\pi q_n e^{-\mu_n \tau} = 0, \tag{17}$$

in fact for all τ in $[0, +\infty)$.

By taking the Laplace transform of (17), we have

$$\sum_{n=1}^{\infty} n\pi q_n (s + \mu_n)^{-1} = 0,$$

for $s > 0$. Defining the function

$$f(z) = \sum_{n=1}^{\infty} n\pi q_n (z + \mu_n)^{-1},$$

we observe that f is a meromorphic function with poles at $z = -\mu_n$ and with the property that $f(z) = 0$ for z real and positive. But this implies that $f(z) \equiv 0$, with zero residues at the poles. Thus, $n\pi q_n = 0$, for $n \geqq 1$ so that $q_n = 0$. Accordingly, $q(x, t) = 0$ in $(0, 1) \times (-\infty, T)$, and this implies that $\xi = 0$. Therefore, $Y(T)$ is dense in $L^2(0, 1)$.

Remark 2. In the proof above, the condition $dq/dx = 0$ is used only at one boundary. Thus, if in (3) we control only one boundary, say $y(0, t) = u_0(t)$, and fix $y(1, t) = 0$, then the problem (1)–(3) remains controllable.

Remark 3. If (3) is replaced by Neumann controls

$$y_x(0, t) = u_0(t) \quad \text{and} \quad y_x(1, t) = u_1(t) \quad \text{in } (0, +\infty), \tag{3$''$}$$

a similar proof establishes controllability. Obviously combinations are possible.

Remark 4. A similar proof holds with $-d^2/dx^2$ replaced by operators of the form $-(d/dx)(m(x)\, d/dx) + \bar{m}(x)$, with $m(x) \geqq c_m > 0$. Generalization to higher dimensions R^n, by using generalized Fourier series, in which $(0, 1)$ is replaced by an open bounded domain G with a sufficiently smooth boundary and $-d^2/dx^2$ is replaced by a uniformly strongly elliptic operator is also possible.

3. An example. In this section we give an example to show that the pseudo-parabolic equation is not controllable if the interval $(0, 1)$ used in (1)–(3) is replaced by the semi-infinite interval $(0, +\infty)$. Here the condition (3) is actually replaced by

$$y(0, t) = u_0(t), \quad \text{and} \quad y(x, t) \to 0 \quad \text{as } x \to \infty, \tag{3$'$}$$

for each $t \in (0, +\infty)$, see [9]. We again consider $Y(T)$ in this case and seek to determine whether it is dense in $L^2(0, +\infty)$.

Here we introduce the problem

$$\begin{aligned} &-q_t(x, t) + q_{xtx}(x, t) - q_{xx}(x, t) = 0, \quad \text{in } (0, +\infty) \times (-\infty, T]; \\ &q(x, T) = \eta(x), \quad \text{in } (0, +\infty); \\ &q(0, t) = q_x(0, t) = 0, \quad \text{in } (-\infty, T]. \end{aligned} \tag{18}$$

The function η is given by

$$\eta(x) = \frac{1}{2} e^{-x} \int_{-\infty}^{x} h(\xi)\, e^{\xi}\, d\xi - \frac{1}{2} e^{x} \int_{-\infty}^{x} h(\xi)\, e^{-\xi}\, d\xi, \tag{19}$$

where h is chosen to have the properties

$$\begin{aligned}&h \in C_0^\infty(-\infty, \infty),\\ &h(x)=0, \quad \text{for } x<1 \text{ and } x>2,\\ &\int_1^2 h(x)e^{-x}\,dx=0.\end{aligned}\tag{20}$$

Note that η is the solution of

$$-\eta_{xx}(x)+\eta(x)=h(x) \quad \text{in } (-\infty, \infty),$$

that satisfies $\eta(x)=0$, for $x \leqq 1$ and $x \geqq 2$. Furthermore, we have that $\eta_x(x)=0$, for $x \leqq 1$ and $x \geqq 2$. As was done previously we multiply (1) by q and integrate. Using (1), (2), (3)′, and (18), we obtain the following equation.

$$\begin{aligned}0=&\int_0^\infty [q(x, T)-q_{xx}(x, T)]y(x, T)\,dx\\ &+\int_0^T\int_0^\infty [-q_t(x, t)+q_{xtx}(x, t)-q_{xx}(x, t)]y(x, t)\,dx\,dt\\ &+\lim_{x\to\infty}\Big\{q_x(x, T)y(x, T)-q(x, T)y_x(x, T)\\ &\qquad-\int_0^T [q_t(x, t)-q(x, t)]y_x(x, t)\,dt\\ &\qquad+\int_0^T [q_{xt}(x, t)-q_x(x, t)]y(x, t)\,dt\Big\}.\end{aligned}\tag{21}$$

Since h and η are infinitely differentiable with compact support, the solution of the pseudo-parabolic equation and its derivatives are rapidly decreasing functions [9]. Thus, the limit is zero, and we have

$$\int_0^\infty h(x)y(x, T; u)\,dx=0,\tag{22}$$

for all u, in fact, in $L^2(0, T)$. Since h is nonzero, $Y(T)$ is not dense in $L^2(0, \infty)$.

Remark 5. The motivation for the choice of h comes from the representation of the solution of a pseudo-parabolic equation by means of an integral that involves a Riemann function [3]. This representation implies that if support of $h(x)$ is contained in $x>x_0$ then support of $q(x, t)$ is contained in $x>x_0$ [10]. The growth of $q(x, t)$ comes from the fact that if $\eta(x)$ is rapidly decreasing then q is rapidly decreasing [9].

REFERENCES

[1] R. W. Carroll and R. E. Showalter, *Singular and Degenerate Cauchy Problems*, Mathematics in Science and Engineering, Vol. 127, Academic Press, New York, 1976.

[2] P. Chen and M. Gurtin, *On a theory of heat conduction involving two temperatures*, Z. Angew. Math. Phys., 19 (1968), pp. 614–627.

[3] D. Colton, *Pseudoparabolic equations in one space variable*, J. Differential Equations, 12 (1972), pp. 559–565.

[4] H. O. Fattorini, *Boundary control systems*, this Journal, 6 (1968), pp. 349–385.

[5] G. Hellwig, *Partial Differential Equations*, Blaisdell Publishing Co., New York, 1964.

[6] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, translated by S. K. Mitter, Springer Verlag, New York, 1971.

[7] E. A. MILNE, *The diffusion of imprisoned radiation through a gas*, J. London Math. Soc., 1 (1926), pp. 40–51.

[8] S. MIZOHATA, *Unicite du prolongement des solutions pour quelques operateurs differentials paraboliques*, Memoirs of the College of Science, University of Kyoto, Series A, Vol. XXXI, 3 (1958), pp. 219–239.

[9] V. R. GOPALA RAO AND T. W. TING, *Solutions of pseudo heat equations in the whole space*, Arch. Rational Mech. Anal., 49 (1972), pp. 57–78.

[10] W. RUNDELL AND M. STECHER, *Remarks concerning the supports of solutions of pseudoparabolic equations*, Proc. Amer. Math. Soc., 63 (1977), pp. 77–81.

[11] R. E. SHOWALTER, *The Sobolev equation, II*, Applied Analysis, 5 (1975), pp. 81–89.

[12] R. E. SHOWALTER AND T. W. TING, *Pseudoparabolic partial differential equations*, SIAM J. Math. Anal., 1 (1970), pp. 1–26.

[13] T. W. TING, *Certain nonsteady flows of second order fluids*, Arch. Rational Mech. Anal., 14 (1963), pp. 1–26.

[14] ———, *Parabolic and pseudoparabolic partial differential equations*, J. Math. Soc. Japan, 21 (1969), pp. 440–453.

[15] L. W. WHITE, *Control problems governed by a pseudoparabolic partial differential equation*, Trans. Amer. Math. Soc., to appear.

[16] ———, *Control of a pseudo-parabolic initial value problem to a target function*, this Journal, 17 (1979), pp. 587–595.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 5, September 1980

0363-0129/80/1805-0006 $01.00/0

STABILITY ANALYSIS OF CONTINUOUS-TIME ADAPTIVE CONTROL SYSTEMS*

BO EGARDT†

Abstract. The stability properties of a fairly general continuous-time adaptive control scheme are analyzed. Sufficient conditions for L^∞-stability in the presence of disturbances are given. The stability results are used to prove convergence of the process outputs in the disturbance-free case, without requiring any a priori stability assumptions.

1. Introduction. Methods for on-line tuning of controllers when the plant is unknown have been frequently discussed during the recent years. Model reference adaptive systems (MRAS) for continuous-time control have focused much attention. See, e.g., the surveys by Landau [13] and Narendra and Valavani [18]. The close connections between this approach and the self-tuning regulator (STR) philosophy have been demonstrated in, e.g., Ljung and Landau [15] and Egardt [5], [7].

The MRAS have been analyzed extensively from different points of view, e.g., convergence, stability and noise rejection properties. In particular, the stability problem has been discussed by many authors since Monopoli's important paper appeared in 1974. See, e.g., Feuer and Morse [9], Narendra and Valavani [20], Feuer et al. [11], and Egardt [6], [8]. Not only is the stability as such important, but a stability condition has also been imposed in most studies of convergence. It seems that the only rigorous convergence proofs without the stability requirement are the ones by Feuer and Morse [9] and Morse [17], treating the disturbance-free case, and Egardt [6] on which this paper is based.

The purpose of the present paper is to give some stability results for a class of continuous-time adaptive schemes in the presence of disturbances. The algorithm considered is based on the STR philosophy, i.e., a separation between identification and control. Slightly modified versions of several MRAS can be treated as special cases of the general scheme. See Egardt [7]. The main result (Theorem 1) states that the closed-loop signals remain bounded under some reasonable conditions. The most important one—boundedness of parameter estimates—can be omitted if the algorithm is slightly modified. When no disturbances affect the plant, the stability result can be used to prove convergence of the output error to zero. This result thus holds without a priori requiring the closed loop to be stable.

The adaptive controller considered is based on a design scheme for known plants. This scheme is briefly described in § 2, and § 3 then gives the formulation of the adaptive controller. The stability results are presented in § 4 and § 5 gives some conclusions.

2. Design scheme for known plants. The adaptive scheme defined in the next section is based on a design method for known systems. A brief description of this method is given below. For more details, one is referred to Åström [1] and Egardt [5], [7].

Problem formulation. The plant is assumed to satisfy the differential equation

$$(1)\qquad y(t)=\frac{b_0B(p)}{A(p)}u(t)=\frac{b_0(p^m+b_1p^{m-1}+\cdots+b_m)}{p^n+a_1p^{n-1}+\cdots+a_n}u(t),$$

where p denotes the differential operator.

* Received by the editors May 3, 1979, and in revised form December 27, 1979.

† Department of Automatic Control, Lund Institute of Technology, Box 725, S-220 07 Lund 7, Sweden.

Remark. It is assumed that there is no disturbance. This is a temporary assumption which will be removed in the stability analysis.

The objective of the controller is to make the closed-loop transfer operator equal to a reference model transfer operator, given by

$$y^M(t)=\frac{B^M(p)}{A^M(p)}u^M(t)=\frac{b_0^Mp^m+\cdots+b_m^M}{p^n+a_1^Mp^{n-1}+\cdots+a_n^M}u^M(t). \tag{2}$$

Here $y^M(t)$ is the desired output of the closed loop system and $u^M(t)$ is the command input. It is assumed that the polynomial $A^M(p)$ is asymptotically stable. Note that the pole excess (i.e., the difference between number of poles and number of zeros) of the reference model is greater than or equal to the pole excess of the plant. This assumption is made to avoid differentiators in the control law.

Design procedure. The design procedure consists of the following steps:

(i) Choose the asymptotically stable monic polynomial $T(p)$,

$$T(p)=p^{n_T}+t_1p^{n_T-1}+\cdots+t_{n_T}, \qquad n_T\geqq n-m-1.$$

(ii) Solve the equation

$$T(p)A^M(p)=A(p)R(p)+S(p) \tag{3}$$

for the unique solutions $R(p)$ and $S(p)$, defined by

$$S(p)=s_0p^{n-1}+\cdots+s_{n-1}$$
$$R(p)=p^{n_T}+r_1p^{n_T-1}+\cdots+r_{n_T}.$$

(iii) Use the control law

$$b_0B(p)R(p)u(t)=T(p)B^M(p)u^M(t)-S(p)y(t). \tag{4}$$

Remark 1. Note that the B-polynomial is canceled, restricting the design method to minimum phase systems.

Remark 2. The polynomial $T(p)$ can be interpreted as the characteristic polynomial of an observer.

3. An adaptive controller. The design procedure given in the preceding section will serve as the starting point when defining an adaptive algorithm in this section. The controller considered is fairly general and several proposed MRAS can be viewed as special cases (see Egardt [7]).

First, introduce a disturbance in the problem formulation. Thus, let the plant be governed by

$$y(t)=\frac{b_0B(p)}{A(p)}u(t)+v(t), \tag{5}$$

where $v(t)$ is a disturbance, which cannot be measured. The polynomials are defined as in (1).

The following assumptions are made:

(A1) The number of poles n and zeros m are known and $m\leqq n-1$.

(A2) The parameter b_0 is nonzero and its sign is known. Without loss of generality b_0 is assumed positive.

(A3) The plant is minimum phase, i.e., the zeros of the polynomial $B(p)$ lie in the open left half plane.

Remark. Notice that it is sufficient to know the pole excess and an upper bound on the number of poles to write the differential equation in the form of (5) with known n

and m. The minimum phase assumption is natural since the underlying design scheme works for minimum phase systems only.

The desired closed-loop response is still given by (2) and the discrepancy between true and desired response is given by

$$e(t)=y(t)-y^M(t).$$

To obtain some flexibility, a filtered version of the output error $e(t)$ will be considered. Thus, define the filtered error

$$e_f(t)=\frac{Q(p)}{P(p)}e(t)=\frac{Q(p)}{P_1(p)P_2(p)}[y(t)-y^M(t)], \tag{6}$$

where

$$\begin{aligned}Q(p)&=p^{n+n_T-1}+q_1p^{n+n_T-2}+\cdots+q_{n+n_T-1},\\ P_1(p)&=p^{n-m-1}+p_{11}p^{n-m-2}+\cdots+p_{1(n-m-1)},\\ P_2(p)&=p^{m+n_T}+p_{21}p^{m+n_T-1}+\cdots+p_{2(m+n_T)},\end{aligned}$$

are all asymptotically stable polynomials.

The algorithm to be considered is an *implicit* algorithm, Åström et al. [3]. This means that the controller parameters are estimated instead of the parameters of the model (5). It is therefore necessary to have a model of the plant, which contains the controller parameters. Use the equations (2), (3) and (5) to write $e_f(t)$ as

$$\begin{aligned}e_f(t)&=\frac{Q}{TA^M}\left[\frac{b_0BR}{P}u(t)+\frac{S}{P}y(t)-\frac{TB^M}{P}u^M(t)\right]+\frac{QAR}{TA^MP}v(t)\\ &=\frac{Q}{TA^M}\left[b_0\frac{u(t)}{P_1}+b_0(BR-P_2)\frac{u(t)}{P}+S\frac{y(t)}{P}-\frac{TB^M}{P}u^M(t)\right]\\ &\quad+\frac{QAR}{TA^MP}v(t).\end{aligned} \tag{7}$$

Let θ be a vector containing the unknown parameters of the polynomials $BR-P_2$ (degree $m+n_T-1$) and S/b_0 (degree $n-1$) and the constant $1/b_0$ as the last element. Note that the vector θ contains the parameters of the controller, described in § 2. Furthermore, define the vector

$$\varphi^T(t)=\left[\frac{p^{m+n_T-1}}{P}u(t),\cdots,\frac{1}{P}u(t),\frac{p^{n-1}}{P}y(t),\cdots,\frac{1}{P}y(t),-\frac{TB^M}{P}u^M(t)\right]. \tag{8}$$

It is then possible to rewrite the expression (7) for the filtered error $e_f(t)$ as

$$\begin{aligned}e_f(t)&=\frac{Q}{TA^M}\left[b_0\frac{u(t)}{P_1}+b_0\theta^T\varphi(t)\right]+\frac{QAR}{TA^MP}v(t)\\ &=b_0\frac{\bar{u}(t)}{P_1}+b_0\theta^T\bar{\varphi}(t)+\frac{AR}{P}\bar{v}(t).\end{aligned} \tag{9}$$

Here $\bar{x}$ denotes the signal obtained by filtering x with Q/TA^M.

The model (9) provides the starting point for a class of implicit adaptive algorithms. Note that b_0 and θ are essentially the parameters of the controller described in the preceding section. Using the idea of the self tuning controllers, the intention is to estimate the parameters b_0 and θ and then to use a control law, derived from the known parameter case (See § 2). Different choices of estimation algorithm and control law are of course possible. It is, however, necessary to be more specific for the analysis. The following adaptive controller, using a stochastic approximation estimation scheme, will be considered in this paper:

–estimation:

(10a) $$\dot{\hat{b}}_0(t)=\left(\frac{\bar{u}(t)}{P_1}+\hat{\theta}^T(t)\bar{\varphi}(t)\right)\frac{\varepsilon(t)}{r(t)},$$

(10b) $$\dot{\hat{\theta}}(t)=\bar{\varphi}(t)\frac{\varepsilon(t)}{r(t)},$$

(10c) $$r(t)=\alpha+|\bar{\varphi}(t)|^2, \qquad \alpha>0,$$

(10d) $$\varepsilon(t)=e_f(t)-\hat{e}_f(t),$$

(10e) $$\hat{e}_f(t)=\hat{b}_0(t)\left(\frac{\bar{u}(t)}{P_1}+\hat{\theta}^T(t)\bar{\varphi}(t)\right).$$

–control:

(10f) $$\frac{\bar{u}(t)}{P_1(p)}=-\left(\frac{P_1(0)}{P_1(p)}\hat{\theta}^T(t)\right)\bar{\varphi}(t).$$

Remark 1. The estimation scheme is analogous to stochastic approximation algorithms in discrete time. The denominator $r(t)$ can also be given by

$$\dot{r}(t)=-\lambda r(t)+|\bar{\varphi}(t)|^2+\alpha, \qquad \lambda>0,$$

but this generalization will not be considered here. Note that when the pole excess is equal to one, $P_1(p)$ is a constant and (10a, f) imply that $\dot{\hat{b}}_0(t)=0$. In this case it is thus not necessary to estimate b_0. It also follows from (10d, e) that $\varepsilon(t)=e_f(t)$.

Remark 2. The control law (10f) is *not* the same as the commonly used one in MRAS, which is given by

$$u(t)=-\hat{\theta}^T(t)P_1(p)\varphi(t).$$

The control laws are identical if $\hat{\theta}(t)$ is constant. Both should be considered as differentiator-free approximations of the control law

$$\frac{\bar{u}(t)}{P_1}=-\hat{\theta}^T(t)\bar{\varphi}(t),$$

which sets the estimate of the filtered error equal to zero and contains differentiators if P_1 is not a constant. Note that the control law (10f) can be written in terms of u as

$$u(t)=-\frac{TA^MP_1}{Q}\left[\left(\frac{P_1(0)}{P_1(p)}\hat{\theta}^T(t)\right)\bar{\varphi}(t)\right].$$

Further comments on the choice of control law are given in Egardt [6], [8].

Remark 3. The polynomials Q and P give the flexibility to cover several earlier proposed MRAS as special cases (see Egardt [7]). It should also be noted that the usually required positive real condition is not imposed here. The reason is that the filtering by the transfer function Q/TA^M above eliminates this condition (see Egardt [7] for details). It should, however, be noted that this filtering makes the behavior in tracking and regulation identical. This is discussed in Landau [14].

4. Stability analysis. Different routes can be taken when investigating stability properties. The most straightforward one is probably to analyze local properties only. This is done in, e.g., Feuer and Morse [10].

Global stability is much more difficult to analyze. The standard method—to find a Lyapunov function—has been used by, e.g., Feuer and Morse [9] to design a globally stable MRAS. The scheme is, however, complicated. Morse [17] gives a stability proof for the deterministic case with a somewhat more involved analysis.

The separation between identification and control suggests an alternative approach to the stability analysis. If the parameter estimates cause the closed-loop system to be unstable, the input and output signals increase. The estimates will then improve and a stabilizing feedback is again achieved.

There are some shortcomings of the argument given above. It takes some time for the estimates to become good. The argument is thus not valid if the signals increase very rapidly or if the parameter adjustment is very slow. The latter situation is avoided by assuming that the gain in the estimation algorithm is nondecreasing, i.e., $\lambda \neq 0$. The possibility that the signals might increase arbitrarily fast is excluded by ensuring bounded parameter estimates. It is shown in Egardt [6], [8] that it is really necessary to introduce this or some other, similar assumption.

With these extra assumptions, it is possible to make a rigorous proof of stability using the ideas above. Here uniform boundedness of the closed-loop signals will be considered. It is then natural to assume that the command signal u^M and the disturbance v are bounded. It is convenient to make the following definition.

Definition. The closed-loop system is said to be *L^∞-stable* if uniformly bounded command (u^M) and disturbance (v) signals give uniformly bounded input (u) and output (y) signals.

In addition to the assumptions A1–A3, introduced in § 3, the following assumption will be needed.

(A4) There exists a solution to the differential equations describing the closed -loop system such that $\bar{\varphi}(t)$ is continuous.

This assumption is of a technical nature. It does not seem to be very restrictive. For example, it can easily be shown that the closed-loop system can be written as a differential equation,

$$\dot{x}(t) = f[x(t), t],$$

where $f \in C^1$ if the noise $v(t)$ is continuous. The existence and continuity of the solution then follows from well-known theorems for ordinary differential equations and A4 is satisfied. We will not, however, go into these details.

The proof of the main result relies on some lemmas to be stated next. In these Lemmas, A1–A4 are assumed to be satisfied and it is also assumed that u^M and v are uniformly bounded. First, define

$$\begin{aligned} \tilde{b}_0(t) &= \hat{b}_0(t) - b_0, \\ \tilde{\theta}(t) &= \hat{\theta}(t) - \theta. \end{aligned} \tag{11}$$

LEMMA 1. *Let $\tilde{b}_0(t)$ and $\tilde{\theta}(t)$ be defined by* (11). *Then the following holds for the algorithm* (10):

$$\frac{d}{dt}(\tilde{b}_0^2(t) + b_0\tilde{\theta}^T(t)\tilde{\theta}(t)) \leqq -\frac{\varepsilon^2(t)}{r(t)} + \frac{1}{r(t)}\left(\frac{AR}{P}\bar{v}(t)\right)^2. \tag{12}$$

Proof. The equations (10a) and (10b) can quivalently be written in terms of $\dot{\tilde{b}}_0$ and $\dot{\tilde{\theta}}$. Then

$$\tilde{b}_0(t)\dot{\tilde{b}}_0(t)=\frac{\tilde{b}_0(t)\varepsilon(t)}{r(t)}\left(\frac{\bar{u}(t)}{P_1}+\hat{\theta}^T(t)\bar{\varphi}(t)\right),$$

$$\tilde{\theta}^T(t)\dot{\tilde{\theta}}(t)=\tilde{\theta}^T(t)\bar{\varphi}(t)\frac{\varepsilon(t)}{r(t)},$$

which from (9) and (10d, e) implies

$$\begin{aligned}\frac{d}{dt}(\tilde{b}_0^2(t)+b_0\tilde{\theta}^T(t)\tilde{\theta}(t))&=2\frac{\varepsilon(t)}{r(t)}\left[\tilde{b}_0(t)\left(\frac{\bar{u}(t)}{P_1}+\hat{\theta}^T(t)\bar{\varphi}(t)\right)+b_0\tilde{\theta}^T(t)\varphi(t)\right]\\&=2\frac{\varepsilon(t)}{r(t)}\left(\frac{AR}{P}\bar{v}(t)-\varepsilon(t)\right)\\&=\frac{1}{r(t)}\left[-\varepsilon^2(t)+\left(\frac{AR}{P}\bar{v}(t)\right)^2-\left(\varepsilon(t)-\frac{AR}{P}\bar{v}(t)\right)^2\right]\\&\leqq-\frac{\varepsilon^2(t)}{r(t)}+\frac{1}{r(t)}\left(\frac{AR}{P}\bar{v}(t)\right)^2.\quad\square\end{aligned}$$

The next lemma gives a useful expression for the evolution of $\bar{\varphi}(t)$.

LEMMA 2. *The vector $\bar{\varphi}(t)$ satisfies the equation*

$$\dot{\bar{\varphi}}(t)=F\bar{\varphi}(t)+ge_f(t)+\psi(t),\tag{13}$$

where F is a constant matrix with its eigenvalues in the left halfplane, g is a constant column vector, and $\psi(t)$ is a uniformly bounded vector sequence.

Proof. Denote the kth element in $\bar{\varphi}(t)$ by $\bar{\varphi}_k(t)$. Using (8), (5), and (6), we have for $0\leqq k\leqq m+n_T-1$:

$$\begin{aligned}\bar{\varphi}_{m+nT-k}(t)&=\frac{p^k}{P}\bar{u}(t)=\frac{p^k}{P}\cdot\frac{A}{b_0B}(\bar{y}(t)-\bar{v}(t))\\&=\frac{p^kA}{b_0TA^MB}e_f(t)+\frac{p^kA}{b_0PB}(\bar{y}^M(t)-\bar{v}(t))\\&\triangleq\frac{p^kA}{b_0TA^MB}e_f(t)+w_{m+n_T-k}(t),\end{aligned}\tag{14}$$

where $w_{m+n_T-k}(t)$ is bounded. Analogously, for $0\leqq k\leqq n-1$,

$$\begin{aligned}\bar{\varphi}_{m+n_T+n-k}(t)&=\frac{p^k}{P}\bar{y}(t)=\frac{p^kB}{TA^MB}e_f(t)+\frac{p^k}{P}\bar{y}^M(t)\\&\triangleq\frac{p^kB}{TA^MB}e_f(t)+w_{m+n_T+n-k}(t).\end{aligned}\tag{15}$$

Now use (3) to get

$$\begin{aligned}\dot{\bar{\varphi}}_1(t)&=\frac{p^{m+n_T}A}{b_0TA^MB}e_f(t)+\dot{w}_1(t)=\frac{1}{b_0}e_f(t)+\frac{p^{m+n_T}A-TA^MB}{b_0TA^MB}e_f(t)+\dot{w}_1(t)\\&=\frac{1}{b_0}e_f(t)+\frac{A(p^{m+n_T}-BR)-BS}{b_0TA^MB}e_f(t)+\dot{w}_1(t).\end{aligned}$$

Since $w_1(t)$ is bounded, it is easy to see from (14) and (15) that the two last terms can be written as a linear combination of $\bar{\varphi}_k : s$ and bounded signals. It is trivial to show that $\dot{\bar{\varphi}}_{m+n_T+1}(t)$ can also be written in this way. Finally, $\bar{\varphi}_{m+n_T+n+1}(t)$ satisfies the equation

$$\dot{\bar{\varphi}}_{m+n_T+n+1}(t) = -\bar{\varphi}_{m+n_T+n+1}(t) - \frac{(p+1)TB^M}{P}\bar{u}^M(t),$$

where the last term is bounded. Summarizing, $\bar{\varphi}(t)$ satisfies the equation

$$\dot{\bar{\varphi}}(t) = F\bar{\varphi}(t) + ge_f(t) + \psi(t),$$

for some constant matrix F, constant vector g, and bounded vector $\psi(t)$. It follows from (14) and (15) that the matrix F has the strictly stable characteristic polynomial $TA^MB(p+1)$. □

The next lemma shows that $|\bar{\varphi}(t)|$ cannot increase arbitrarily fast if the estimates are assumed to be bounded.

LEMMA 3. *Assume that the parameter estimates are uniformly bounded. Then there exist constants K_1 and K_2 so that*

$$|\bar{\varphi}(t)| \leqq e^{K_1(t-s)}(|\bar{\varphi}(s)| + K_2), \qquad \forall t \geqq s. \tag{16}$$

Proof. Using (9), (10f), and (13), we get

$$\begin{aligned}\dot{\bar{\varphi}}(t) &= F\bar{\varphi}(t) + g\left[b_0\left(\theta - \frac{P_1(0)}{P_1(p)}\hat{\theta}(t)\right)^T\bar{\varphi} + \frac{AR}{P}\bar{v}(t)\right] + \psi(t)\\ &\triangleq A(t)\bar{\varphi}(t) + b(t),\end{aligned}$$

where $A(t)$ is bounded because $\hat{\theta}(t)$ is bounded and $b(t)$ is bounded because $\psi(t)$ and $AR/P\bar{v}(t)$ are. Integrating this differential equation and applying the Groenwall-Bellman lemma gives (16). □

From Lemma 3 the following estimate of $\hat{e}_f(t)$ can be derived.

LEMMA 4. *Assume that the parameter estimates are uniformly bounded. Then, for arbitrary $T \geqq 0$,*

$$\hat{e}_f(t) = \xi^T(t)\bar{\varphi}(t) + \eta(t), \qquad t \geqq T, \tag{17}$$

where the vector ξ and the scalar η satisfy

$$|\xi(t)| \leqq K_4 e^{-cT}, \tag{18}$$

$$|\eta(t)| \leqq K_5 e^{K_1T}\int_{t-T}^{t} e^{-c(t-s)}|\varepsilon(s)|\,ds, \tag{19}$$

for some $c > 0$ and K_4, K_5 independent of T.

Proof. It is seen from (10e, f, b) that

$$\begin{aligned}\hat{e}_f(t) &= \hat{b}_0(t)\left(\frac{\bar{u}(t)}{P_1} + \hat{\theta}^T(t)\bar{\varphi}(t)\right) = \hat{b}_0(t)\left(\hat{\theta}^T(t) - \frac{P_1(0)}{P_1}\hat{\theta}^T(t)\right)\bar{\varphi}(t)\\ &= \hat{b}_0(t)[G(p)\dot{\hat{\theta}}^T(t)]\bar{\varphi}(t) = \hat{b}_0(t)\left(G(p)\frac{\bar{\varphi}^T(t)\varepsilon(t)}{r(t)}\right)\bar{\varphi}(t),\end{aligned}$$

where

$$G(p) = \frac{\left[\dfrac{(P_1(p) - P_1(0))}{p}\right]}{P_1(p)}$$

is a strictly proper, asymptotically stable transfer operator and $\hat{b}_0(t)$ is bounded. If the output of the filter $G(p)$ is expressed as a convolution integral which is split into two parts, we thus get

$$\hat{e}_f(t) = \xi^T(t)\bar{\varphi}(t) + \eta(t),$$

where

$$|\xi(t)| \leqq K_3\left(e^{-ct} + \int_0^{t-T} e^{-c(t-s)} \frac{|\bar{\varphi}(s)| \cdot |\varepsilon(s)|}{r(s)}\, ds\right),$$

$$|\eta(t)| \leqq K_3 |\bar{\varphi}(t)| \int_{t-T}^{t} e^{-c(t-s)} \frac{|\bar{\varphi}(s)| \cdot |\varepsilon(s)|}{r(s)}\, ds.$$

The two estimates (18) and (19) will be derived separately. First use the boundedness of the estimates and the noise to conclude from (9) and (10d, e, f) that for some K_ε and K_v,

$$|\varepsilon(t)| \leqq K_\varepsilon |\bar{\varphi}(t)| + K_v.$$

Hence,

$$\begin{aligned} |\xi(t)| &\leqq K_3\, e^{-cT}\left(1 + \int_0^{t-T} e^{-c(t-T-s)} \frac{|\bar{\varphi}(s)| \cdot |\varepsilon(s)|}{r(s)}\, ds\right) \\ &\leqq K_3\, e^{-cT}\left(1 + \frac{1}{c} \sup_{s \leqq t-T} \frac{|\bar{\varphi}(s)|(K_\varepsilon |\bar{\varphi}(s)| + K_v)}{\alpha + |\bar{\varphi}(s)|^2}\right) \\ &\leqq K_4\, e^{-cT}. \end{aligned}$$

The second term is estimated using Lemma 3:

$$\begin{aligned} |\eta(t)| &\leqq K_3\, e^{K_1 T} \int_{t-T}^{t} e^{-c(t-s)} [|\bar{\varphi}(s)| + K_2] \frac{|\bar{\varphi}(s)| \cdot |\varepsilon(s)|}{r(s)}\, ds \\ &\leqq K\, e^{K_1 T} \int_{t-T}^{t} e^{-c(t-s)} \frac{(K_2+1)|\bar{\varphi}(s)|^2 + K_2}{\alpha + |\bar{\varphi}(s)|^2} |\varepsilon(s)|\, ds \\ &\leqq K_5\, e^{K_1 T} \int_{t-T}^{t} e^{-c(t-s)} |\varepsilon(s)|\, ds. \quad \square \end{aligned}$$

Combining Lemmas 2 and 4 now gives an expression for $\bar{\varphi}(t)$ in terms of $\varepsilon(t)$.

LEMMA 5. *Assume that the parameter estimates are uniformly bounded. Then $\bar{\varphi}(t)$ satisfies*

$$(20) \qquad |\bar{\varphi}(t)| \leqq K\left(e^{-c(t-s)} |\bar{\varphi}(s)| + (t-s) \int_{s-T}^{t} |\varepsilon(\sigma)|\, d\sigma\right), \qquad \forall t \geqq s+1,$$

where K and T are positive constants and c is defined in Lemma 4.

Proof. It follows directly from Lemmas 2 and 4 that

$$\dot{\bar{\varphi}}(t) = F(t)\bar{\varphi}(t) + g(\varepsilon(t) + \eta(t)) + \psi(t),$$

where $F(t) \triangleq F + g\xi^T(t)$ can be made exponentially stable by choosing T appropriately. If the transition matrix of $F(t)$ is denoted $\Phi(t, s)$, we have

$$\|\Phi(t, s)\| < K\, e^{-r_F(t-s)}, \qquad t \geqq s,$$

for some positive K and r_F. There is no loss of generality in assuming $r_F = c$, with c in

Lemma 4 (take the largest). Thus, using Lemma 4, we have for $t \geqq s+1$:

$$|\bar{\varphi}(t)| \leqq \|\Phi(t,s)\| \cdot |\bar{\varphi}(s)| + \left| \int_s^t \Phi(t,\sigma) g(\varepsilon(\sigma)+\eta(\sigma))\, d\sigma \right| + \left| \int_s^t \Phi(t,\sigma)\psi(\sigma)\, d\sigma \right|$$

$$\leqq K\Big(e^{-c(t-s)}|\bar{\varphi}(s)| + \int_s^t e^{-c(t-\sigma)}|\varepsilon(\sigma)|\, d\sigma$$

$$+ \int_s^t e^{-c(t-\sigma)}\Big(K_5\, e^{K_1 T} \int_{\sigma-T}^{\sigma} e^{-c(\sigma-\tau)}|\varepsilon(\tau)|\, d\tau \Big)\, d\sigma \Big)$$

$$\leqq K\Big(e^{-c(t-s)}|\bar{\varphi}(s)| + (t-s) \int_{s-T}^{t} e^{-c(t-\sigma)}|\varepsilon(\sigma)|\, d\sigma \Big),$$

where the constant K is different in the different expressions. Thus (20) follows. □

Main results. Lemmas 1 and 5 are the main ingredients when proving the main result, given by the following theorem.

THEOREM 1. *Consider the plant* (5) *controlled by the algorithm* (10). *Assume that* A1–A4 *are satisfied and that the parameter estimates are uniformly bounded. Then the closed-loop system is L^∞-stable.*

Proof. The proof is given in the Appendix. □

Theorem 1 can be specialized in different ways. One motive for the stability investigations were the problems with MRAS. In that case noise is generally not included in the problem formulation and Theorem 1 can be specialized to give the solution.

THEOREM 2. *Consider the plant* (5) *with no noise, i.e.,* $v(t)=0$, *controlled by the algorithm* (10). *Assume that* A1–A4 *are satisfied. Then the closed-loop system is L^∞-stable.*

Proof. The boundedness of the parameter estimates follows immediately from Lemma 1, and Theorem 1 can be applied. □

Theorem 2 gives a fairly satisfactory stability result for the deterministic case. A natural question is whether it is possible to extend the result in Theorem 2 to the case of disturbances which are not zero. The assumption of bounded estimates in Theorem 1 is, however, difficult to verify a priori. Two possibilities to modify the algorithm to ensure bounded estimates are presented below.

THEOREM 3. *Consider the plant* (5) *controlled by the algorithm* (10) *modified in the following way*:

$$\left.\begin{aligned} \dot{\hat{b}}_0(t) &= 0 \\ \dot{\hat{\theta}}(t) &= 0 \end{aligned}\right\} \quad \text{if} \quad |\varepsilon(t)| < K_v, \tag{21}$$

where K_v is a positive constant, satisfying

$$\sup_t \left| \frac{AR}{P} \bar{v}(t) \right| \leqq K_v. \tag{22}$$

Assume that A1–A4 *are satisfied. Then the closed-loop system is L^∞-stable.*

Proof. Two problems have to be considered in order to apply Theorem 1. It must be shown that the estimates are bounded and the consequences on Theorem 1 of the modification (21) must be examined. First, Lemma 1 gives

$$\frac{d}{dt}(\tilde{b}_0^2(t) + b_0 \tilde{\theta}^T(t)\tilde{\theta}(t)) \leqq -\frac{\varepsilon^2(t)}{r(t)} + \frac{1}{r(t)}\left(\frac{AR}{P}\bar{v}(t) \right)^2 \leqq 0, \tag{23}$$

if

$$|\varepsilon(t)| \geqq K_v \geqq \left|\frac{AR}{P}\bar{v}(t)\right|.$$

On the other hand, it follows from (21) that

$$\frac{d}{dt}(\tilde{b}_0^2(t) + b_0\tilde{\theta}^T(t)\tilde{\theta}(t)) \leqq 0,$$

if $|\varepsilon(t)| < K_v$.

The parameter estimates are thus bounded.

For the second problem, some minor changes are needed in the proof of Theorem 1. See Egardt [6], [8] for details.

The conclusion is that Theorem 1 can be applied and the theorem is proven. □

Remark. The modification (21) introduces a discontinuity in the differential equations. In order to guarantee existence and continuity of the solutions, this modification can be made smoother. This technicality will be left with these remarks.

It is true that the modification of the algorithm involves an unknown quantity, namely the upper bound on the disturbance K_v. However, if the noise level is low, the modification could be of practical value anyhow. Similar modifications are well-known in discrete time adaptive control, see Egardt [6], [8].

Another possibility to obtain bounded estimates is to project them into a bounded area. This idea is exploited in the following theorem.

THEOREM 4. *Consider the plant* (5) *controlled by the algorithm* (10), *modified in the following way*:

$$\left.\begin{aligned} \dot{\hat{b}}_0(t) &= \left[\frac{\bar{u}(t)}{P_1} + \hat{\theta}^T(t)\bar{\varphi}(t)\right]\frac{\varepsilon(t)}{r(t)} - \frac{\gamma}{r(t)}\hat{b}_0(t) \\ \dot{\hat{\theta}}(t) &= \bar{\varphi}(t)\frac{\varepsilon(t)}{r(t)} - \frac{\gamma}{r(t)}\hat{\theta}(t) \end{aligned}\right\}, \quad \text{if } \left|\begin{pmatrix} \hat{b}_0(t) \\ \hat{\theta}(t) \end{pmatrix}\right| \geqq C,$$

where γ *is a positive constant and the constant C satisfies*

$$C > 2\sqrt{\frac{\max(1, b_0)}{\min(1, b_0)}} \cdot \left|\begin{pmatrix} b_0 \\ \theta \end{pmatrix}\right|,$$

where b_0 *and* θ *are the true plant parameters. Assume that* A1–A4 *are satisfied. Then the closed-loop system is* L^∞*-stable.*

The proof of this theorem essentially consists of a verification that Lemma 1 still holds. The proof is found in Egardt [6], [8] and is omitted here.

Remark. The same comment on the discontinuity of the differential equations as above can be made here.

Stability conditions are crucial in the convergence analysis of adaptive schemes. For example, convergence of the output error in the absence of noise could not readily be solved except for the case with pole excess equal to one or two. Compare with the discussion in the introduction.

Theorem 2 proves the boundedness of the closed-loop signals in the disturbance-free case. It thus follows that the output error converges to zero.

THEOREM 5. *Consider the plant* (5) *with no noise, i.e.,* $v(t) = 0$, *controlled by the algorithm* (10). *Assume that* A1–A4 *are satisfied and that the command input* $u^M(t)$ *is uniformly bounded. Then the output error converges to zero, i.e.,*

$$y(t) - y^M(t) \to 0, \qquad t \to \infty.$$

Proof. Lemma 1 gives ($v(t)=0$)

$$\int_0^\infty \frac{\varepsilon^2(t)}{r(t)}\,dt<\infty.$$

But $|\bar{\varphi}(t)|$ is bounded from Theorem 2 and $r(t)$ is therefore also bounded. Hence,

$$\int_0^\infty \varepsilon^2(t)\,dt<\infty. \tag{24}$$

This does not, however, imply that $\varepsilon(t)$ tends to zero. A bound on the derivative of $\varepsilon^2(t)$ is necessary for $\varepsilon(t)$ to converge to zero. It follows from (9) and (10d, e, f) that

$$\varepsilon(t)=-\tilde{b}_0(t)\left(\hat{\theta}(t)-\frac{P_1(0)}{P_1(p)}\hat{\theta}(t)\right)^T\bar{\varphi}(t)-b_0\tilde{\theta}^T(t)\bar{\varphi}(t).$$

Thus, $\varepsilon(t)$ is bounded, because the parameter estimates and $|\bar{\varphi}(t)|$ are bounded. Define

$$H(p)=1-\frac{P_1(0)}{P_1(p)},$$

and differentiate the expression for $\varepsilon(t)$ to get

$$\begin{aligned}\dot{\varepsilon}(t)=&-[\dot{\tilde{b}}_0(t)(H(p)\hat{\theta}(t))^T\bar{\varphi}(t)+\tilde{b}_0(t)(H(p)\dot{\hat{\theta}}(t))^T\bar{\varphi}(t)\\&+\tilde{b}_0(t)(H(p)\hat{\theta}(t))^T\dot{\bar{\varphi}}(t)+b_0\dot{\tilde{\theta}}^T(t)\bar{\varphi}(t)+b_0\tilde{\theta}^T(t)\dot{\bar{\varphi}}(t)]\\=&-\Bigg[(H(p)\hat{\theta}(t))^T\bar{\varphi}(t)\frac{\varepsilon(t)}{r(t)}(H(p)\hat{\theta}(t))^T\bar{\varphi}(t)\\&+\tilde{b}_0(t)\left[H(p)\left(\bar{\varphi}(t)\frac{\varepsilon(t)}{r(t)}\right)\right]^T\bar{\varphi}(t)+\tilde{b}_0(t)(H(p)\hat{\theta}(t))^T\dot{\bar{\varphi}}(t)\\&+b_0\bar{\varphi}^T(t)\frac{\varepsilon(t)}{r(t)}\bar{\varphi}(t)+b_0\tilde{\theta}^T(t)\dot{\bar{\varphi}}(t)\Bigg].\end{aligned}$$

The parameter estimates and $|\bar{\varphi}(t)|$ are bounded. Also, $\varepsilon(t)$ is bounded as was seen above. Furthermore, $H(p)$ is asymptotically stable and $r(t)$ is bounded from below by α. Finally $|\dot{\bar{\varphi}}(t)|$ is bounded from the proof of Lemma 3. It is thus possible to conclude that $\dot{\varepsilon}(t)$ is bounded. Hence,

$$\frac{d}{dt}[\varepsilon^2(t)]=2\varepsilon(t)\dot{\varepsilon}(t)$$

is bounded. It then follows from (24) that

$$\varepsilon(t)\to 0,\qquad t\to\infty.$$

In the same way as in Lemma 4, we have

$$\hat{e}_f(t)=\hat{b}_0(t)\left(G(p)\frac{\bar{\varphi}^T(t)\varepsilon(t)}{r(t)}\right)\bar{\varphi}(t),$$

where $G(p)$ is a strictly proper, asymptotically stable transfer operator. Since $|\hat{b}_0(t)|$ and $|\bar{\varphi}(t)|$ are bounded and $r(t)\geqq\alpha$, $\forall t$, it thus follows that

$$\hat{e}_f(t)\to 0,\qquad t\to\infty.$$

This implies that

$$e_f(t)=\varepsilon(t)+\hat{e}_f(t)\to 0,\qquad t\to\infty.$$

Hence

$$y(t)-y^M(t)=e(t)=\frac{P(p)}{Q(p)}e_f(t)\to 0, \qquad t\to\infty,$$

because $Q(p)$ is asymptotically stable and $P(p)/Q(p)$ is proper. □

The above result proves the convergence of the output error to zero in the disturbance-free case for a class of adaptive algorithms. Apart from the result by Feuer and Morse [9] this seems to be the first rigorous proof of convergence without a priori assuming closed-loop stability. A similar result has been given by Morse [17].

5. Conclusions. The paper has presented some stability theorems on a fairly general adaptive algorithm in continuous time. It is shown in Egardt [5], [7] that slightly modified MRAS algorithms by Monopoli [16], Narendra and Valavani [19], Bénéjean [4], and Feuer and Morse [9] can be treated as special cases of the algorithm considered here.

Theorems 2 and 5 prove the convergence of the plant output to the desired output in the disturbance-free case. Unlike most earlier convergence studies, the result does not require any assumption of closed-loop stability.

In the case with disturbances it has been pointed out that some additional assumption is needed to guarantee global stability. The approach taken here is to assume that the parameter estimates are bounded and two different means to ensure this were considered in Theorems 3 and 4. Another possibility is to put more conditions on the noise and/or command signal. It does not seem unreasonable that some kind of persistently exciting condition (see, e.g., Åström and Bohlin [2], Kudva and Narendra [12]) might be sufficient to ensure the boundedness of the parameter estimates. This is however still an open problem. It should finally be pointed out that the case with decreasing gains ($\lambda=0$) in the estimation algorithm has not been treated at all.

Some comments should also be made on the structure of the estimation scheme. A model structure which is *bilinear* in the unknown parameters b_0 and θ is used. It has been pointed out in Egardt [6], [8] that it is not straightforward to extend the stability results to models which are *linear* in the unknown parameters. This is an interesting observation which perhaps deserves further investigation.

Finally note that the results are valid for minimum phase systems only. This is a consequence of the choice of design method. It is naturally of interest to investigate the properties of algorithms which are capable of controlling nonminimum phase systems. It seems that such analysis has not been carried out so far.

Acknowledgments. The material presented here is part of a Ph.D. thesis, Egardt [6]. I would like to thank my adviser Professor K. J. Åström who proposed the problem and provided stimulating support.

Appendix—proof of theorem 1. A single realization will be considered throughout the proof. The boundedness of $|\bar{\varphi}(t)|$ will be proved by contradiction. Thus, assume that

$$\sup_{t\geqq 0}|\bar{\varphi}(t)|>NM,$$

for N and M arbitrarily large. This assumption will be contradicted for some N and M. Assuming the unboundedness, t_{NM} and t_M are well-defined if $N>1$ and $M>|\bar{\varphi}(0)|$:

$$t_{NM}=\min\{t\,|\,|\bar{\varphi}(t)|=NM\}$$

$$t_M=\max\left\{t\,\middle|\,t<t_{NM};\ |\bar{\varphi}(t)|=M;\ |\bar{\varphi}(s)|<M\ \forall s\in\left(\max\left(0,t-\frac{1}{c}\ln N\right),t\right)\right\}.$$

Here the continuity of $|\varphi(t)|$ is used. The constant c is defined in Lemma 4. A typical realization of $|\bar{\varphi}(t)|$ in the time interval $[t_M, t_{NM}]$ is shown in Fig. A.1.

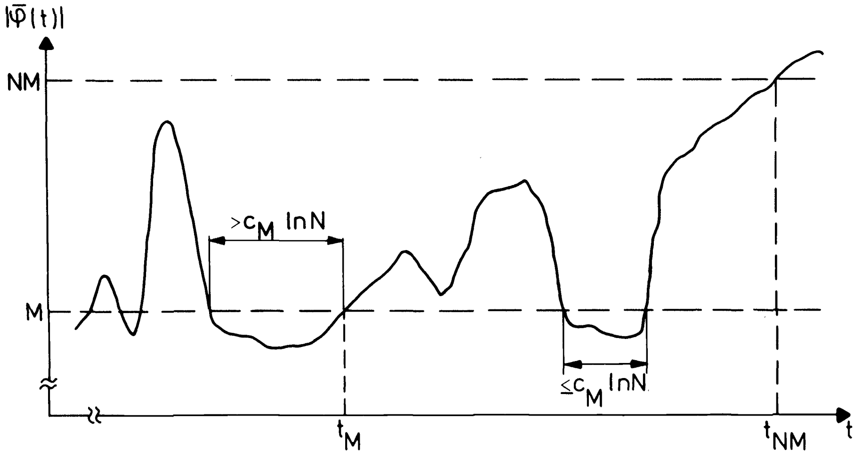


FIG. A.1. *The behavior of* $|\bar{\varphi}(t)|$ *in the interval* $[t_M, t_{NM}]$.

The contradiction will follow from thorough analysis of the algorithm in the interval $[t_M, t_{NM}]$. An outline of the proof is as follows. In Step 1 an increasing sequence $\{|\bar{\varphi}(\tau_i)|\}_{i=1}^{N_\tau}$ in the interval $[t_M, t_{NM}]$ is defined and a lower bound on N_τ is given. Step 2 derives an upper bound on $\tau_{N_\tau} - \tau_1$. This is used in Step 3 to derive an upper bound on N_τ which is in disagreement with the result in Step 1 and the boundedness of $|\bar{\varphi}(t)|$ is thereby proved. The boundedness of $u(t)$ and $y(t)$ is then easily concluded in the last step of the proof.

Before proceeding to the first step of the proof, just note that the following inequality follows from the definition of t_M and t_{NM} and Lemma 3:

$$\text{(A.1)} \qquad \min_{t_M \leq s \leq t_{NM}} |\bar{\varphi}(s)| \geq \frac{M}{N^{K_1/c}} - K_2.$$

This follows from simple calculations which are omitted here.

Step 1. *Characterization of the sequence* $\{\bar{\varphi}(\tau_i)\}$. The sequence $\{\tau_i\}_{i=1}^{N_\tau}$ is defined recursively from

$$\tau_1 = t_M$$

$$\tau_{i+1} = \inf \{t \mid \tau_i + n_\tau \leq t < t_{NM}, |\bar{\varphi}(t)| \geq \sup_{t_M \leq s \leq t} |\bar{\varphi}(s)|\},$$

where n_τ is chosen to satisfy the conditions:

$$\text{(A.2)} \qquad \begin{aligned} &\text{(i)} \;\; n_\tau \geq \max(T, 1) \\ &\text{(ii)} \;\; K e^{-cn_\tau} \leq \tfrac{1}{2}. \end{aligned}$$

Here T, c, and K are defined in Lemma 5.

Let M satisfy the condition $M \geq K_2$, with K_2 as in Lemma 3. It should be noted that N and M can be chosen arbitrarily. A number of conditions of the type above will appear in the proof. They are however easy to fulfill by choosing N and M appropriately. It is, however, important that the constants appearing in the conditions do not depend on the choice of intervals $[t_M, t_{NM}]$, i.e., on N and M themselves. This fact will not be commented upon in the sequel.

If M is chosen to fulfill the condition above, Lemma 3 gives rise to an inequality in the following way. Separate between two cases:

(i) $\tau_{i+1}=\tau_i+n_\tau$; then

$$|\bar{\varphi}(\tau_{i+1})|\leqq 2\,e^{K_1 n_\tau}|\bar{\varphi}(\tau_i)|.$$

(ii) $\tau_{i+1}>\tau_i+n_\tau$; the definition of $\{\tau_i\}$ then implies

$$|\bar{\varphi}(\tau_i+n_\tau)|<\sup_{\tau_i\leqq s\leqq\tau_i+n_\tau}|\bar{\varphi}(s)|$$

and the continuity gives

$$|\bar{\varphi}(\tau_{i+1})|=\sup_{\tau_i\leqq s\leqq\tau_i+n_\tau}|\bar{\varphi}(s)|\leqq 2\,e^{K_1 n_\tau}|\bar{\varphi}(\tau_i)|.$$

The same inequality thus holds in both cases. Using this together with the fact that

$$t_{NM}-\tau_{N_\tau}<n_\tau,$$

which follows from the definition of $\{\tau_i\}$ and the continuity, the following is obtained:

$$\begin{aligned}NM&=|\bar{\varphi}(t_{NM})|\leqq 2\,e^{K_1 n_\tau}|\bar{\varphi}(\tau_{N_\tau})|\leqq\cdots\leqq 2^{N_\tau}e^{K_1 n_\tau N_\tau}|\bar{\varphi}(\tau_1)|\\&=2^{N_\tau}e^{K_1 n_\tau N_\tau}M,\end{aligned}$$

which implies

$$N_\tau\leqq\frac{\ln N}{\ln 2+K_1 n_\tau}. \tag{A.3}$$

This is the lower bound on N_τ sought for in Step 1.

Step 2. *Derivation of an upper bound on* $\tau_{N_\tau}-\tau_1$. Define intervals

$$I_i=[\tau_{i-1},\tau_{i+1}],\qquad i=2,4,\cdots,2N_I,$$

where the number of intervals N_I satisfies

$$N_I=\begin{cases}\frac{1}{2}(N_\tau-1), & N_\tau \text{ odd},\\ \frac{1}{2}(N_\tau-2), & N_\tau \text{ even}.\end{cases} \tag{A.4}$$

Consider an interval I_i and define the sequence $\{T_j^i\}_{j=0}^{N_T^i}$ inside the interval through

$$\begin{aligned}T_0^i&=\tau_{i-1},\\ T_j^i&=\min\{t\,|\,t\geqq T_{j-1}^i+n_T,\ |\bar{\varphi}(t)|\geqq M\},\qquad j=1,\cdots,N_T^i,\end{aligned} \tag{A.5}$$

where N_T^i satisfies

$$\tau_{i+1}-n_T\leqq T_{N_T^i}^i\leqq\tau_{i+1}. \tag{A.6}$$

The left inequality follows because $|\bar{\varphi}(\tau_{i+1})|\geqq M$. The constant n_T is defined as

$$n_T=\frac{2}{c}\ln N. \tag{A.7}$$

Let ΔT be the maximal distance between any T_j^i and T_{j+1}^i. It follows from the definition of t_M (cf. Fig. A.1) and (A.5) that

$$\Delta T\leqq n_T+\frac{1}{c}\ln N=\left(\frac{2}{c}+\frac{1}{c}\right)\ln N\triangleq K_\Delta\ln N, \tag{A.8}$$

where K_Δ is independent of N.

Define intervals

$$J_j^i = [T_{j-1}^i, T_{j+1}^i], \qquad j = 1, 2, \cdots, 2N_j^i - 1,$$

where the number of intervals N_j^i satisfies

$$\text{(A.9)} \qquad N_J^i = \begin{cases} \frac{1}{2}(N_T^i - 1), & N_T^i \text{ odd}, \\ \frac{1}{2}N_T^i, & N_T^i \text{ even}. \end{cases}$$

The behavior of the algorithm in an interval J_j^i will now be examined. Distinguish between two cases.

The case $N_T^i \leqq 2$. From (A.9) it is seen that there is at least one interval J_j^i in the interval I_i. Suppose that

$$\text{(A.10)} \qquad \int_{T_{j-1}^i}^{T_{j+1}^i} |\varepsilon(s)|\, ds < \frac{M}{K\, \Delta T}\ .$$

This will lead to a contradiction. First note that from (A.7)

$$K e^{-cnT} = KN^{-2} \leqq \frac{1}{2N},$$

for large N. It is thus possible to use Lemma 5 to obtain

$$\begin{aligned} |\bar\varphi(T_{j+1}^i)| &\leqq \frac{|\bar\varphi(T_j^i)|}{2N} + K\, \Delta T \int_{T_j^i - T}^{T_{j+1}^i} |\varepsilon(s)|\, ds \\ &\leqq \frac{NM}{2N} + K\, \Delta T \int_{T_{j-1}^i}^{T_{j+1}^i} |\varepsilon(s)|\, ds < M, \end{aligned}$$

where the fact that $n_T \geqq T$ (for large N) and the assumption (A.10) have been used in the last two steps.

We have thus arrived at a contradiction and the conclusion is that

$$\text{(A.11)} \qquad \int_{T_{j-1}^i}^{T_{j+1}^i} |\varepsilon(s)|\, ds \geqq \frac{M}{K\, \Delta T}.$$

The inequality holds for every interval J_j^i. Define

$$\text{(A.12)} \qquad V(t) = \tilde b_0^2(t) + b_0 \tilde\theta^T(t) \tilde\theta(t).$$

Lemma 1 gives, for some K_v,

$$\begin{aligned} V(T_{j+1}^i) - V(T_{j-1}^i) &\leqq -\int_{T_{j-1}^i}^{T_{j+1}^i} \frac{\varepsilon^2(s)}{r(s)}\, ds + \int_{T_{j-1}^i}^{T_{j+1}^i} \left(\frac{AR}{P} \bar v(s)\right)^2 \frac{ds}{r(s)} \\ &\leqq -\int_{T_{j-1}^i}^{T_{j+1}^i} \frac{\varepsilon^2(s)}{r(s)}\, ds + \int_{T_{j-1}^i}^{T_{j+1}^i} \frac{K_v^2}{r(s)}\, ds, \end{aligned}$$

since the disturbance is bounded. Note that, for $T_{j-1}^i \leqq s \leqq T_{j+1}^i$,

$$r(s) = \alpha + |\bar\varphi(s)|^2 \leqq 2(NM)^2,$$

if N and M are chosen sufficiently large. Now choose

$$\text{(A.13)} \qquad M = N^p = N^{K_1/c+2}.$$

It then follows from (A.1) that, for large N,

$$r(s) \geqq |\bar{\varphi}(s)|^2 \geqq \left(\frac{M}{N^{K_1/c}} - K_2\right)^2 \geqq \tfrac{1}{2} N^{2(p-K_1/c)} = \tfrac{1}{2} N^4. \tag{A.14}$$

Apply these two inequalities above to obtain

$$V(T^i_{j+1}) - V(T^i_{j-1}) \leqq -\frac{1}{2(NM)^2} \int_{T^i_{j-1}}^{T^i_{j+1}} \varepsilon^2(s)\, ds + \frac{4\,\Delta T K_v^2}{N^4}.$$

Now, use Schwarz' inequality to obtain for N sufficiently large:

$$\begin{aligned} V(T^i_{j+1}) - V(T^i_{j-1}) &\leqq -\frac{1}{2(NM)^2} \cdot \frac{1}{(T^i_{j+1} - T^i_{j-1})} \left[\int_{T^i_{j-1}}^{T^i_{j+1}} |\varepsilon(s)|\, ds\right]^2 + \frac{4\,\Delta T K_v^2}{N^4} \\ &\leqq -\frac{1}{4\,\Delta T (NM)^2} \left[\int_{T^i_{j-1}}^{T^i_{j+1}} |\varepsilon(s)|\, ds\right]^2 + \frac{4\,\Delta T K_v^2}{N^4} \\ &\leqq -\frac{1}{4\,\Delta T N^2 (K\,\Delta T)^2} + \frac{4\,\Delta T K_v^2}{N^4} \\ &\triangleq -\frac{c_1}{N^2\,\Delta T^3} + \frac{c_2\,\Delta T}{N^4}, \end{aligned} \tag{A.15}$$

where c_1 and c_2 are independent of N.

It follows from (A.8) that for large N

$$\Delta T^3 \leqq K_\Delta^3 N.$$

Inserting this inequality into (A.15) and also using (A.8) gives

$$\begin{aligned} V(T^i_{j+1}) - V(T^i_{j-1}) &\leqq -\frac{c_1}{N^2 K_\Delta^3 N} + \frac{c_2 K_\Delta \ln N}{N^4} \\ &= -\frac{1}{N^3}\left[\frac{c_1}{K_\Delta^3} - c_2 K_\Delta \frac{\ln N}{N}\right] \\ &\leqq -\frac{c_0}{N^3}, \quad N \text{ sufficiently large,} \end{aligned}$$

where c_0 is a constant, independent of N.

This inequality holds for every interval J^i_j, i.e., N^i_j times, whence

$$V(T^i_{2N^i_J}) - V(T^i_0) \leqq -c_0 \frac{N^i_J}{N^3}.$$

But V is positive and also from the assumptions bounded, by $\tilde{K}_v$ say, so that

$$N^i_J \leqq \frac{\tilde{K}_v}{c_0} \cdot N^3. \tag{A.16}$$

The case $N^i_T < 2$. The inequality (A.16) is trivially satisfied also in this case, because $N^i_T < 2$ implies $N^i_J = 0$.

The conclusion is thus that (A.16) holds in every interval I_i provided N is chosen large enough. From (A.6) and (A.9) it follows that

$$\tau_{i+1} - T^i_{2N^i_J} = (\tau_{i+1} - T^i_{N^i_T}) + (T^i_{N^i_T} - T^i_{2N^i_J}) \leqq n_T + \Delta T \leqq 2\,\Delta T,$$

which together with (A.16) gives

$$\tau_{i+1}-\tau_{i-1}=(\tau_{i+1}-T^i_{2N^i_j})+(T^i_{2N^i_j}-T^i_0)\leqq 2\,\Delta T+2N^i_J\,\Delta T$$

$$\leqq 2\,\Delta T\left(1+\frac{\tilde{K}_v}{c_0}\cdot N^3\right)\leqq\frac{4\tilde{K}_v}{c_0}\Delta T\cdot N^3.$$

Summing for $i=2,4,\cdots,2N_I$ gives

$$\tau_{2N_I+1}-\tau_1\leqq\frac{4\tilde{K}_v}{c_0}\Delta T\cdot N^3\cdot N_I, \tag{A.17}$$

which concludes Step 2 of the proof.

Step 3. *Derivation of an upper bound on* N_τ *and the contradiction.* Consider an interval I_i defined in Step 2. Suppose that

$$\int_{\tau_{i+1}-2n_\tau}^{\tau_{i+1}}|\varepsilon(s)|\,ds<\frac{|\bar{\varphi}(\tau_{i+1})|}{2Kn_\tau}. \tag{A.18}$$

This assumption will lead to a contradiction in much the same way as in Step 2. From Lemma 5 we have

$$|\bar{\varphi}(\tau_{i+1})|\leqq\frac{|\bar{\varphi}(\tau_{i+1}-n_\tau)|}{2}+Kn_\tau\cdot\int_{\tau_{i+1}-n_\tau-T}^{\tau_{i+1}}|\varepsilon(s)|\,ds$$

$$\leqq\frac{|\bar{\varphi}(\tau_{i+1})|}{2}+Kn_\tau\cdot\int_{\tau_{i+1}-2n_\tau}^{\tau_{i+1}}|\varepsilon(s)|\,ds<|\bar{\varphi}(\tau_{i+1})|,$$

where the properties (A.2i, ii) of n_τ and the assumption (A.18) have been used.

We have thus arrived at a contradiction and the conclusion is that

$$\int_{\tau_{i+1}-2n_\tau}^{\tau_{i+1}}|\varepsilon(s)|\,ds\geqq\frac{|\bar{\varphi}(\tau_{i+1})|}{2Kn_\tau}. \tag{A.19}$$

The inequality holds for every interval I_i. We have, for $\tau_{i-1}\leqq s\leqq\tau_{i+1}$,

$$r(s)=\alpha+|\bar{\varphi}(s)|^2\leqq 2|\bar{\varphi}(\tau_{i+1})|^2,$$

for N sufficiently large. Applying Lemma 1 in the same way as in Step 2 now gives a result analogous with (A.15) for large N:

$$V(\tau_{i+1})-V(\tau_{i-1})\leqq-\int_{\tau_{i-1}}^{\tau_{i+1}}\frac{\varepsilon^2(s)}{r(s)}\,ds+\int_{\tau_{i-1}}^{\tau_{i+1}}\left(\frac{AR}{P}\bar{v}(s)\right)^2\frac{ds}{r(s)}$$

$$\leqq-\int_{\tau_{i+1}-2n_\tau}^{\tau_{i+1}}\frac{\varepsilon^2(s)}{r(s)}\,ds+\int_{\tau_{i-1}}^{\tau_{i+1}}\frac{K_v^2}{r(s)}\,ds$$

$$\leqq-\frac{1}{2|\varphi(\tau_{i+1})|^2}\cdot\frac{1}{2n_\tau}\left[\int_{\tau_{i+1}-2n_\tau}^{\tau_{i+1}}|\varepsilon(s)|\,ds\right]^2+\frac{2K_v^2(\tau_{i+1}-\tau_{i-1})}{N^4}$$

$$\leqq-\frac{1}{4n_\tau[2Kn_\tau]^2}+\frac{2K_v^2(\tau_{i+1}-\tau_{i-1})}{N^4}$$

$$\triangleq -c_3+c_4\frac{\tau_{i+1}-\tau_{i-1}}{N^4},$$

where c_3 and c_4 are independent of N and (A.19) has been used in the second last step. Summing the inequality for $i = 2, 3, \cdots, 2N_I$ gives

$$\begin{aligned} V(\tau_{2N_1+1}) - V(\tau_1) &\leqq -c_3N_I + c_4\frac{\tau_{2N_I+1} - \tau_1}{N^4} \\ &\leqq -c_3N_I + c_4\frac{4\tilde{K}_v}{c_0}\cdot\frac{\Delta T\cdot N^3\cdot N_I}{N^4} \\ &= -N_I\left(c_3 - c_4\frac{4\tilde{K}_v\,\Delta T}{c_0N}\right), \end{aligned}$$

where (A.17) has been used. But V is positive and bounded by $\tilde{K}_v$ as in Step 2, so that

$$-\tilde{K}_v \leqq -N_I\left(c_3 - c_4\frac{4\tilde{K}_v\,\Delta T}{c_0N}\right),$$

which by (A.4) and (A.8) implies

$$N_\tau \leqq 2N_I + 2 \leqq \frac{2\tilde{K}_v}{\left|c_3 - c_4\dfrac{4\tilde{K}_v\,\Delta T}{c_0N}\right|} + 2 \leqq \frac{2\tilde{K}_v}{c_3 - c_{3/2}} + 2 = \frac{4\tilde{K}_v}{c_3} + 2,$$

for N sufficiently large. This result obviously violates the inequality (A.3) obtained in Step 1 for N large enough. The existence of the sequence $\{|\bar{\varphi}(\tau_i)|\}$ for N arbitrarily large is thus contradicted and the boundedness of $|\bar{\varphi}(t)|$ is proved.

Step 4. *Boundedness of u and y.* It remains to conclude boundedness of $u(t)$ and $y(t)$ from the boundedness of $|\bar{\varphi}(t)|$. From (9) and (10f) it is clear that $e_f(t) = (Q/P)[y(t) - y^M(t)]$ is bounded. But $y^M(t)$ is bounded and Q and P are asymptotically stable polynomials of the same degree, which implies that $y(t)$ is bounded.

The boundedness of $u(t)$ is possible to establish from (10f), which can be written

$$P_2\frac{\bar{u}(t)}{P} = -\left(\frac{P_1(0)}{P_1}\hat{\theta}^T(t)\right)\bar{\varphi}(t),$$

or, using the definition of P_2,

$$\begin{aligned} p^{m+n_T}\frac{\bar{u}(t)}{P} = &-\left(p_{21}p^{m+n_T-1}\frac{\bar{u}(t)}{P} + \cdots + p_{2(m+n_T)}\frac{\bar{u}(t)}{P}\right) \\ &-\left(\frac{P_1(0)}{P_1}\hat{\theta}^T(t)\right)\bar{\varphi}(t). \end{aligned} \tag{A.20}$$

Here all terms in the first bracket are components of $\bar{\varphi}(t)$ and it follows that $p^{m+n_T}\bar{u}(t)/P$ is bounded. Differentiating (A.20) $1, 2, \cdots, n-m-1$ times gives recursively boundedness of $p^{m+n_T+1}\bar{u}(t)/P, \cdots, P^{n+n_T-1}\bar{u}(t)/P$. Notice that $\hat{\theta}(t)/P_1$ is possible to differentiate because P_1 is of degree $n-m-1$ and also that the derivatives of $\bar{\varphi}(t)$ are bounded because of earlier steps in the recursion and boundedness of $y(t)$ and $u^M(t)$, cf. (8). Finally, boundedness of $p^{n+n_T}\bar{u}(t)/P$ follows by an additional differentiation of (A.20) but then the boundedness of $\dot{\hat{\theta}}(t)$, which follows from (10b, c, d) is also used. As a result, the first $n + n_T$ derivatives of $\bar{u}(t)/P = (Q/TA^MP)u(t)$ have shown to be bounded. But the pole excess of Q/TA^MP is exactly $n+n_T$ and Q is asymptotically stable. Hence boundedness of $u(t)$ follows readily. The theorem is thus proven. □

REFERENCES

[1] K. J. Åström, *Reglerteori*, Almqvist & Wiksell, Stockholm, 1976.

[2] K. J. Åström and T. Bohlin, *Numerical identification of linear dynamic systems from normal operating records*, IFAC Symp. on Theory of Self-Adaptive Control Systems, Teddington, England (1965).

[3] K. J. Åström, B. Westerberg and B. Wittenmark, *Self-tuning controllers based on pole-placement design*, Department of Automatic Control, Lund Institute of Technology, Lund, Sweden (1978).

[4] R. Bénéjean, *La commande adaptive à modèle de référence évolutif*, Université Scientifique et Médicale de Grenoble, France (1977).

[5] B. Egardt, *Unification of some adaptive control schemes*, Department of Automatic Control, Lund Institute of Technology, Lund, Sweden (1978).

[6] ———, *Stability of model reference adaptive and self-tuning regulators*, Department of Automatic Control, Lund Institute of Technology, Lund, Sweden (1978).

[7] ———, *Unification of some continuous-time adaptive control schemes*, IEEE Trans. Automatic Control, AC–24, (1979), pp. 588–592.

[8] ———, *Stability of Adaptive Controllers*, Springer-Verlag, Heidelberg, Berlin (1979).

[9] A. Feuer and A. S. Morse, *Adaptive control of single-input, single-output linear systems.* Proceedings of the 1977 IEEE Conference on Decision and Control, New Orleans, USA (1977), pp. 1030–1035.

[10] ———, *Local stability of parameter-adaptive control systems*, John Hopkin Conference on Information Science and Systems (1978).

[11] A. Feuer, B. R. Barmish and A. S. Morse, *An unstable dynamical system associated with model reference adaptive control*, IEEE Trans. Automatic Control, AC–23 (1978),, pp. 499–500.

[12] P. Kudva and K. S. Narendra, *An identification procedure for discrete multivariable systems*, IEEE Trans. Automatic Control, AC-19 (1974), pp. 549–552.

[13] I. D. Landau, *A survey of model reference adaptive techniques—theory and applications*, Automatica, 10 (1974), pp. 353–379.

[14] ———, *Adaptive controllers with explicit and implicit reference models and stochastic self-tuning regulators—equivalence and duality aspects*, Proc. 17th IEEE-CDC Conf., San Diego (1979).

[15] L. Ljung and I. D. Landau, *Model reference adaptive systems and self-tuning regulators—some connections*, Preprints of the 7th IFAC World Congress, Helsinki, Finland (1978), pp. 1973–1980.

[16] R. V. Monopoli, *Model reference adaptive control with an augmented error signal*, IEEE Trans. Automatic Control, AC–19 (1973), pp. 474–484.

[17] A. S. Morse, *Global stability of parameter-adaptive control systems*, Department of Engineering and Applied Science, Yale University (1979).

[18] K. S. Narendra and L. S. Valavani, *Stable adaptive observers and controllers*, Proc. IEEE, 64 (1976), pp. 1198–1208.

[19] ———, *Stable adaptive controller design, part I: Direct control*, Proceedings of the IEEE Conference on Decision and Control, New Orleans, USA (1977), pp. 881–886.

[20] ———, *Direct and indirect adaptive control*, Preprints of the 7th IFAC World Congress, Helsinki, Finland (1978), pp. 1981–1987.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 5, September 1980

0363-0129/80/1805-0007 $01.00/0

OPTIMAL STOCHASTIC CONTROL WITH SPECIAL INFORMATION PATTERNS*

NORBERT CHRISTOPEIT†

Abstract. This paper treats the problem of existence of optimal controls in stochastic systems when observations can be taken only at certain discrete times. The technique applied is based on weak convergence of probability measures and the Girsanov measure transformation method.

1. Introduction.

This paper concerns the control of a system whose dynamics are governed by the nonlinear stochastic differential equation

$$dx = g(t, x)h(t, u)\,dt + \sigma(t, x)\,dw, \qquad 0 \leqq t \leqq 1, \tag{1.1}$$

with initial condition

$$x(0) = x_0. \tag{1.2}$$

Here w is an r-dimensional standard Brownian motion, g, h, and σ are nonanticipating matrix valued functions, and the control u is a function whose value at time t depends on specified information about the history of $x(\cdot)$ previous to time t. Control is to be chosen such as to minimize the expected cost

$$J(x, u) = E\left\{\int_0^1 k(t, x)l(t, u)\,dt\right\} + E\{m(x)\}. \tag{1.3}$$

Existence results for this type of problem greatly involve the information pattern available to the controller. The techniques based on the Girsanov measure transformation method (cf. [2], [9]) seem to work only in the case of complete information about the past.

In this paper, we consider information patterns which allow observations to be taken only at finitely many observation times, including the effect of forgetting part of the past information as the time goes on. In the model considered in § 3, the space of observable outcomes is divided into decision regions, and control action is taken according to which decision region is actually hit. The results obtained are then used in § 4 to approximate a larger class of admissible controls.

The techniques applied involve a method used by Kushner (cf. [12]), which is based on weak convergence of probability measures, together with the Girsanov measure transformation method in § 4. The underlying concept of solution to (1.1), (1.2) is hence that of weak solutions (cf. [14]).

Note that the system equation (1.1) allows for drifts of the form

$$f_1(t, x) + f_2(t, u) + \tilde{g}(t, x)\tilde{h}(t, u),$$

by suitable choice of g and h. The same applies to (1.3) if we allow k and l to be vector valued. Since this offers no additional difficulty, we shall confine ourselves to the case of scalar valued functions k and l.

2. Assumptions and formulation of the problem.

The following notations and assumptions will be used throughout.

C^k = space of $\mathbb{R}^k$-valued continuous functions on [0, 1] with the sup norm topology.

* Received by the editors June 18, 1979, and in revised form January 8, 1980.

† Institut fur Ökonometrie und Operations Research, University of Bonn, D-53 Bonn, West Germany.

$\mathscr{C}_t^k$ = σ-algebra on C^k induced by the continuous functions on $[0, t]$, $0 \leqq t \leqq 1$; i.e., $\mathscr{C}_t^k$ is the σ-algebra generated by all sets of the form $\{\xi : \xi(s) \in \Gamma\}$, where $0 \leqq s \leqq t$, Γ is an arbitrary Borel set in $\mathbb{R}^k$ and ξ denotes the generic element of C^k.

Let $\mathscr{U}$ denote the compact metric space of control points, $\mathscr{B}$ and $\mathscr{B}_{\mathscr{U}}$ the Borel σ-fields on $[0, 1]$ and $\mathscr{U}$, respectively.

(A1) $g : [0, 1] \times C^r \to \mathbb{R}^{r \times s}$ is measurable with respect to $\mathscr{B} \otimes \mathscr{C}_1^r$. For each $t \in [0, 1]$, the function $g(t, \cdot)$ is continuous, $\mathscr{C}_t^r$-measurable and bounded on bounded C^r-sets uniformly in t.

(A2) $h : [0, 1] \times \mathscr{U} \to \mathbb{R}^s$ is measurable with respect to $\mathscr{B} \otimes \mathscr{B}_{\mathscr{U}}$ and bounded. For each $t \in [0, 1]$, the function $h(t, \cdot)$ is continuous.

(A3) $k : [0, 1] \times C^r \to \mathbb{R}$ and $l : [0, 1] \times \mathscr{U} \to \mathbb{R}$ are nonnegative functions satisfying (A1) and (A2), respectively.

(A4) For each $t \in [0, 1]$, the extended velocity set

$$\mathscr{V}(t) = \{(h(t, u), l(t, u)) : u \in \mathscr{U}\}$$

is compact and convex.

(A5) $\sigma : [0, 1] \times C^r \to$ nonsingular $r \times r$-matrices is measurable with respect to $\mathscr{B} \otimes \mathscr{C}_1^r$. For each $t \in [0, 1]$, $\sigma(t, \cdot)$ is $\mathscr{C}_t^r$-measurable. Let σ satisfy the Ito conditions

(I1) for every $N > 0$ there exists a constant K_N such that

$$|\sigma(t, \xi) - \sigma(t, \xi')| \leqq K_N \|\xi - \xi'\|_t,$$

for all $t \in [0, 1]$ and all $\|\xi\|, \|\xi'\| \leqq N$ ($\|\cdot\|_t$ = sup norm on $[0, t]$);

(I2) there is a constant K such that

$$|\sigma(t, \xi)| \leqq K(1 + \|\xi\|_t)$$

for all $t \in [0, 1]$ and all $\xi \in C^r$.

Let $\sigma^{-1}(t, \cdot)$ be bounded on bounded C^r-sets uniformly in t.

If (B_t) is a standard Brownian motion defined on some probability space $(\Omega, \mathscr{F}, P)$ then (I1) and (I2) imply that the stochastic differential equation

$$dx_t = \sigma(t, x)\, dB_t, \qquad x(0) = x_0, \tag{2.1}$$

has a unique strong solution. Let us fix an initial distribution F_0 admitting finite second moments. Then the solution of (2.1) with initial value distributed according to F_0 defines a unique probability measure μ on C^r by

$$\mu(A) = P[x \in A], \tag{2.2}$$

(cf. [19]) as well as unique finite dimensional distributions

$$\mu_{t_1, \cdots, t_N}(B) = P[(x_{t_1}, \cdots, x_{t_N}) \in B],$$

$0 < t_1 \leqq t_2 \leqq \cdots \leqq t_N \leqq 1$, B Borel set in $\mathbb{R}^N$, $N = 1, 2, \cdots$.

(A6) Let the finite dimensional distributions of μ be absolutely continuous with respect to the appropriate Lebesgue measure.

(A7) $m : C^r \to \mathbb{R}$ is continuous and nonnegative

Let us now define what we shall understand by an admissible control. Suppose that observations can only be taken at certain times $0 < t_1 < \cdots < t_p = 1$, and that at each of these times certain functionals of the state at the present and all past observation times can be observed. More precisely:

(A8) Let

$$\varphi_j : \mathbb{R}^{rj} \to \mathbb{R}^{k_j}, \qquad j = 1, \cdots, p,$$

be continuous functions (with positive integers k_j).

Then a function $u : [0, 1] \times \Omega \to \mathscr{U}$ defined on some probability space $(\Omega, \mathscr{F}, P)$ will be called an admissible control if there exists a process $x(t)$, $0 \leqq t \leqq 1$, on $(\Omega, \mathscr{F}, P)$ with continuous trajectories and $x(0)$ having the prescribed distribution F_0, such that the following conditions (i)–(iii) are satisfied.

(i) For each $1 \leqq j \leqq p$, $1 \leqq i \leqq k_j$, there exists a finite partition,

$$-\infty = a_{j,i,1} < \cdots < a_{j,i,\nu_{ij}} = \infty,$$

of the real axis, $1 \leqq \nu_{ij} \leqq \bar{\nu}$, together with functions

$$u_{j;\mu_1,\cdots,\mu_{k_j}} \in L_1^m[t_j, t_{j+1}], \qquad 1 \leqq \mu_i \leqq \nu_{ij} \quad \text{for all } i = 1, \cdots, k_j,$$

($L_1^m[t_j, t_{j+1}]$ = space of integrable functions on $[t_j, t_{j+1}]$ with values in $\mathbb{R}^m$) taking on values in $\mathscr{U}$ for almost all $t \in [t_j, t_{j+1}]$, such that

$$u(t, \omega) = u_{j;\mu_1,\cdots,\mu_{k_j}}(t),$$

for $t \in [t_j, t_{j+1}]$ and

$$\varphi_{ji}(x(t_1, \omega), \cdots, x(t_j, \omega)) \in (a_{j,i,\mu_i}, a_{j,i,\mu_i+1}], \qquad i = 1, \cdots, k_j.$$

(ii) There exists an r-dimensional standard Brownian motion $(w(t), \mathscr{F}_t)$, $0 \leqq t \leqq 1$, on $(\Omega, \mathscr{F}, P)$ such that $x(t)$ is nonanticipating with respect to $(\mathscr{F}_t)$ and the Ito equation,

$$x(t) = x(0) + \int_0^t g(s, x)h(s, u(s))\, ds + \int_0^t \sigma(s, x)\, dw(s), \tag{2.3}$$

holds with probability one for all $0 \leqq t \leqq 1$.

The process $x(t)$ will be called a solution of (2.3) corresponding to the control u. Note that no uniqueness of the solution is required.

(iii) There exists a constant K such that, uniformly in $0 \leqq t \leqq t + \Delta \leqq 1$,

(a) $E\int_0^1 |g(s, x)|^2\, ds \leqq K$ and

$$E\left(\int_t^{t+\Delta} |g(s, x)|\, ds\right)^2 \leqq K\Delta^2;$$

(b) $E\int_0^1 |\Sigma(s, x)|^4\, ds \leqq K$, where $\Sigma = \sigma\sigma'$.

An alternative formulation of (i) providing a better link to the commonly used classes of controls is the following. There exists a measurable function $\mathbf{u} : [0, 1] \times C^r \to \mathscr{U}$ with the property that for each $t \in [t_j, t_{j+1})$, $j = 1, \cdots, p-1$, $\mathbf{u}(t, \cdot)$ is measurable with respect to the σ-field $\mathscr{G}_j$ generated by the C^r sets

$$\{\varphi_{ji}(\xi(t_1), \cdots, \xi(t_j)) \in (a_{j,i,\mu_i}, a_{j,i,\mu_i+1}], i = 1, \cdots, k_j\},$$
$$1 \leqq \mu_1 \leqq \nu_{1j}, \cdots, 1 \leqq \mu_{k_j} \leqq \nu_{k_j,j},$$

and

$$u(t, \omega) = \mathbf{u}(t, x(\omega)).$$

Note, however, that $\mathcal{G}_j$ may differ from control to control. Hence $\mathcal{G}_j$ cannot be interpreted as the σ-field containing the information available at time t. Actually, in each time interval $[t_j, t_{j+1})$, $\varphi_j(x(t_1), \cdots, x(t_j))$ can be observed. A control action consists in dividing the range space of φ_j into a finite number of decision regions and choosing a deterministic control function for every region. As to the case where the decision regions are fixed in advance, note the remark at the end of § 3.

Denoting by $\mathcal{A}$ the class of admissible controls, a rigorous formulation of the control problem is the following:

$$\text{(P)}\qquad \text{minimize}\quad J(x,u)=E\left\{\int_0^1 k(t,x)l(t,u(t))\,dt\right\}+E\{m(x)\}$$

in the class $\mathcal{A}$ of admissible controls u and corresponding solutions x.

Note that by virtue of (A3) and (A6) $\hat{J}=\inf\{J(x,u): u\in\mathcal{A}, x \text{ corresponding solution}\}>-\infty$. Let us assume once and for all that $\hat{J}$ is finite.

3. Existence of optimal controls. For an admissible control u with corresponding solution x define functions

$$\text{(3.1)}\qquad \begin{aligned} F(t)&=\int_0^t g(s,x)h(s,u(s))\,ds, & B(t)&=\int_0^t \sigma(s,x)\,dw(s),\\ H(t)&=\int_0^t h(s,u(s))\,ds, & L(t)&=\int_0^t l(s,u(s))\,ds,\end{aligned}$$

and

$$\Phi(t)=(x(t),F(t),B(t),H(t),L(t)).$$

Then Φ is a measurable process on Ω with paths in $S=C^r\times C^{r\times s}\times C^s\times C^r\times C$. It induces on the Borel σ-field on S a probability measure Q by

$$Q(A)=P_\Phi(A)=P[\Phi\in A].$$

Denote by $\mathcal{P}$ the class of all probability measures on S generated in this way with u ranging in $\mathcal{A}$. Then it follows from (iii) and the boundedness of h and l that $\mathcal{P}$ is tight (cf. [3], [4]), hence every sequence in $\mathcal{P}$ contains a weakly convergent subsequence.

The following considerations will show that $\mathcal{P}$ is weakly closed.

Start with a weakly convergent sequence $\Phi^n=(x^n,F^n,B^n,H^n,L^n)$ defined on probability spaces $(\Omega^n,\mathcal{F}^n,P^n)$, i.e.,

$$Q^n\to Q^0,$$

where $Q^n=P^n_{\Phi^n}$ and Q^0 is some probability measure on S. Then, by a theorem of Skorokhod (cf. [16]), there exist measurable processes $\tilde{\Phi}^n=(\tilde{x}^n,\tilde{F}^n,\tilde{B}^n,\tilde{H}^n,\tilde{L}^n)$, $n=0, 1,\cdots$, all defined on the same probability space $(\tilde{\Omega},\tilde{\mathcal{F}},\tilde{P})=([0,1],\mathcal{B},$ Lebesgue measure) such that

$$\text{(3.2)}\qquad \tilde{P}_{\tilde{\Phi}^n}=Q^n,\qquad n=0,1,\cdots,$$

and

$$\text{(3.3)}\qquad \tilde{\Phi}^n\to\tilde{\Phi}^0\qquad \tilde{P}\text{-a.e.},$$

in the topology of S. Since the u^n are of the form described in (i) (with indices n properly affixed), define

$$\tilde{u}^n(t,\tilde{\omega})=u^n_{j;\mu_1,\cdots,\mu_{k_j}}(t),\quad \text{for } t\in[t_j,t_{j+1})$$

and

$$\varphi_{ji}(\tilde{x}^n(t_1, \tilde{\omega}), \cdots, \tilde{x}^n(t_j, \tilde{\omega})) \in (a^n_{j,i,\mu_i}, \cdots, a^n_{j,i,\mu_i+1}], \qquad i = 1, \cdots, k_j.$$

Then it readily follows from (3.2) by arguments used in [12] and [4] that the $\tilde{u}^n$, $n = 1, 2, \cdots$, are admissible controls with corresponding solutions $\tilde{x}^n$. In particular,

$$\begin{aligned}\tilde{x}^n(t) &= \tilde{x}^n(0) + \tilde{F}^n(t) + \tilde{B}^n(t)\\ &= \tilde{x}(0) + \int_0^t g(s, \tilde{x}^n)h(s, \tilde{u}^n(s))\,ds + \int_0^t \sigma(s, \tilde{x}^n)\,d\tilde{w}^n(s)\end{aligned}$$

holds for all t with probability one, and

$$J(\tilde{x}^n, \tilde{u}^n) = J(x^n, u^n), \quad \text{for all } n = 1, 2, \cdots.$$

Moreover, with probability one

$$\tilde{x}^0(t) = \tilde{x}^0(0) + \tilde{F}^0(t) + \tilde{B}^0(t), \tag{3.4}$$

and

$$\tilde{B}^0(t) = \int_0^t \sigma(s, \tilde{x}^0)\,d\tilde{w}^0(s), \tag{3.5}$$

for all t with some standard Brownian motion $(\tilde{w}^0(t), \mathscr{F}^0_t)$ with respect to which $\tilde{x}^0(t)$ is nonanticipative.

Since from now on we will be working only with the tilded processes, let us omit the tilde in the sequel. Let μ^0 denote the measure induced on C^r by x^0, i.e.,

$$\mu^0(A) = P[x^0 \in A].$$

LEMMA 1. *μ^0 is absolutely continuous with respect to the measure μ.*

Proof. Arguing as in [12] it can be shown that there is an integrable function $\bar{f}(t, \omega)$ such that

$$F^0(t) = \int_0^t f(s, \omega)\,ds, \quad P\text{-a.e.},$$

for all t. Note that by (3.4), $F^0(t)$ is measurable and $(\mathscr{F}^0_t)$-adapted. Since $(w^0(t), \mathscr{F}^0_t)$ is a Brownian motion, so is $(w^0(t), \mathscr{F}^0_{t+})$; hence we may assume that the family $(\mathscr{F}^0_t)$ is right continuous and completed, and, by Lemma 5.2 in [14], that $\bar{f}(t, \omega)$ is measurable and adapted to $(\mathscr{F}^0_t)$. Now let $\mathscr{F}^{x^0}_t$ denote the σ-algebra generated by $x^0(s)$, $0 \leqq s \leqq t$, and let $\bar{\bar{f}}(t, \omega)$ be a measurable version of $E\{\bar{f}(t, \cdot)/\mathscr{F}^{x^0}_t\}$. Since (3.4) can be written

$$x^0(t) = x^0(0) + \int_0^t \bar{f}(s, \omega)\,ds + \int_0^t \sigma(s, x^0)\,dw^0(s), \tag{3.6}$$

by Theorem 4.3 in [18] and Lemma 10.4 in [14] there exists a standard Brownian motion $w(t)$ with respect to which $x^0(t)$ is nonanticipating such that

$$x^0(t) = x^0(0) + \int_0^t \bar{\bar{f}}(s, \omega)\,ds + \int_0^t \sigma(s, x^0)\,dw(s) \tag{3.7}$$

holds for all t with probability one. Since the process $\bar{\bar{f}}(t, \omega)$ is adapted to $(\mathscr{F}^{x^0}_t)$, this means that $x^0(t)$ is a process of diffusion type in the sense of [14]. Hence, according to the multidimensional version of Theorem 7.20 in [14], μ^0 is absolutely continuous with respect to the measure induced on C^r by the solution of

$$dx_t = \sigma(t, x)\, dw_t,$$
$$x_0 = x^0(0),$$

which is just μ.

To go on, we need a mild regularity assumption about the observation functions φ_j.

(A9) $\varphi_j^{-1}(A)$ has Lebesgue measure 0 in $\mathbb{R}^{r_j}$ for every set A of Lebesgue measure 0, $j = 1, \cdots, p$.

To simplify the notation let us introduce the processes

$$y_j^n = \varphi_j(x^n(t_1), \cdots, x^n(t_j)), \qquad n = 0, 1, \cdots,$$

which take on values in $\mathbb{R}^{k_j}$. Then, as an immediate consequence of (A4), (A9) and Lemma 1, we find

COROLLARY. *The measure induced on $\mathbb{R}^{k_j}$ by y_j^0 is absolutely continuous with respect to the k_j-dimensional Lebesgue measure.*

LEMMA 2. *There exists an admissible control u^0 such that*

$$F^0(t) = \int_0^t g(s, x^0) h(s, u^0(s))\, ds$$

and

$$\int_0^t k(s, x^n) l(s, u^n(s))\, ds \to \int_0^t k(s, x^0) l(s, u^0(s))\, ds$$

hold for all t with probability one.

Proof. Denote $R^n(t) = (H^n(t), L^n(t))$, $n = 0, 1, \cdots$, and $r(t, u) = (h(t, u), l(t, u))$. Then, for $t \in [t_j, t_{j+1})$, $n = 1, 2, \cdots$,

$$\begin{aligned} R^n(t) - R^n(t_j) &= \int_{t_j}^t r(s, u^n(s))\, ds \\ &= \sum \int_{t_j}^t r(s, u^n_{j;\mu_1, \cdots, \mu_{k_j}}(s))\, ds \cdot \chi_{A^n_{j;\mu_1, \cdots, \mu_{k_j}}}, \end{aligned} \tag{3.8}$$

where

$$A^n_{j;\mu_1, \cdots, \mu_{k_j}} = \{y^n_{ji} \in (a^n_{j,i,\mu_i}, a^n_{j,i,\mu_i+1}], \quad i = 1, \cdots, k_j\},$$

χ_A denotes the indicator of A and the sum is over all combinations $(\mu_1, \cdots, \mu_{k_j})$ with $1 \leqq \mu_i \leqq \nu_{ij}$. Note that (by passing to a further subsequence) we may assume that the ν_{ij} and hence the number of terms under the sum are the same for all n and, moreover, that

$$a^n_{j,i,\mu_i} \to a^0_{j,i,\mu_i},$$

for all $j = 1, \cdots, p$; $i = 1, \cdots, k_j$, $\mu_i = 1, \cdots, \nu_{ij}$. Let us show that

$$\chi_{A^n_{j;\mu_1, \cdots, \mu_{k_j}}} \to \chi_{A^0_{j;\mu_1, \cdots, \mu_{k_j}}}, \qquad P\text{-a.e.} \tag{3.9}$$

To this end, forget the indices for a moment and write

$$A^n = \{y^n \in (a^n, b^n]\},$$

with $a^n \to a^0$, $b^n \to b^0$. Since $y^n \to y^0$ for all ω (remove the exceptional set),

$$\limsup_{n \to \infty} A^n \subset \{y^0 \in [a^0, b^0]\}. \tag{3.10}$$

On the other hand,

$$\{y^0 \in [a^0+\varepsilon, b^0-\varepsilon]\} \subset \liminf_{n\to\infty} A^n,$$

for all $\varepsilon > 0$, hence

$$\bigcup_{\varepsilon>0} \{y^0 \in [a^0+\varepsilon, b^0-\varepsilon]\} = \{y^0 \in (a^0, b^0)\} \subset \liminf_{n\to\infty} A^n. \tag{3.11}$$

By the corollary to Lemma 1, $\{y^0 \in [a^0, b^0]\}$ and $\{y^0 \in (a^0, b^0)\}$ differ only by a P-nullset from $\{y^0 \in (a^0, b^0]\}$; hence, from (3.10), (3.11) and

$$\chi^n_{A_{j;\mu_1,\cdots,\mu_{k_j}}} = \prod_{i=1}^{k_j} \chi_{\{y^n_{ji} \in (a^n_{j,i,\mu_i}, a^n_{j,i,\mu_{i+1}}]\}},$$

(3.9) follows. From this and (3.8) it follows that for $t \in [t_j, t_{j+1})$,

$$\begin{aligned} R^0(t) - R^0(t_j) &= \lim_{n\to\infty} [R^n(t) - R^n(t_j)] \\ &= \lim_{n\to\infty} \int_{t_j}^t r(s, u^n_{j;\mu_1,\cdots,\mu_{k_j}}(s))\,ds, \end{aligned} \tag{3.12}$$

P-a.e. on $A^0_{j;\mu_1,\cdots,\mu_{k_j}}$. Using standard methods from optimal control theory (cf. [13], Chapter 4) it can be deduced from (3.12) and (A4) that there exist measurable functions $u^0_{j;\mu_1,\cdots,\mu_{k_j}}(\cdot)$ defined on $[t_j, t_{j+1}]$ taking a.e. values in $\mathcal{U}$ such that

$$R^0(t) - R^0(t_j) = \int_{t_j}^t r(s, u^0_{j;\mu_1,\cdots,\mu_{k_j}}(s))\,ds,$$

P-a.e. on $A^0_{j;\mu_1,\cdots,\mu_{k_j}}$. Define the control $u^0 : [0, T] \times C^r \to \mathcal{U}$ by

$$u^0(t, \omega) = u^0_{j;\mu_1,\cdots,\mu_{k_j}}(t), \quad \text{for } t \in [t_j, t_{j+1}) \quad \text{and} \quad \omega \in A^0_{j;\mu_1,\cdots,\mu_{k_j}}.$$

Then

$$R^0(t) = \int_0^t r(s, u^0(s))\,ds = \lim_{n\to\infty} \int_0^t r(s, u^n(s))\,ds, \tag{3.13}$$

for all t with probability one. The second equality implies that for a.e. ω,

$$r(\cdot, u^n(\cdot, \omega)) \to r(\cdot, u^0(\cdot, \omega)),$$

weakly in $L_1[0, 1]$. From this

$$\begin{aligned} \left| F^n(t) - \int_0^t g(s, x^0) h(s, u^0(s))\,ds \right| &\leqq \left| \int_0^t g(s, x^0)[h(s, u^n(s)) - h(s, u^0(s))]\,ds \right. \\ &\quad \left. + \int_0^t |h(s, u^n(s))|\,|g(s, x^n) - g(s, x^0)| \right| ds \\ &\to 0, \quad \text{a.e. for all } t \end{aligned}$$

since for a fixed ω the $g(t, x^n(\omega))$ are uniformly bounded by virtue of (A1). Hence

$$F^0(t) = \int_0^t g(s, x^0) h(s, u^0(s))\,ds, \tag{3.14}$$

for all t with probability one, proving that u^0 is an admissible control with corresponding solution x^0. The relation

$$I^n(t) := \int_0^t k(s, x^n) l(s, u^n(s))\, ds \to \int_0^t k(s, x^0) l(s, u^0(s))\, ds =: I^0(t)$$

is proved in the same way.

Suppose now that the sequence (u^n, x^n) is minimizing, i.e.,

$$J(x^n, u^n) \to \hat{J}.$$

Since the $I^n(t)$ are nonnegative and $m(x^n) \to m(x^0)$ a.e.,

$$E\{I^0(1) + m(x^0)\} \leqq \liminf_{n\to\infty} E\{I^n(1) + m(x^n)\} \tag{3.15}$$

by Fatou's lemma, thus showing that

$$J(x^0, u^0) = \hat{J}.$$

Hence we have proved

THEOREM 1. *Under assumptions* (A1)–(A8) *problem* (P) *has an optimal solution.*

If $h(t, u)$ and $l(t, u)$ are continuous in both variables assumption (A4) can be relaxed to

(A4′) For every $t \in [0, 1]$, the set

$$\mathcal{V}^+(t) = \{\tilde{z} = (z, z_{r+1}): z = h(t, u),\ z_{r+1} \geqq l(t, u),\ u \in \mathcal{U}\}$$

is compact and convex.

The proof of Theorem 1 can then be carried out in basically the same way as above. Instead of using the standard methods from deterministic control theory in [13, § 4], we proceed as in [11, chapter III.5 (especially Lemmas 5.4 to 5.6)] to obtain a control u^0 such that

$$H^0(t) = \int_0^t h(s, u^0(s))\, ds = \lim_{n\to\infty} \int_0^t h(s, u^n(s))\, ds, \tag{3.16}$$

and

$$L^0(t) = \int_0^t [l(s, u^0(s)) + v(s)]\, ds = \lim_{n\to\infty} \int_0^t l(t, u^n(s))\, ds, \tag{3.17}$$

for all t a.e. with some nonnegative function v. These two relations replace (3.13). As above, it follows from the weak convergence of the integrands in (3.16) and the boundedness of the $g(t, x^n)$ that (3.14) holds. With $I^n(t)$, $n = 0, 1, \cdots$, defined as above,

$$\begin{aligned} I^n(t) - I^0(t) = \int_0^t & l(s, u^n(s))[k(s, x^n) - k(s, x^0)]\, ds \\ &+ \int_0^t k(s, x^0)[l(s, u^n(s)) - l(s, u^0(s)) - v(s)]\, ds \\ &+ \int_0^t k(s, x^0) v(s)\, ds. \end{aligned}$$

The first two integrals tend to 0, hence

$$I^0(t) \leqq \lim_{n\to\infty} I^n(t),$$

and, by Fatou's lemma, (3.15) results.

So far we have excluded $x(0)$ from our observations. This is no loss if the initial distribution F_0 is a point mass in $\mathbb{R}^r$. If only some components of $x(0)$—say the first k—are degenerate while the other r–k have a distribution which is absolutely continuous with respect to $(r$–$k)$ dimensional Lebesgue measure, then we may allow the φ_j to depend also on these nondegenerate components.

Remark 1. Suppose that the decision regions are fixed in advance, i.e., the numbers a_{j,i,μ_i} in (i) are the same for all admissible controls. This may be interpreted by saying that not the actual value of the y_{ji} can be observed, but only which decision region is hit. Let $\mathscr{A}\{a_{j,i,\mu_i}, j=1,\cdots,p, i=1,\cdots,k_j, \mu_i=1,\cdots,\nu_{ji}\}$ denote the corresponding class of admissible controls. Then it is easy to see that the proof of Theorem 1 remains valid if $\mathscr{A}$ is replaced by $\mathscr{A}\{\cdot\}$.

4. A different class of controls. In this section we shall consider the same system (1.1)–(1.3) but with a different class of admissible controls. A measurable function $u:[0,1]\times\Omega\rightarrow\mathscr{U}$ defined on some probability space $(\Omega, \mathscr{F}, P)$ will now be called an admissible control if there exists a measurable process $x(t)$ on $(\Omega, \mathscr{F}, P)$ such that the following conditions (i)–(iii) are satisfied.

(i) For every $t\in[t_j, t_{j+1})$, $j=1,\cdots,p-1$, $u(t,\cdot)$ is measurable with respect to the completed σ-algebra

$$\mathscr{Y}_j = \sigma\{y_j\},$$

where $y_j=\varphi_j(x(t_1),\cdots,x(t_j))$ as above.

(ii) and (iii): same as in § 2.

(i) is equivalent to saying that for $t\in[t_j, t_{j+1})$, u has the form

$$u(t,\omega)=\mathbf{u}_j(t, \varphi_j(x(t_1,\omega),\cdots,x(t_j,\omega))), \quad \text{a.e.},$$

for some function $\mathbf{u}_j:[0,1]\times\mathbb{R}^{k_j}\rightarrow\mathscr{U}$.

Putting it in still another way which will turn out useful, let $\mathscr{G}_j$ denote the σ-algebra on C^r generated by the function $\xi\rightarrow\varphi_j(\xi(t_1),\cdots,\xi(t_j))$. Then (i) is equivalent to

(i') There is a function $\mathbf{u}:[0,1]\times C^r\rightarrow\mathscr{U}$ such that for each j $\mathbf{u}|_{[t_j,t_j+1)\times C^r}$ is measurable with respect to $\mathscr{B}_{[t_j,t_{j+1})}\otimes\mathscr{G}_j$ and for each $t\in[0,1]$,

$$u(t,\omega)=\mathbf{u}(t,x(\omega)), \quad \text{a.e.}$$

Controls of this form have been considered in [10], where the functions u_j are assumed to be Lipschitz in all variables. Note that the σ-algebras $\mathscr{Y}_j$ need not be an increasing family, thus reflecting the fact that past observations may get lost.

Let $\mathscr{A}^0$ denote the class of admissible controls and let (P^0) denote the problem posed at the end of § 2 with $\mathscr{A}$ replaced by $\mathscr{A}^0$. We shall use the results obtained in § 3 to approximate controls in $\mathscr{A}^0$ by controls with a finite number of decision regions. For simplicity, let us confine the following discussions to the case $k_j=1$, for all $j=1,\cdots,p$, the general case following the same line of argument.

Choose a sequence of partitions,

$$-\infty=a_1^n<\cdots<a_n^n=\infty,$$

of the real axis such that $\{a_1^n, \cdots, a_n^n\} \subset \{a_1^{n+1}, \cdots, a_{n+1}^{n+1}\}$, and the (increasing) fields $\mathcal{B}^n$ generated by the sets $(a_i^n, a_{i+1}^n]$, $i = 1, \cdots, n-1$, generate the Borel σ-field on the line, i.e.,

$$\mathcal{B} = \bigvee_{n=1}^{\infty} \mathcal{B}^n.$$

Let (P^n) denote the problem posed in § 2 with fixed decision regions given by the same partition $\{a_1^n, \cdots, a_n^n\}$, for all j, i.e., in the notation of Remark 1 the class of admissible controls is $\mathcal{A}^n = \mathcal{A}\{a_1^n, \cdots, a_n^n\}$.

Remark 2. Let $\eta_j : C^r \to \mathbb{R}$ denote the mapping

$$\eta_j(\xi) = \varphi_j(\xi(t_1), \cdots, \xi(t_j)),$$

and define fields

$$\mathcal{G}_j^n = \eta_j^{-1}(\mathcal{B}^n),$$

$j = 1, \cdots, p$. Then, for $u^n \in \mathcal{A}^n$, property (i) in § 2 is equivalent to saying that there exists a function $\mathbf{u}^n : [0, 1] \times C^r \to \mathcal{U}$ such that $\mathbf{u}^n|_{[t_i, t_{i+1}) \times C^r}$ is measurable with respect to $\mathcal{B}_{[t_j, t_{j+1})} \otimes \mathcal{G}_j^n$ and

$$u^n(t, \omega) = u^n(t, x(\omega)),$$

for all (t, ω). Moreover, it is easily verified that

$$\mathcal{G}_j = \bigvee_{n=1}^{\infty} \mathcal{G}_j^n.$$

In the following, let us assume the validity of all the assumptions (A1)–(A9). Then, according to Remark 1, each problem (P^n), $n = 1, 2, \cdots$, possesses an optimal solution $(\bar{u}^n, \bar{x}^n)$ defined on some probability space $(\Omega^n, \mathcal{F}^n, P^n)$. Define processes $\Phi^n = (\bar{x}^n, F^n, B^n, H^n, L^n)$ as in § 2 and observe that the family (Φ^n) is tight (meaning that the family of probability measures $P_{\Phi^n}^n$ induced on the sample space S is tight). Hence, by Skorokhod's theorem, we may assume (after passing to a subsequence, which will be denoted by the same index n) that the Φ^n are all defined on the same probability space $(\Omega, \mathcal{F}, P) = ([0, 1], \mathcal{B}, \text{Lebesgue measure})$ and that

$$\Phi^n \to \Phi^0, \qquad \text{a.e.},$$

in S for some measurable process $\Phi^0 = (\bar{x}^0, F^0, B^0, H^0, L^0)$. As in § 3, it can be shown that the representation (3.6) is valid, and hence x^0 is a process of diffusion type and Corollary 1 applies. Again, denote

$$R^n(t) = (H^n(t), L^n(t)), \qquad \Delta_j^n(t) = R^n(t) - R^n(t_j), \qquad y_j^n = \eta_j(\bar{x}^n(\cdot)), \qquad n = 0, 1, \cdots.$$

The basic result is:

LEMMA 3. *For* $t \in [t_j, t_{j+1})$, $\Delta^0(t)$ *is measurable with respect to the completed* σ*-algebra generated by* y_j^0.

Proof. It suffices to show the assertion for the components of $\Delta^0(t)$; hence, let us suppose that $\Delta^0(t)$ is scalar valued. Denote

$$\mathcal{Y}_j^n = (y_j^n)^{-1}(\mathcal{B}^n) = (\bar{x}^n)^{-1}(\mathcal{G}_j^n), \qquad n = 1, 2, \cdots,$$

and let $\mathcal{Y}_j^0$ denote the completion with respect to P of

$$(y_j^0)^{-1}(\mathcal{B}) = (\bar{x}^0)^{-1}(\mathcal{G}_j).$$

Then $\Delta^n(t)$ is $\mathscr{Y}_j^n$-measurable and, since $\Delta^0(t)=\lim_{n\to\infty}\Delta^n(t)$ (remove the nullset of no· convergence), for every $\lambda\in R$,

$$\{\Delta^0(t)<\lambda\}=\bigcup_{k=1}^{\infty}\bigcup_{N=1}^{\infty}\bigcap_{n=N}^{\infty}\left\{\Delta^n(t)<\lambda-\frac{1}{k}\right\}.$$

This implies that $\Delta^0(t)$ is measurable with respect to the σ-algebra $\mathscr{S}$ generated by the sets

$$\bigcup_{N=1}^{\infty}\bigcap_{n=N}^{\infty}\{\Delta^n(t)<\mu\},$$

with μ ranging in $\mathbb{R}$. But

$$\{\Delta^n(t)<\mu\}=\{y_j^n\in A^n\},\quad \text{for some } A^n\in\mathscr{B}^n;$$

hence,

$$\mathscr{S}=\sigma\left\{\bigcup_{N=1}^{\infty}\bigcap_{n=N}^{\infty}\{y_j^n\in a^n\},\quad A^n\in\mathscr{B}^n\right\}. \tag{4.1}$$

Now, for every n, A^n is a finite union of disjoint intervals of the form $(a,b]$ with endpoints in $\{a_1^n,\cdots,a_n^n\}$. Let

$$A^n=\bigcup_{(\underline{i},\bar{\imath})\in I_n}(a^n_{\underline{i}},{}^n_{\bar{\imath}}],\qquad I_n\subset\{1,2,\cdots,n\}^2,$$

be the (unique) representation with the smallest possible number of intervals. It is easily checked that for a fixed sequence (A^n),

$$\bigcup_{N=1}^{\infty}\bigcap_{n=N}^{\infty}\{y_j^n\in A^n\}=\bigcup_{(\underline{i}_1,\bar{\imath}_1,\underline{i}_2,\bar{\imath}_2,\cdots)\in\prod_{j=1}^{\infty}I_j}\bigcup_{N=1}^{\infty}\bigcap_{n=N}^{\infty}\{y_j^n\in(a^n_{\underline{i}_n},a^n_{\bar{\imath}_n}]\}. \tag{4.2}$$

For a fixed sequence $\sigma=(\underline{i}_1,\bar{\imath}_1,\cdots)\in\prod_{j=1}^{\infty}I_j$ denote

$$S_\sigma=\bigcup_{N=1}^{\infty}\bigcap_{n=N}^{\infty}\{y_j^n\in(a^n_{\underline{i}_n},a^n_{\bar{\imath}_n}]\}. \tag{4.3}$$

Then, by the same sort of argument used in the proof of Lemma 2,

$$S_\sigma=\{y_j^0\in(\underline{a},\bar{a}]\},\qquad P\text{-a.e.}, \tag{4.4}$$

where $\underline{a}=\limsup_{n\to\infty}a^n_{\underline{i}_n}$, $\bar{a}=\liminf_{n\to\infty}a^n_{\bar{\imath}_n}$. Hence

$$S_\sigma\in\mathscr{Y}_j^0,$$

for every sequence $\sigma\in\prod_{j=1}^{\infty}I_j$. But

$$\bigcup_{\sigma\in\prod_{j=1}^{\infty}I_j}S_\sigma$$

is obtained from $\mathscr{Y}_j^0$ by a Suslin operation (cf. [15], [17]). To see this, let $\mathscr{V}$ and $\mathscr{V}^*$ denote the respective sets of infinite and finite sequences of positive integers (with generic elements σ and σ^*). Observe that each sequence $\sigma=(\underline{i}_1,\bar{\imath}_1,\underline{i}_2,\bar{\imath}_2,\cdots)\in\prod_{j=1}^{\infty}I_j$ uniquely determines numbers

$$\underline{a}=\limsup_{n\to\infty}a^n_{\underline{i}_n},\qquad \bar{a}=\liminf_{n\to\infty}a^n_{\bar{\imath}_n},$$

such that (4.4) holds, thus determining a mapping $l:\prod_{j=1}^{\infty}I_j\to\mathscr{V}$ by

$$l(\sigma)=(\underline{a},\bar{a},\underline{a},\bar{a},\cdots).$$

Define a mapping $\mathscr{A}: \mathscr{V}^* \to \mathscr{Y}_j^0$ by putting

$$\mathscr{A}(\sigma^*) = \begin{cases} \Omega & \text{if } \sigma^* = (\sigma_1^*) \text{ (one element sequence);} \\ S_\sigma & \text{if } \sigma^* = l(\sigma)|n \text{ for } n \geqq 2 \text{ and some } \sigma \in \prod_{j=1}^{\infty} I_j; \\ \varnothing & \text{elsewhere.} \end{cases}$$

Here, for $\sigma \in \mathscr{V}^*$, $\sigma|n = (\sigma_1, \cdots, \sigma_n) \in \mathscr{V}$. Then

$$\bigcup_{\sigma \in \prod_{j=1}^{\infty} I_j} S_\sigma = \bigcup_{\sigma \in \mathscr{V}} \bigcap_{n=1}^{\infty} \mathscr{A}(\sigma|n).$$

Since $\mathscr{Y}_j^0$ is complete, it is a Suslin family (cf. [15]), hence

$$\bigcup_{\sigma \in \prod_{j=1}^{\infty} I_j} S_\sigma \in \mathscr{Y}_j^0.$$

Then it follows from (4.1)–(4.4) that

$$\mathscr{S} \subset \mathscr{Y}_j^0,$$

thus proving the assertion.

Let $r(t, u)$ be defined as in § 3. Then from

$$R^n(t) = \int_0^t f(s, \bar{u}^n(s))\, ds \to R^0(t), \qquad \text{a.e.},$$

it follows that there is an integrable function $\bar{r}(t, \omega)$ with $\bar{r}(t, \omega) \in \mathscr{V}(t) = r(t, \mathscr{U})$ for almost all (t, ω) such that

$$R^0(t) = \int_0^t \bar{r}(s, \omega)\, ds,$$

for all t with probability one (cf. [12]). Since $R^0(t) - R^0(t_j)$ is measurable with respect to $\mathscr{Y}_j^0$ for $t \in [t_j, t_{j+1})$, $\bar{r}(t, \omega)$ can be chosen in such a way that $\bar{r}(t, \cdot)$ is $\mathscr{Y}_j^0$-measurable for $t \in [t_j, t_{j+1})$ (cf. [14, Lemma 5.2]). It may then be assumed that $\bar{r}|_{[t_j, t_{j+1}) \times \Omega}$ is measurable with respect to $\mathscr{B}_{[t_j, t_{j+1})} \otimes \mathscr{Y}_j^0$. Modify $\bar{r}$ on a nullset such that $\bar{r}(t, \omega) \in \mathscr{V}(t)$ holds for all (t, ω). Then, by Lemma 5 in [2], there exists a function $\bar{u}_j^0 : [t_j, t_{j+1}) \times \Omega \to \mathscr{U}$ which is measurable with respect to $\mathscr{B}_{[t_j, t_{j+1})} \otimes \mathscr{Y}_j^0$ such that

$$\bar{r}(t, \omega) = r(t, \bar{u}_j^0(t, \omega)),$$

for almost all $(t, \omega) \in [t_j, t_{j+1}) \times \Omega$.

Define $\bar{u}^0 : [0, 1] \times \Omega \to \mathscr{U}$ by

$$\bar{u}^0|_{[t_j, t_{j+1}) \times \Omega} = \bar{u}_j^0.$$

Then

$$R^0(t) = \int_0^t r(s, \bar{u}^0(s))\, ds = \lim_{n \to \infty} \int_0^t r(s, \bar{u}^n(s))\, ds,$$

for all t with probability one.

In the same way as in § 3 it can now be shown that (3.14) and (3.15) hold. Hence $\bar{u}^0$ is an admissible control for (P^0) with corresponding solution $\bar{x}^0$, and

$$J(\bar{x}^0, \bar{u}^0) \leqq \liminf_{n \to \infty} J(\bar{x}^n, \bar{u}^n).$$

Let us now introduce the following

HYPOTHESIS. *For every $u \in \mathscr{A}^0$ with corresponding solution x there exists a sequence of controls $u^n \in \mathscr{A}^n$ with corresponding solutions x^n such that*

$$J(x^n, u^n) \to J(x, u).$$

THEOREM 2. *Assume* (A1)–(A9) *and suppose that the hypothesis is valid. Then problem* (P^0) *possesses an optimal solution.*

Proof. Construct $(\bar{x}^0, \bar{u}^0)$ as above. Let u be any control in $\mathscr{A}^0$ with corresponding solution x, and let (x^n, u^n) be the sequence from the hypothesis. Recall that the $(\bar{x}^n, \bar{u}^n)$ were optimal for (P^n). Hence

$$J(\bar{x}^0, \bar{u}^0) \leqq \liminf_{n\to\infty} J(\bar{x}^n, \bar{u}^n) \leqq \lim_{n\to\infty} J(x^n, u^n) = J(x, u).$$

Of course, Theorem 2 is useful only if reasonable sufficient conditions for the hypothesis to hold can be given. We shall show that for a wide class of problems the hypothesis is actually satisfied. These problems are essentially those where the Girsanov measure transformation method for defining solutions of (1.1) works.

To see this, we introduce the following assumption:

(A10) σ^{-1} is bounded, and there is a constant K such that for all $\xi \in C^r$ and all $t \in [0, 1]$,

$$|k(t, \xi)| + |g(t, \xi)| \leqq K(1 + \|\xi\|_t).$$

Let $u(t, x)$ be an admissible control in $\mathscr{A}^0$ with corresponding solution x (according to (i′) we shall work with the control functions defined on $[0, 1] \times C^r$ from now on). It will turn out convenient to introduce the notations

$$f_t^u(x) = g(t, x)h(t, u(t, x)), \qquad c_t^u(x) = k(t, x)l(t, u(t, x))$$

and

$$\sigma_t(x) = \sigma(t, x), \qquad \sigma_t^{-1}(x) = \sigma^{-1}(t, x).$$

Then (1.1) takes the form

$$dx = f_t^u(x)\, dt + \sigma_t(x)\, dw. \tag{4.5}$$

LEMMA 4. *Let u be an admissible control with corresponding solution x. Then, under assumptions* (A1)–(A2), (A5) *and* (A10), *for every positive integer m,*

$$E\|x\|_t^{2m} \leqq C_m(1 + e^{C_m t}),$$

for all $t \in [0, 1]$, with the same constant C_m for all admissible controls.

Proof. Define

$$\chi_N(t) = \begin{cases} 1, & \text{if } \|x\|_t \leqq N, \\ 0, & \text{if } \|x\|_t > N. \end{cases}$$

Then, since $\chi_N(t) = \chi_N(t)\chi_N(r)$, for $r \leqq t$,

$$x(s)\chi_N(t) = \int_0^s \chi_N(r) f_r^u(x)\, dr + \int_0^s \chi_N(r)\sigma_r(x)\, dw(r),$$

for all $s \leqq t$. Hence

$$|x(s)|^{2m}\chi_N(t) \leqq \text{const} \cdot \left[\int_0^s \chi_N(r)(1 + \|x\|_r^{2m})\, dr + \left| \int_0^s \chi_N(r)\sigma_r(x)\, dw(r) \right|^{2m} \right],$$

where the constant is independent of the particular control chosen (it may depend on m). It follows that

$$\|x\|_t^{2m}\chi_N(t)\text{const}\cdot\left[\int_0^t \chi_N(r)(1+\|x\|_r^{2m})\,dr+\sup_{0\leqq s\leqq t}\left|\int_0^s \chi_N(r)\sigma_r(x)\,dw(r)\right|^{2m}\right],$$

and

$$E\|x\|_t^{2m}\chi_N(t)\leqq\text{const}\cdot\left[\int_0^t (1+E\|x\|_r^{2m}\chi_N(r))\,dr+\int_0^t (1+E\|x\|_r^{2m}\chi_N(r))\,dr\right].$$

The last inequality follows from the estimate

$$E\sup_{0\leqq s\leqq t}\left|\int_0^s \phi(r)\,\mathrm{d}w(r)\right|^{2m}\leqq\left(\frac{2m}{2m-1}\right)^{2m}E\left|\int_0^t \phi(r)\,dw(r)\right|^{2m}=\text{const}\cdot\int_0^t E\phi^{2m}(r)\,dr,$$

which holds for every nonanticipating bounded functional ϕ. By the Gronwall-Bellman inequality,

$$E\|x\|_t^{2m}\chi_N(t)\leqq C_m(1+e^{C_m t}),$$

and Fatou's lemma (for $N\to\infty$) accomplishes the proof.

Note that by virtue of Lemma 4, for every admissible control u with corresponding solution x the tightness conditions (iii) in § 2 are automatically satisfied.

PROPOSITION 1. *Under assumptions* (A1)–(A10) *the hypothesis holds, hence problem* (P^0) *has an optimal solution.*

Proof. Let $u:[0,1]\times C^r\to\mathcal{U}$ be an admissible control with corresponding solution $\tilde{x}$ (cf. (i′)). Let x be the unique solution of (2.1), defined on some probability space $(\Omega,\mathcal{F},P)$ carrying a Brownian motion (B_t). Then x is a solution of (4.5) for the probability measure

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

where

$$\zeta(f^u)=\exp\left[\int_0^1(\sigma_t^{-1}f_t^u)'\,dB_t-\tfrac{1}{2}\int_0^1|\sigma_t^{-1}f_t^u|^2\,dt\right],$$

and the Brownian motion in (4.5) is given by

$$\begin{aligned}dw_t&=dB_t-\sigma_t^{-1}(x)f_t^u(x)\,dt\\&=\sigma_t^{-1}(x)[dx_t-f_t^u(x)\,dt].\end{aligned}$$

For details cf. [6], [8]. Hence x is a solution corresponding to u. Let E denote expectation with respect to μ, the measure induced on C^r by x under P, and define controls u^n by

$$u^n(t)=E\{u(t)/\mathcal{G}_j^n\},\quad\text{for } t\in[t_j,t_{j+1}),$$

(take a measurable version). Then x is a solution of (4.5) with u replaced by u^n with respect to the probability measure

$$dP^n=\zeta(f^{u^n})\,dP,$$

and the Brownian motion

$$dw_t^n=\sigma_t^{-1}(x)[dx_t-f_t^{u^n}(x)\,dt].$$

Hence $u^n \in \mathcal{A}^n$ with corresponding solution x. Now, for fixed $t \in [t_j, t_{j+1})$, $(u^n(t), \mathcal{G}_j^n)$ is a bounded martingale, and

$$u^n(t) \to E\left\{u(t) \Big/ \bigvee_{n=1}^{\infty} \mathcal{G}_j^n\right\} = E\{u(t)/\mathcal{G}_j\}$$
$$= u(t), \qquad \mu\text{-a.e.},$$

(cf. [7], VII, Thm 4.3). It follows that for all t

$$c_t^{u^n}(x) \to c_t^u(x), \qquad P\text{-a.e.},$$

and

$$f_t^{u^n}(x) \to f_t^u(x), \qquad P\text{-a.e.}$$

By (A1), (A2), (A10) and Lemma 4, the $|\sigma_t^{-1} f_t^{u^n}|^2$ are uniformly integrable with respect to $dt \times dP$ and P-a.e., bounded uniformly in n and t; hence, the last relation implies

$$E \int_0^1 |\sigma_t^{-1}(f_t^{u^n} - f_t^u)|^2 \, dt \to 0$$

and

$$\int_0^1 |\sigma_t^{-1} f_t^{u^n}|^2 \, dt \to \int_0^1 |\sigma_t^{-1} f_t^u|^2 \, dt, \qquad P\text{-a.e.},$$

from which it follows that for a subsequence (n')

$$\zeta(f^{u^{n'}}) \to \zeta(f^u), \qquad P\text{-a.e.}$$

Using Lemma 1 in [2] (or rather the extension to the case $\sigma \neq I$, as it is used in [5] and [6]), (A10), Lemma 4 and a Hölder estimate, it can be shown that the functions $c_t^{u^{n'}} \zeta(f^{u^{n'}})$ are uniformly integrable with respect to $dt \times dP$; consequently,

$$\begin{aligned} J(x, u^{n'}) &= E^{n'}\{m(x)\} + E^{n'}\left\{\int_0^1 c_t^{u^{n'}} \, dt\right\} \\ &= E\{m(x)\zeta(f^{u^{n'}})\} + E\left\{\int_0^1 c_t^{u^{n'}} \zeta(f^{u^{n'}}) \, dt\right\} \\ &\to E\{m(x)\zeta(f^u)\} + E\left\{\int_0^1 c_t^u \zeta(f^u) \, dt\right\} \\ &= E^u\{m(x)\} + E^u\left\{\int_0^1 c_t^u \, dt\right\} = J(x, u), \end{aligned}$$

E^n, E^u and E denoting integration with respect to P^n, P^u and P, respectively. The assertion then follows from Lemma 5 below.

LEMMA 5. *Under the assumptions of Proposition* 1 *the solution of* (1.1), (1.2) *is unique in law.*

Proof. The proof can be copied almost literally from the proof of Theorem 10.3, chapter V, in [11], noting that the Girsanov formula is valid under our assumptions and using an approximating sequence of step functions instead of the Riemann-Stieltjes sums. Moreover, for the special case $\sigma = I$ (identity matrix), the result can be found in Theorem 4.12 in [14].

5. Concluding remarks. So far we have assumed that the times at which observations can be taken are fixed in advance. A natural question to ask is: What will happen if not only the net of decision (or rather observation) regions but also the net of observation times is getting finer and finer? Certainly the class of controls that can be approximated in this way will contain the class $\mathscr{A}^0$. The question is what this wider class looks like and whether it contains an optimal solution.

What one would like is to admit as controls all measurable functions $u:[0,1]\times C^r \to \mathscr{U}$ which are adapted to $(\mathscr{G}_t)$, where $\mathscr{G}_t$ is the σ-field on C^r generated by the past of the observable component η of $\xi=(\zeta,\eta)$ (ξ the generic element of C^r), i.e., if l is the dimension of η,

$$\mathscr{G}_t=\sigma\{[\xi:\eta(s)\in A], \text{a Borel set in } \mathbb{R}^l,\quad 0\leqq s\leqq t\}.$$

Denote this class of admissible controls by $\mathscr{A}^*$ and the corresponding control problem by (P*). Then, under the assumptions of Proposition 1, every control in $\mathscr{A}^*$ determines a unique (in law) solution of (1.1), (1.2) such that the tightness conditions (iii) in § 2 hold. Taking sequences $\mathscr{I}^n=\{t_1^n,\cdots,t_n^n\}$ and $\mathscr{A}^n=\{a_1^n,\cdots,a_n^n\}$ of refining partitions of [0, 1] and the real axis, respectively, we get a sequence of problems (P^n) of the type discussed in §§ 2 and 3, with observation times $\mathscr{I}^n$, observation functions

$$\varphi_j^n(\xi)=(\eta(t_1^n),\cdots,\eta(t_j^n)),\qquad j=1,\cdots,n,\qquad \xi=(\zeta,\eta),$$

and decision regions given by $\mathscr{A}^n$ for all components of φ_j^n and all $j=1,\cdots,n$. As in § 4, let $\mathscr{B}^n$ denote the algebra generated by $\mathscr{A}^n$ on the real axis and $(\mathscr{B}^n)^{lj}$, its lj-fold product, and let $\mathscr{G}_t^n$, $t\in[t_j,t_{j+1})$, be defined by

$$\mathscr{G}_t^n=(\varphi_j^n)^{-1}((\mathscr{B}^n)^{lj}).$$

If the $\mathscr{I}^n$ and $\mathscr{A}^n$ are chosen such that $\bigcup\mathscr{I}^n$ and $\bigcup\mathscr{A}^n$ are dense in [0, 1] and R, respectively,

$$\mathscr{G}_t=\bigvee_{n=1}^{\infty}\mathscr{G}_t^n,$$

and the proof of Proposition 1 can be copied to show that the hypotheses remains valid for controls in $\mathscr{A}^*$.

Next, under the assumptions of Proposition 1, every problem (P^n) possesses an optimal solution $(\bar{x}^n,\bar{u}^n)$ and, by tightness, we arrive at processes $\Phi^n=(\bar{x}^n,F^n,B^n,H^n,L^n)$, $n=0,1,\cdots$, such that $\Phi^n\to\Phi^0$ a.e. and $\bar{x}^0$ is a process of diffusion type, (3.6) holds and Corollary 1 applies. With $R^0(t)$ defined as in § 4, the crucial point is to show that the assertion of Lemma 4 holds, i.e., that $R^0(t)$ is measurable with respect to $(\bar{x}^0)^{-1}(\mathscr{G}_t)$. Here it turns out that the mechanism of § 4 does not continue to work, basically because now an infinite number of observation times has to be considered and no infinite dimensional counterpart of the left-hand inclusion in (3.9) seems to be available.

REFERENCES

[1] V. E. Benes, *Existence of optimal strategies based on specified information, for a class of stochastic decision problems*, this Journal, 8 (1970), pp. 179–188.

[2] ———, *Existence of optimal stochastic control laws*, Ibid. 9 (1971), pp. 446–472.

[3] P. Billingsley, *Convergence of Probability Measures*, Wiley, New York, 1968.

[4] N. Christopeit, *Existence of Optimal Stochastic Controls under Partial Observation*, Technical Report, Bonn, 1978.

[5] M. H. A. DAVIS, *On the existence of optimal policies in stochastic control*, this Journal, 11 (1973), pp. 587–594.

[6] M. H. A. DAVIS AND P. VARAIYA, *Dynamic programming conditions for partially observable stochastic systems*, Ibid. 11 (1973), pp. 226–261.

[7] J. L. DOOB, *Stochastic Processes*, Wiley, New York, 1967.

[8] R. J. ELLIOTT, *The optimal control of a stochastic system*, Ibid. 15 (1977), pp. 756–778.

[9] T. DUNCAN AND P. VARAIYA, *On the solutions of a stochastic control system*, Ibid. 9 (1971), pp. 354–371.

[10] W. H. FLEMING, *Optimal control of partially observable diffusions*, Ibid. 6 (1968), pp. 194–214.

[11] W. H. FLEMING AND R. W. RISHEL, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, 1975.

[12] H. J. KUSHNER, *Existence results for optimal stochastic controls*, J. Optimization Theory Appl., 15 (1975), pp. 347–359.

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

[14] R. S. LIPSTER AND A. N. SHIRYAYEV, *Statistics of Random Processes I*, Springer-Verlag, New York, 1977.

[15] S. SAKS, *Theory of the Integral*, Dover, New York, 1968.

[16] A. V. SKOROKHOD, *Studies in the Theory of Random Processes*, Addison-Wesley, Reading, Mass., 1965.

[17] D. H. WAGNER, *Survey of measurable selection theorems*, this Journal, 15 (1977), pp. 859–903.

[18] E. WONG, *Representation of martingales, quadratic variations and applications*, Ibid. 9 (1971), pp. 621–633.

[19] T. YAMADA AND S. WATANABE, *On the uniqueness of solutions to stochastic differential equations*, J. Math. Kyoto Univ., 11 (1971), pp. 155–167.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 5, September 1980

0363-0129/80/1805-0008 $01.00/0

ON CHARACTERIZING OPTIMAL POLICIES IN INFINITE-HORIZON STOCHASTIC CONTROL*

D. P. KENNEDY†

Abstract. An infinite-horizon sequential-decision model in discrete time is studied. Throughout, the minimal conditional expected loss at time n is defined as the essential infimum of the conditional expected losses at n associated with those policies that may be used after n. Necessary and sufficient conditions for conserving policies to be optimal are developed. An outline of how the model extends to continuous time is given, and the corresponding characterization of optimal conserving policies in continuous time is presented.

1. Introduction. We will consider an infinite-horizon sequential-decision model in discrete time which generalizes to continuous time quite naturally. The model is based on those presented in Striebel [5] and Kreps [2], and considers an arbitrary loss function.

Let $(\Omega, \mathscr{F})$ be a measurable space and let Π be a nonempty set. For each $\pi \in \Pi$, assume there is,

(i) a probability measure, P^π, on $\mathscr{F}$.

(ii) an extended-real-valued random variable, L_π, defined on $(\Omega, \mathscr{F})$, which is quasi-P^π-integrable,

(iii) a filtration, $\{\mathscr{F}_n^\pi, n \geqq 0\}$; that is, a nondecreasing sequence of sub σ-fields of $\mathscr{F}$ with $\mathscr{F}_\infty^\pi$ denoting the smallest σ-field containing each $\mathscr{F}_n^\pi$, $n \geqq 0$.

Here, Π is to be interpreted as the set of available policies, each policy π determining a probability distribution over the underlying sample space and a loss L_π. The σ-field $\mathscr{F}_n^\pi$ represents the information available to the controller if policy π has been used until time n; this formulation enables the model to cover the usual case when randomized policies are permitted.

Furthermore, assume that there is a nested sequence of equivalence relations, $\{\underset{n}{\sim}, n \geqq 0\}$, on Π; that is, for $\nu, \pi \in \Pi$, $\nu \underset{n}{\sim} \pi$ implies that $\nu \underset{k}{\sim} \pi$ for $0 \leqq k \leqq n$. If policy π is used up to time n, then the equivalence class containing π under $\underset{n}{\sim}$ is to be interpreted as the set of those policies which may be used after time n. We say that policy ν is *compatible* with π up to time n if $\nu \underset{n}{\sim} \pi$, and we will require that if $\nu \underset{n}{\sim} \pi$ then

(i) $\mathscr{F}_n^\nu = \mathscr{F}_n^\pi$, and

(ii) $P_n^\nu = P_n^\pi$, where P_n^π denotes the restriction of P^π to $\mathscr{F}_n^\pi$.

We now define the minimal conditional expected loss given that π has been used up to time n, by

$$V_n^\pi = \operatorname*{ess\,inf}_{\nu \underset{n}{\sim} \pi} E_\nu[L_\nu | \mathscr{F}_n^\nu], \qquad P_n^\pi \text{ a.s.} \tag{1.1}$$

Here and in the following, unless there is some indication to the contrary, it may be assumed that an essential infimum is taken with respect to the probability measure that qualifies almost surely the statement in which it appears; typically this will be the generic probability measure P_n^π. Also, E_π denotes expectation with respect to P^π.

For $\nu \in \Pi$, a policy π is said to be *optimal for ν at n* if $\pi \underset{n}{\sim} \nu$ and

$$V_m^\pi = E_\pi[L_\pi | \mathscr{F}_m^\pi], \qquad P^\pi \text{ a.s., for all } m \geqq n. \tag{1.2}$$

Generally, we will say that π is *optimal at n* if (1.2) holds. So if π is optimal at n then it is

* Received by the editors August 9, 1979, and in final form January 30, 1980.

† Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, Cambridge, CB2 1SB, England.

optimal for all ν at n such that $\nu \underset{\bar{n}}{\sim} \pi$. If each V_m^π, $m \geqq n$, is quasi-P^π-integrable, π is said to be *conserving at n* if $\{V_m^\pi, \mathcal{F}_m^\pi, m \geqq n\}$ is a P^π-martingale. Throughout, we will not require martingales (or submartingales) to be integrable, but only that they be quasi-integrable and satisfy the usual martingale equality (or submartingale inequality).

Notice that if π is optimal at n then it is conserving at n; the converse is not true in general. Kreps [2] has investigated conditions which are sufficient to ensure that a conserving policy is optimal for a slightly more specialized model. In § 3, under certain conditions, we give necessary and sufficient conditions for a conserving policy to be optimal, and we show that a uniform integrability condition related to one given by Kreps is necessary as well as sufficient for optimality. We also develop some properties of the model along the lines of Striebel ([5, Ch. 4]), which are useful in considering a continuous-time version of the model. This version is presented in § 4 where it is seen that the necessary and sufficient conditions for conserving policies to be optimal carry across directly from discrete time. We begin in § 2 by listing some easily established or known results which will be required subsequently.

2. Preliminaries. Consider a family of extended-real-valued random variables $\{X_\alpha; \alpha \in A\}$ defined on a probability space $(\Omega, \mathcal{F}, P)$. We will say that this family has the *(downwards) countable lattice property* if for any $\varepsilon > 0$ and countable subset $B \subseteq A$ there exists $\beta \in A$ with

$$X_\beta \leqq \inf_{\alpha \in B} X_\alpha + \varepsilon, \qquad \text{a.s.} \tag{2.1}$$

The family has the *(downwards) finite lattice property* if (2.1) holds for any finite subset B, while it is *directed downwards* if for any finite subset $B \subseteq A$ there exists $\beta \in A$ with $X_\beta \leqq \min_{\alpha \in B} X_\alpha$, a.s. Trivially, if the family is directed downwards it has the (downwards) finite lattice property. The (upwards) countable and finite lattice properties are defined in the obvious manner by reversing the sign of ε and the inequality in (2.1) and replacing the infimum by supremum.

If each X_α is quasi-integrable and $\operatorname{ess\,inf}_{\alpha \in A} X_\alpha$ is quasi-integrable, then for any sub σ-field $\mathcal{G} \subseteq \mathcal{F}$ it is immediate that

$$E[\operatorname*{ess\,inf}_{\alpha \in A} X_\alpha | \mathcal{G}] \leqq \operatorname*{ess\,inf}_{\alpha \in A} E[X_\alpha | \mathcal{G}] \qquad \text{a.s.} \tag{2.2}$$

On the left-hand side the essential infimum is that taken with respect to P, and on the right-hand side it is with respect to the restriction of P to $\mathcal{G}$. Conditions sufficient to ensure that $\operatorname{ess\,inf}_{\alpha \in A} X_\alpha$ is quasi-integrable are that either

$$EX_\alpha^+ < \infty, \quad \text{for some } \alpha \in A, \tag{2.3}$$

or

$$X_\alpha \geqq Y, \text{ a.s.} \quad \text{for each } \alpha, \text{ with } EY^- < \infty. \tag{2.4}$$

Recall that (Neveu [4, p. 121]) there always exists a countable subset $B \subseteq A$ such that

$$\inf_{\alpha \in B} X_\alpha = \operatorname*{ess\,inf}_{\alpha \in A} X_\alpha, \qquad \text{a.s.}$$

Using this fact, it is straightforward to establish the following (cf. Striebel [5, p. 198]).

PROPOSITION 2.5. *If either* (i) (2.3) *holds and the family* $\{X_\alpha; \alpha \in A\}$ *has the* (*downwards*) *finite lattice property or* (ii), *the family* $\{X_\alpha; \alpha \in A\}$ *has the* (*downwards*) *countable lattice property, then equality holds in* (2.2).

Returning to the sequential-decision system introduced in the previous section, following Striebel [5], we will say that the system has the *finite* (*or countable*) *lattice property* if for each $\pi \in \Pi$ and $n \geqq 0$ the family $\{E_\nu[L_\nu|\mathcal{F}_n^\nu];\ \nu \underset{n}{\sim} \pi\}$ has the (downwards) finite (or countable) lattice property with respect to P^π.

The archetypal situation where the system has either the finite or countable lattice property is obtained as follows. Let $(C, \mathcal{C})$ be a measurable space, usually referred to as the control space, and suppose that each $\pi \in \Pi$ is of the form $\pi = \{\pi_n\}_{n=0}^\infty$, where π_n is a C-valued random variable on $(\Omega, \mathcal{F})$. Typically, either π_n is $\mathcal{F}_{n+1}^\pi$-measurable, or $\mathcal{F}_n^\pi = \mathcal{F}_n$ for all π and π_n is $\mathcal{F}_n$-measurable (the two eventualities corresponding to randomized and non-randomized policies respectively). Then we will let $\nu \underset{n}{\sim} \pi$ if $\nu_k \equiv \pi_k$, $0 \leqq k \leqq n-1$ for $n \geqq 1$, with $\nu \underset{0}{\sim} \pi$ for all $\nu, \pi \in \Pi$. If, for each $n \geqq 0$, $\pi \in \Pi$, for all choices of $\nu^j \underset{n}{\sim} \pi$, $j = 1, \cdots, k$, and if for each partition $\Omega_j \in \mathcal{F}_n^\pi$, $j = 1, \cdots, k$ of Ω, defining

$$\nu_r = \sum_{j=1}^k \nu_r^j I(\Omega_j), \qquad r \geqq 0, \tag{2.6}$$

gives $\{\nu_r\}_{r=0}^\infty \in \Pi$, then we will say that the set of policies Π is *stable* (cf. Hinderer [1] and Streibel [5]); here $I(\Omega_j)$ denotes the indicator of the event Ω_j. Say that the system is *stable* if Π is stable and if

$$L_\nu = \sum_{j=1}^k L_{\nu^j} I(\Omega_j), \tag{2.7}$$

when $\nu = \{\nu_r\}_{r=0}^\infty$ is defined by (2.6).
If (2.6) and (2.7) hold for countably many policies $\nu^1, \nu^2, \cdots$ and countable partitions of Ω, then we say that Π is *countably stable* and the system is *countably stable*. A standard argument (Striebel [5, p. 87]) shows that if the system is stable (or countably stable) then it has the finite (or countable) lattice property. Notice that the model of Kreps [2] is countably stable.

Finally, we will require the next simple result on uniform integrability. Here $\{X_{\alpha\beta};\ (\alpha, \beta) \in A\}$ denotes a family of extended-real-valued random variables defined on the underlying probability space; for each α, $A_\alpha = \{\beta: (\alpha, \beta) \in A\}$ represents the section of A at α, while $B = \{\alpha: (\alpha, \beta) \in A \text{ for some } \beta\}$.

LEMMA 2.8. *If for each α the family $\{X_{\alpha\beta};\ \beta \in A_\alpha\}$ has the* (*downwards*) *finite lattice property then the random variables* $[\operatorname{ess\,inf}_{\beta \in A_\alpha} X_{\alpha\beta}]^-$, $\alpha \in B$, *are uniformly integrable if and only if the random variables $X_{\alpha\beta}^-$, $(\alpha, \beta) \in A$, are uniformly integrable.*

Proof. The sufficiency is immediate (and does not require the lattice property). For the necessity, for each α, there exist $\beta_1, \beta_2, \cdots \in A_\alpha$ with

$$\inf_i X_{\alpha\beta_i} = \operatorname*{ess\,inf}_{\beta \in A_\alpha} X_{\alpha\beta} \qquad \text{a.s.}$$

For each $\varepsilon > 0$, by the finite lattice property there exist $\gamma_i \in A_\alpha$ with

$$X_{\alpha\gamma_i} \leqq \bigwedge_{j=1}^i X_{\alpha\beta_j} + \varepsilon.$$

So for $a > 0$,

$$\inf_{\beta \in A_\alpha} \int_{\{X_{\alpha\beta} < -a\}} X_{\alpha\beta}\, dP \leqq \int_{\{\bigwedge_1^i X_{\alpha\beta_j} + \varepsilon < -a\}} \left[\bigwedge_{j=1}^i X_{\alpha\beta_j} + \varepsilon\right] dP,$$

letting $i \to \infty$ and $\varepsilon \to 0$ gives by monotone convergence that

$$\inf_{\beta \in A_\alpha} \int_{\{X_{\alpha\beta} < -a\}} X_{\alpha\beta}\, dP \leqq \int_{\{\operatorname{ess\,inf}_{\beta \in A_\alpha} X_{\alpha\beta} < -a\}} \operatorname*{ess\,inf}_{\beta \in A_\alpha} X_{\alpha\beta}\, dP.$$

Taking the infimum over α and letting $a \to \infty$ gives the result.

3. The discrete-time model. Throughout this section, for the model of § 1, we will assume that at least one of the following hold; either

$$E_\pi L_\pi^+ < \infty, \quad \text{for each } \pi \in \Pi, \tag{3.1}$$

or

$$\sup_\pi E_\pi L_\pi^- < \infty, \quad \text{and for each } \pi \in \Pi \tag{3.2}$$

and $n \geqq 0$ the family $\{E_\nu[L_\nu^- | \mathcal{F}_n^\nu];\ \nu \underset{n}{\sim} \pi\}$ has the (upwards) finite lattice property with respect to P_n^π.

It is then immediate that V_n^π is quasi-P^π-integrable with either of these assumptions since (3.1) implies that $E_\pi(V_n^\pi)^+ < \infty$, while Proposition 2.5 gives $E_\pi(V_n^\pi)^- < \infty$ when (3.2) holds. If the system is stable in the sense of the previous section the second part of (3.2) holds. In the following all the results established under the assumption (3.2) may be seen to hold when the random variables L_π are uniformly bounded below, i.e., when there exists a real M, with $L_\pi \geqq M$, P^π a.s., for all $\pi \in \Pi$.

Before discussing optimal policies we develop several elementary properties of the model following Striebel [5]; we omit those arguments which are standard or which are readily reconstructed.

Now, for each $\pi \in \Pi$, $N \geqq 0$ and $0 \leqq n < N$, define

$$F_{N,N}^\pi = V_N^\pi, \qquad F_{n,N}^\pi = \operatorname*{ess\,inf}_{\nu \underset{n}{\sim} \pi} E_\nu[F_{n+1,N}^\nu | \mathcal{F}_n^\nu], \qquad P_n^\pi \text{ a.s.} \tag{3.3}$$

By backwards induction on $n = N, N-1, \cdots$, it is easily established that each $F_{n,N}^\pi$ is quasi-P^π-integrable, and thus $F_{n,N}^\pi$ is well-defined by (3.3); in addition it follows that for $0 \leqq n < N$,

$$F_{n,N}^\pi \leqq \operatorname*{ess\,inf}_{\nu \underset{n}{\sim} \pi} E_\nu[V_{n+1}^\nu | \mathcal{F}_n^\nu] \leqq V_n^\pi, \qquad P_n^\pi \text{ a.s.} \tag{3.4}$$

Again, by induction on $N - n$ we may see that for $0 \leqq n \leqq N$,

$$F_{n,N}^\pi \geqq F_{n,N+1}^\pi, \qquad P_n^\pi \text{ a.s.,} \tag{3.5}$$

whence $F_n^\pi = \lim_{N \to \infty} F_{n,N}^\pi$, exists, P_n^π a.s.; with the assumptions above each F_n^π is quasi-P^π-integrable and $F_n^\pi = F_n^\nu$, P^π a.s. if $\nu \underset{n}{\sim} \pi$.

As in Striebel ([5, Ch. 4]), it is immediate from the definition (3.3) that for each $N > 0$ and π, $\{F_{n,N}^\pi, \mathcal{F}_n^\pi, 0 \leqq n \leqq N\}$, is a P^π-submartingale; furthermore it is the maximal submartingale with $F_{N,N}^\pi$ dominated by V_N^π, in the following sense. Say that a family of random variables $\{G_n^\pi; n \geqq 0, \pi \in \Pi\}$ is *compatible* if $G_n^\pi = G_n^\nu$, P^π a.s. whenever $\nu \underset{n}{\sim} \pi$. Then, if a compatible family is such that for each π, $\{G_n^\pi, \mathcal{F}_n^\pi, 0 \leqq n \leqq N\}$ is a P^π-submartingale, by backwards induction on n, $G_N^\pi \leqq V_N^\pi$, P^π a.s., for each π, implies that $G_n^\pi \leqq F_{n,N}^\pi$, P^π a.s., for each π and $n = 0, 1, \cdots, N$. This result may be extended to F_n^π when (3.1) holds by the next result.

PROPOSITION 3.6.

(a) *For all* $m \geqq n \geqq 0$ *and* $\pi \in \Pi$,

$$\operatorname*{ess\,inf}_{\nu \underset{n}{\sim} \pi} E_\nu[F^\nu_{n+1} | \mathcal{F}^\nu_n] \leqq F^\pi_n \tag{3.7}$$

$$\leqq \operatorname*{ess\,inf}_{\nu \underset{n}{\sim} \pi} E_\nu[V^\nu_m | \mathcal{F}^\nu_n] \leqq V^\pi_n, \qquad P^\pi_n \text{ a.s.}$$

(b) *If* (3.1) *holds, for* $m \geqq n+1$, *we have*

$$F^\pi_n = \operatorname*{ess\,inf}_{\nu \underset{n}{\sim} \pi} E_\nu[F^\nu_{n+1} | \mathcal{F}^\nu_n] \leqq \operatorname*{ess\,inf}_{\nu \underset{n}{\sim} \pi} E_\nu[F^\nu_m | \mathcal{F}^\nu_n], \qquad P^\pi_n \text{ a.s.}, \tag{3.8}$$

and thus $\{F^\pi_n, \mathcal{F}^\pi_n, n \geqq 0\}$ *is a* P^π*-submartingale.*

Proof. The relations (3.7) follow immediately from the above remarks; while if (3.1) holds, since for $N \geqq n$, when $\nu \underset{n}{\sim} \pi$,

$$F^\pi_n \leqq F^\pi_{n,N} \leqq E_\nu[F^\nu_{n+1,N} | \mathcal{F}^\nu_n], \qquad P^\pi_n \text{ a.s.},$$

using monotone convergence as $N \to \infty$, we have (since $E_\nu(F^\nu_{n+1,N})^+ < \infty$),

$$F^\pi_n \leqq E_\nu[F^\nu_{n+1} | \mathcal{F}^\nu_n], \qquad P^\pi_n \text{ a.s.};$$

the left-hand side of (3.8) follows from (3.7), showing the submartingale property. Finally the right-hand side of (3.8) follows directly from the submartingale property.

The next result follows immediately.

COROLLARY 3.9. *The following statements are equivalent.*

(i) *For each* $\pi \in \Pi$ *and* $n \geqq 0$, $V^\pi_n = F^\pi_n$, P^π a.s.

(ii) *For each* $\pi \in \Pi$ *and* $n \geqq 0$, $V^\pi_n = \text{ess inf}_{\nu \underset{n}{\sim} \pi} E_\nu[V^\nu_{n+1} | \mathcal{F}^\nu_n]$, P^π_n a.s.

(iii) *For each* $\pi \in \Pi$ *and* $m \geqq n \geqq 0$, $V^\pi_n = \text{ess inf}_{\nu \underset{n}{\sim} \pi} E_\nu[V^\nu_m | \mathcal{F}^\nu_n]$, P^π_n a.s.

(iv) *For each* $\pi \in \Pi$, $\{V^\pi_n, \mathcal{F}^\pi_n, n \geqq 0\}$ *is a* P^π*-submartingale.*

Conditions which are sufficient to ensure that (i)–(iv) hold follow from the remarks in § 2 (cf. Striebel [5, p. 68]).

COROLLARY 3.10. *If either* (a) (3.1) *holds and the system has the finite lattice property, or* (b) *the system has the countable lattice property, then* (i)–(iv) *of Corollary* 3.9 *hold.*

Recall that a submartingale $\{X_n, \mathcal{F}_n, n \geqq 0\}$ with $\sup_n EX^+_n < \infty$ is *regular* if $E[X_\infty | \mathcal{F}_n] \geqq X_n$, a.s., for each n, where $X_\infty = \lim_{n\to\infty} X_n$, a.s.

COROLLARY 3.11. *When* (3.1) *holds,* $\{F^\pi_n ; n \geqq 0, \pi \in \Pi\}$ *is the maximal compatible family such that for each* π, $\{F^\pi_n, \mathcal{F}^\pi_n, n \geqq 0\}$ *is a regular* P^π*-submartingale satisfying*

$$F^\pi_\infty \leqq E_\pi[L_\pi | \mathcal{F}^\pi_\infty], \qquad P^\pi \text{ a.s.} \tag{3.12}$$

Proof. Since $F^\pi_n \leqq E_\pi[L_\pi | \mathcal{F}^\pi_n]$, with $E_\pi L^+_\pi < \infty$, $\{F^\pi_n, \mathcal{F}^\pi_n, n \geqq 0\}$ is a regular P^π-submartingale satisfying (3.12). This follows from [4, Lemma IV-2-4], and [6, Theorem 2(ii), p. 234]. If $\{G^\pi_n ; n \geqq 0, \pi \in \Pi\}$ is a compatible family such that for each π, $\{G^\pi_n, \mathcal{F}^\pi_n, n \geqq 0\}$ is a regular P^π-submartingale satisfying

$$G^\pi_\infty \leqq E_\pi[L_\pi | \mathcal{F}^\pi_\infty], \qquad P^\pi \text{ a.s.},$$

then

$$G^\pi_n = G^\nu_n \leqq E_\nu[G^\nu_\infty | \mathcal{F}^\nu_n] \leqq E_\nu[L_\nu | \mathcal{F}^\nu_n], \qquad P^\pi_n \text{ a.s.},$$

for all ν such that $\nu \underset{n}{\sim} \pi$. Hence $G^\pi_n \leqq V^\pi_n$ and by the remarks preceding Proposition 3.6, $G^\pi_n \leqq F^\pi_{n,N}$, P^π a.s., for all $N \geqq n$; letting $N \to \infty$ gives $G^\pi_n \leqq F^\pi_n$, P^π a.s.

We now consider the characterization of optimal policies. Notice that if π is optimal at n then $E_\pi L_\pi = \inf_{\nu \underset{n}{\sim} \pi} E_\nu L_\nu$; the converse is not true in general, though we do have the following.

PROPOSITION 3.13. *If $E_\pi|L_\pi| < \infty$ and the system has the finite lattice property then the following are equivalent.*

(i) π is optimal at n.

(ii) $V^\pi_n = E_\pi[L_\pi|\mathscr{F}^\pi_n]$, P^π_n a.s.

(iii) $E_\pi L_\pi = \inf_{\nu \underset{n}{\sim} \pi} E_\nu L_\nu$.

Proof. It is immediate that (i) implies (ii) and (iii), and that (ii) implies (iii). If (iii) holds, for $m \geq n$

$$E_\pi L_\pi = \inf_{\nu \underset{n}{\sim} \pi} E_\nu L_\nu \leqq \inf_{\nu \underset{m}{\sim} \pi} E_\nu L_\nu \leqq E_\pi L_\pi;$$

thus there is equality, so by Proposition 2.5,

$$E_\pi V^\pi_m = E_\pi \,[\operatorname*{ess\,inf}_{\nu \underset{m}{\sim} \pi} E_\nu[L_\nu|\mathscr{F}^\nu_m]]$$

$$= \inf_{\nu \underset{m}{\sim} \pi} E_\nu L_\nu = E_\pi L_\pi.$$

But

$$V^\pi_m \leqq E_\pi[L_\pi|\mathscr{F}^\pi_m], \qquad P^\pi_m \text{ a.s.};$$

hence there is equality, P^π_m a.s., giving (i) and (ii).

Turning to conserving policies we observe that if π is conserving then $V^\pi_\infty = \lim_{n\to\infty} V^\pi_n$ exists P^π a.s.; this follows because a martingale is both a submartingale and a supermartingale, while $\sup_n E_\pi(V^\pi_n)^+ \leqq E_\pi L^+_\pi < \infty$, if (3.1) holds and $\sup_n E_\pi(V^\pi_n)^- < \infty$ if (3.2) holds. We may now determine when conserving policies are optimal.

THEOREM 3.14. *If* (3.2) *holds, then if π is conserving at n it is optimal at n if and only if*

(i) $V^\pi_\infty \geqq E_\pi[L_\pi|\mathscr{F}^\pi_\infty]$, P^π a.s., *and*

(ii) $(V^\pi_m)^-$, $m \geqq n$ *,are uniformly P^π-integrable.*

Proof. If π is conserving and (ii) holds, then $\{V^\pi_m, \mathscr{F}^\pi_m, m \geqq n\}$ is a regular P^π-supermartingale (Neveu, [4, p. 92]) so for $m \geqq n$, by (i),

$$V^\pi_m \geqq E_\pi[V^\pi_\infty|\mathscr{F}^\pi_m] \geqq E_\pi[L_\pi|\mathscr{F}^\pi_m], \qquad P^\pi \text{ a.s.},$$

showing that π is optimal.

Conversely, if π is optimal at n, $V^\pi_m = E_\pi[L_\pi|\mathscr{F}^\pi_m]$, P^π a.s. for all $m \geqq n$. But $E_\pi L^-_\pi < \infty$ since (3.2) holds, and so $(V^\pi_m)^-$, $m \geqq n$, are uniformly P^π-integrable and by Neveu [4, p. 31],

$$V^\pi_\infty = \lim_{m\to\infty} E_\pi[L_\pi|\mathscr{F}^\pi_m] \geqq E_\pi[L_\pi|\mathscr{F}^\pi_\infty], \qquad P^\pi \text{ a.s.}$$

We can investigate the condition (ii) of Theorem 3.14 a little further.

PROPOSITION 3.15. *For fixed $\pi \in \Pi$ and $n \geqq 0$, if the system has the finite lattice property then the following are equivalent.*

(i) *$(V^\pi_m)^-$, $m \geqq n$ are uniformly P^π-integrable.*

(ii) $\lim_{a\to\infty} \inf_{m \geqq n} \inf_{\nu \underset{m}{\sim} \pi} E_\nu L_\nu I(E_\nu[L_\nu|\mathscr{F}^\nu_m] < -a) = 0.$

Proof. By Lemma 2.8, (i) is equivalent to $E_\nu[L_\nu|\mathscr{F}^\nu_m]^-$, $\nu \underset{m}{\sim} \pi$, $m \geqq n$, being uniformly P^π-integrable, which is equivalent to (ii).

Kreps has shown that under a uniform integrability condition ([2, (1)]), which for his model implies (3.2), that (i) of Theorem 3.14 is sufficient for a conserving policy to be optimal. Together Theorem 3.14 and Proposition 3.15 show that a similar uniform integrability condition on the losses is also necessary. A simpler sufficient condition is given by (3.17).

LEMMA 3.16. *If*

$$\liminf_{a\to\infty\ \nu \underset{n}{\sim} \pi} E_\nu L_\nu I(L_\nu < -a) = 0, \tag{3.17}$$

then Proposition 3.15 (ii) *holds.*

Proof. Suppose that $\nu \underset{m}{\sim} \pi$, $m \geqq n$, $a > 0$, $b > 0$; then

$$\begin{aligned}\int_{\{E_\nu[L_\nu|\mathscr{F}^\nu_m]<-a\}} L_\nu\,dP^\nu &= \int_{\{L_\nu<-b,E_\nu[L_\nu|\mathscr{F}^\nu_m]<-a\}} L_\nu\,dP^\nu + \int_{\{L_\nu\geqq -b,E_\nu[L_\nu|\mathscr{F}^\nu_m]<-a\}} L_\nu\,dP^\nu \\ &\geqq \int_{\{L_\nu<-b\}} L_\nu\,dP^\nu - bP^\nu\{E_\nu[L_\nu|\mathscr{F}^\nu_m] < -a\} \\ &\geqq \int_{\{L_\nu<-b\}} L_\nu\,dP^\nu - bE_\nu L_\nu^-/a.\end{aligned}$$

Taking the infimum over ν and $m \geqq n$, and letting $a \to \infty$, $b \to \infty$ in such a way that $b/a \to 0$, gives the result since (3.17) implies that $\sup_{\nu \underset{n}{\sim} \pi} E_\nu L_\nu^- < \infty$.

Finally, when only (3.1) holds one can be less explicit.

PROPOSITION 3.18. *If* (3.1) *holds it is sufficient for π to be optimal at n that*

(i) $\{V^\pi_m, \mathscr{F}^\pi_m, n \leqq m \leqq \infty\}$ *is a P^π-martingale, and*

(ii) $V^\pi_\infty \geqq E_\pi[L_\pi|\mathscr{F}^\pi_\infty]$ $\quad P^\pi$ a.s.

If in addition, $\sup_m E_\pi[L_\pi|\mathscr{F}^\pi_m] > -\infty$, P^π a.s., *then* (i) *and* (ii) *are also necessary for π to be optimal at n.*

Proof. If (i) and (ii) hold, then for $m \geqq n$,

$$V^\pi_m = E_\pi[V^\pi_\infty|\mathscr{F}^\pi_m] \geqq E_\pi[L_\pi|\mathscr{F}^\pi_m], \qquad P^\pi \text{ a.s.},$$

showing that π is optimal at n. Conversely, if $\sup_m E_\pi[L_\pi|\mathscr{F}^\pi_m] > -\infty$, P^π a.s., and π is optimal at n, then $V^\pi_m = E_\pi[L_\pi|\mathscr{F}^\pi_m]$, $m \geqq n$,

$$V^\pi_\infty = \lim_{m\to\infty} E_\pi[L_\pi|\mathscr{F}^\pi_m] = E_\pi[L_\pi|\mathscr{F}^\pi_\infty], \qquad P^\pi \text{ a.s.},$$

by Neveu ([4, p. 31]), and (i), (ii) hold.

Notice that when (3.1) holds, since $V^\pi_m \leqq E_\pi[L_\pi|\mathscr{F}^\pi_m]$, P^π a.s., we have $V^\pi_\infty \leqq E_\pi[L_\pi|\mathscr{F}^\pi_\infty]$ and so (ii) of Proposition 3.18 is equivalent to $V^\pi_\infty = E_\pi[L_\pi|\mathscr{F}^\pi_\infty]$, P^π a.s.

4. The continuous-time model. One of the virtues of the model presented in the previous sections is that a continuous-time version may be given and analogous results established with essentially no extra work. We will simply just point out where the appropriate changes must be made. As before, we have an underlying measurable space $(\Omega, \mathscr{F})$, and a nonempty set Π. For each $\pi \in \Pi$ assume that there is (i) a probability measure, P^π, on $\mathscr{F}$ so that $(\Omega, \mathscr{F}, P^\pi)$ is a complete probability space, (ii) an extended-real-valued random variable, L_π, defined on $(\Omega, \mathscr{F})$, and (iii) a left-continuous filtration $\{\mathscr{F}^\pi_t, 0 \leqq t \leqq T\}$, with $\mathscr{F}^\pi_0$ containing all the P^π-null sets of $\mathscr{F}$. Here, we may have $T \leqq \infty$.

Throughout this section we will assume that

$$E_\pi L^+_\pi < \infty, \quad \text{for all } \pi \in \Pi. \tag{4.1}$$

Again we assume that there is a nested family of equivalence relations $\underset{t}{\sim}$, $t \geqq 0$ on Π, so that $\nu \underset{t}{\sim} \pi$ implies that $\nu \underset{s}{\sim} \pi$, for $0 \leqq s \leqq t$. As in discrete time we will require that if $\nu \underset{t}{\sim} \pi$

then $\mathscr{F}_t^\nu = \mathscr{F}_t^\pi$ and $P_t^\nu = P_t^\pi$, where P_t^π is the restriction of P^π to $\mathscr{F}_t^\pi$. Define the minimal conditional expected loss by

$$V_t^\pi = \operatorname*{ess\,inf}_{\nu \underset{t}{\sim} \pi} E_\nu[L_\nu | \mathscr{F}_t^\nu], \qquad P_t^\pi \text{ a.s.,} \tag{4.2}$$

The next result, as well as being of some interest in its own right, demonstrates how the development of § 3 may be extended to continuous time and how the continuous-time model (over a finite or infinite horizon) may be approximated by the discrete-time model in natural fashion. Define a compatible family of random variables in the obvious way.

THEOREM 4.3. *There exists a maximal compatible family* $\{F_{t-}^\pi; 0 \leqq t < T, \pi \in \Pi\}$ *such that for each* π, $\{F_{t-}^\pi, \mathscr{F}_t^\pi, 0 \leqq t < T\}$ *is a left-continuous* P^π*-submartingale dominated by* $V_t^\pi, 0 \leqq t < T$.

Proof. Consider the dyadic rationals $k/2^n$, $k \geqq 0$, $n \geqq 0$. For each $\pi \in \Pi$ and $N \geqq 0$, define

$$F_{N/2^n, N/2^n}^\pi(n) = V_{N/2^n}^\pi$$

and

$$F_{k/2^n, N/2^n}^\pi(n) = \operatorname*{ess\,inf}_{\nu \underset{k/2^n}{\sim} \pi} E_\nu[F_{(k+1)/2^n, N/2^n}^\nu(n) | \mathscr{F}_{k/2^n}^\nu], \qquad P_{k/2^n}^\pi \text{ a.s.}$$

for $0 \leqq k < N$. By the previous section $F_{k/2^n}^\pi(n) = \lim_{N\to\infty} F_{k/2^n, N/2^n}^\pi(n)$ exists, P^π a.s., for each $k/2^n$; if q is any dyadic rational, then $F_q^\pi(n)$ is nonincreasing in n, P^π a.s., by (3.8) (since (4.1) holds), so that $F_q^\pi = \lim_{n\to\infty} F_q^\pi(n)$ exists, P^π a.s. It is immediate from (3.7) that $F_q^\pi \leqq V_q^\pi$, P^π a.s.

Furthermore, if $q_1 \leqq q_2$ are dyadic rationals and n is large enough so that $q_1 = p/2^n$, $q_2 = r/2^n$ for some $p, r \geqq 0$, by the submartingale property of $F_q^\pi(n)$ we have

$$E_\pi[F_{q_2}^\pi(n) | \mathscr{F}_{q_1}^\pi] \geqq F_{q_1}^\pi(n), \qquad P^\pi \text{ a.s.,}$$

and monotone convergence gives (using (4.1) again) that as $n \to \infty$,

$$E_\pi[F_{q_2}^\pi | \mathscr{F}_{q_1}^\pi] \geqq F_{q_1}^\pi, \qquad P^\pi \text{ a.s.}$$

Since $E_\pi(F_q^\pi)^+ \leqq E_\pi L_\pi^+ < \infty$ for each q, using the submartingale convergence theorem, we may define for each real t, $0 < t < T$,

$$F_{t-}^\pi = \lim_{q \uparrow\uparrow t} F_q^\pi, \qquad P^\pi \text{ a.s.,}$$

where $q \uparrow\uparrow t$ denotes $q \uparrow t$ through dyadic rationals q with $q < t$; set $F_{0-}^\pi = F_0^\pi$. It may be checked immediately that $\{F_{t-}^\pi, \mathscr{F}_t^\pi, 0 \leqq t < T\}$ is a left-continuous P^π-submartingale (cf. Meyer [3, p. 95]). To see that it is dominated by V_t^π, suppose that $\nu \underset{t}{\sim} \pi$ so that $\nu \underset{q}{\sim} \pi$ for all $q < t$. Then

$$F_q^\pi \leqq V_q^\pi \leqq E_\nu[L_\nu | \mathscr{F}_q^\nu], \qquad P^\pi \text{ a.s.}$$

Letting $q \uparrow t$, using Neveu ([4, p. 31]) and the left continuity of the filtration we have

$$\begin{aligned} F_{t-}^\pi &\leqq E_\nu[L_\nu | \mathscr{F}_{t-}^\nu] \\ &= E_\nu[L_\nu | \mathscr{F}_t^\nu], \qquad P^\pi \text{ a.s.,} \end{aligned}$$

which implies that $F_{t-}^\pi \leqq V_t^\pi$, P^π a.s. For the maximality assume that $\{G_t^\pi; 0 \leqq t < T, \pi \in \Pi\}$ is compatible, with $\{G_t^\pi, \mathscr{F}_t^\pi, 0 \leqq t < T\}$ a left-continuous P^π-submartingale with $G_t^\pi \leqq V_t^\pi, P^\pi$ a.s., for each π. By Corollary 3.11, for each dyadic rational q, $G_q^\pi \leqq F_q^\pi$, P^π a.s., letting $q \uparrow\uparrow t$ gives $G_t^\pi \leqq F_{t-}^\pi$, P^π a.s., by the left-continuity of G_t^π, thus completing the proof.

As in discrete time define a policy π to be optimal at t if

$$V_s^\pi = E_\pi[L_\pi | \mathscr{F}_s^\pi], \qquad P^\pi \text{ a.s.},$$

for each s, $t \leqq s < T$. The analogue of Proposition 3.13 then carries over to the continuous-time case immediately. Let us now assume that the system has the finite lattice property. Exactly as in Corollary 3.10 it follows from (4.1) that for each $\pi, \{V_t^\pi, \mathscr{F}_t^\pi, 0 \leqq t < T\}$ is a P^π-submartingale. If we also assume the corresponding statement to (3.2) got by replacing n by t, we have that for each t, V_t^π is P^π-integrable. Using the finite lattice property we have that $E_\pi V_t^\pi = \inf_{\nu \overset{\frown}{t} \pi} E_\nu L_\nu$. By standard results (Meyer [3, p. 95]), $\{V_t^\pi, \mathscr{F}_t^\pi, 0 \leqq t < T\}$ has a left-continuous (P^π) modification if and only if the map

$$t \rightarrow \inf_{\nu \overset{\frown}{t} \pi} E_\nu L_\nu \quad \text{is left-continuous in } 0 < t < T. \tag{4.4}$$

By Corollary 3.9 this left-continuous modification, V_{t-}^π, coincides with F_{t-}^π defined in Theorem 4.3.

Then, assuming (4.4) holds, (4.1) ensures that $\sup_t E_\pi(V_t^\pi)^+ \leqq E_\pi L_\pi^+ < \infty$, and so (Meyer, [3, p. 96]) for $T \leq \infty$

$$V_T^\pi = \lim_{t \rightarrow T} V_{t-}^\pi, \qquad \text{exists}, \qquad P^\pi \text{ a.s.}$$

Again, if we define π to be conserving at t if $\{V_s^\pi, \mathscr{F}_s^\pi, t \leqq s < T\}$ is a P^π-martingale, we may conclude that if π is conserving at t, it is optimal at t if and only if

$$V_T^\pi \geqq E_\pi[L_\pi | \mathscr{F}_T^\pi], \qquad P^\pi \text{ a.s.}, \tag{4.5}$$

and

$$(V_s^\pi)^-, t \leqq s < T, \tag{4.6}$$

are uniformly P^π-integrable.

Then by virtue of the finite lattice property (4.6) is equivalent to condition (ii) of Proposition 3.15 with m, n replaced by s, t respectively. Exactly as before a sufficient condition for (4.6) is that

$$\lim_{a \rightarrow \infty} \inf_{\nu \overset{\frown}{t} \pi} E_\nu L_\nu I(L_\nu < -a) = 0.$$

REFERENCES

[1] K. Hinderer, *Foundations of Non-stationary Dynamic Programming with Discrete Time Parameter*, Lecture Notes in Operations Research and Mathematical Systems Vol. 33, Springer-Verlag, New York, 1970.

[2] D. M. Kreps, *Sequential decision problems with expected utility criteria III: upper and lower transience*, this Journal, 16 (1978), pp. 420–428.

[3] P. A. Meyer, *Probability and Potentials*, Blaisdell, Waltham, MA, 1966.

[4] J. Neveu, *Discrete-Parameter Martingales*, North-Holland, Amsterdam, 1975.

[5] C. Striebel, *Optimal Control of Discrete Time Stochastic Systems*, Lecture Notes in Economics and Mathematical Systems, Vol. 110, Springer-Verlag, New York, 1975.

[6] Y. S. Chow and H. Teicher, *Probability Theory*, Springer-Verlag, New York, 1978.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 5, September 1980

ON EXISTENCE OF PERIODIC MOTIONS IN NONLINEAR CONTROL SYSTEMS WITH PERIODIC INPUTS*

R. K. MILLER† AND A. N. MICHEL‡

Abstract. The existence of periodic solutions of nonlinear control systems subjected to sinusoidal forcing functions, using the describing function method, is studied. The setting is general enough to allow systems with delays, systems with discontinuous nonlinearities, systems with hysteresis nonlinearities, and so forth. The present results state that if the linearized describing function problem can be solved and if certain bounds (which depend on the exact form of the solution of the describing function problem) can be satisfied, then there is a periodic solution of the exact problem. Furthermore, the present results provide relative error bounds between the response of the exact problem and the associated linearized describing function problem. To demonstrate the applicability of the method advanced, a specific example is considered.

1. Introduction.

We study the existence of periodic solutions of nonlinear systems subjected to sinusoidal forcing functions using describing function techniques. We establish specific computable conditions which guarantee that when the usual describing function method predicts a periodic response, then the nonlinear system has a periodic response which is nearly equal to the predicted one. Our conditions are quite natural in the sense that they involve precise statements of the usual criteria, i.e., the linear part of the system must constitute a good low pass filter and the describing function should provide a good approximation to the response of the nonlinear element for nearly sinusoidal inputs.

In our results, we allow discontinuous nonlinearities. For this reason, it is necessary to work in the space of continuous functions (rather than in an L_2-space), it is necessary to use a generalized notion of solution (in the sense of Filippov [4], [9], [10], [16]), and uniqueness of the predicted periodic solution cannot be assured.

Many previous results address various aspects of the theoretical justification of describing functions. Specifically, the results of Bass [1], Bergen and Franks [2], and Williamson [17] are concerned with the prediction of sustained oscillations under zero forcing functions. This is a different problem from the one considered in this paper. On the other hand, the interesting work of Sandberg [15] (see also Mees and Bergen [11]) involves systems with Lipschitz continuous nonlinearities. Under their hypotheses, one can show that the basic system under consideration and also the corresponding describing function approximate system have unique solutions, which are locally L_2, and furthermore, the L_2-norm of the difference between these solutions is estimated. Furthermore, Holtzman [7] treats a problem similar to the one considered in [15], using a very different analysis. His results are local, replacing global Lipschitz conditions by local differentiability, and he obtains error estimates in the uniform norm over continuous periodic functions. Our results differ from those in [7] and [15] since we allow discontinuous nonlinear elements such as relays with or without dead zones and/or hysteresis. Moreover, we emphasize that our results differ in kind since they state that *if the linearized describing function problem can be solved and if certain bounds (which depend on the exact form of the solution of the describing function problem) can be satisfied, then there is a periodic solution of the exact problem.*

* Received by the editors June 27, 1979, and in revised form November 29, 1979. This research was supported by the National Science Foundation under grant ENG 77-28446.

† Department of Mathematics, Iowa State University, Ames, Iowa 50011.

‡ Department of Electrical Engineering and Engineering Research Institute, Iowa State University, Ames, Iowa 50011.

Other somewhat related work is contained in the absolute stability results of Popov [14]. A good discussion of describing function methods and theory (up to 1968) can be found in the book by Gelb and Vander Velde [5]. Background material on the theory of oscillations can be found, e.g., in Hale [18].

The remainder of this paper consists of four parts. In § 2, we establish our results for very general systems in an abstract setting while in § 3 we consider systems which can appropriately be described by integrodifferential equations. To demonstrate the applicability of these results, we consider a specific example in § 4. In § 5, we discuss the physical implications of the principal hypotheses used in our results.

2. General systems. Let $X = \{\phi : R^1 \to R^1 | \phi$ is continuous, T-periodic, $\phi(t+\pi) = -\phi(t)\}$ with norm defined by $\|\phi\| = \sup\{|\phi(t)| : 0 \leqq t \leqq T\}$. It is routine to check that X is a Banach space. Moreover, with $\omega = 2\pi/T$, we have

$$\phi = \underset{N\to\infty}{\text{l.i.m.}} \sum_{\substack{n=-N\\ n \text{ odd}}}^{N} \phi_n e^{in\omega t},$$

where

$$\phi_n = T^{-1} \int_0^T \phi(t)\, e^{-in\omega t} dt, \qquad n \text{ odd}, \quad \phi_{-n} = \bar{\phi}_n$$

are the Fourier coefficients of ϕ, and Parseval's identity is

$$\frac{1}{T} \int_0^T |\phi(t)|^2\, dt = \sum_{n \text{ odd}} |\phi_n|^2.$$

We define a projection P on X by

$$(P\phi)(t) = \phi_1 e^{i\omega t} + \phi_{-1} e^{-i\omega t},$$

and we define $B(\psi, \varepsilon) = \{\phi \in X : \|\phi - \psi\| \leqq \varepsilon\}$.

We shall study periodic solutions of a nonlinear system of the form.

$$(E) \qquad x = h + GF(x)$$

by analyzing the associated approximate system

$$(\hat{E}) \qquad \hat{x} = h + GPF(\hat{x}).$$

We shall require the following assumptions for (E) and $(\hat{E})$:

(A1) $h(t) = a \sin(\omega t + b)$ for some real constants a and b;

(A2) System $(\hat{E})$ has a solution $\hat{x} \in PX$ where $PX = \{\hat{\phi} : \hat{\phi} = P\phi, \phi \in X\}$;

(A3) $F : X \to X$ is a continuous map;

$(A4)$ $G : X \to X$ is a linear map defined by

$$(G\phi)(t) = \sum_{n \text{ odd}} G_n \phi_n e^{in\omega t},$$

where $G_{-n} = \bar{G}_n$ and $\sup\{|nG_n| : n \text{ odd}\} = \gamma < \infty$.

Before we can remove the continuity condition (A3) and before we can prove the main result of this section (Theorem 1) we need to establish two preliminary results (Lemmas 1 and 2).

LEMMA 1. *The map G defined by* (A4) *is a completely continuous linear map on X and satisfies $PG = GP$.*

Proof. Since $\|G\phi\| \leqq \sum_{n \text{ odd}} |G_n\phi_n| \leqq \sqrt{\sum |G_n|^2}\sqrt{\sum |\phi_n|^2}$ for any $\phi \in X$, then $G\phi \in X$. An elementary argument involving Fourier coefficients will show that G is a closed linear map. Thus G is continuous.

Given a bounded set $S \subset B(0, M)$ for some M, then for $\phi \in S$,

$$\frac{1}{T}\int_0^T |\phi(t)|^2\, dt = \sum_{n \text{ odd}} |\phi_n|^2 \leqq M^2.$$

Thus $\|G\phi\| \leqq \sqrt{\sum |G_n|^2} M$, i.e., GS is a uniformly bounded set. If $\phi \in S$, then $G\phi$ has L_2-derivative

$$(G\phi)' = \underset{N \to \infty}{\text{l.i.m.}} \sum_{\substack{n=-N \\ n \text{ odd}}}^{N} G_n\phi_n(in\omega)\, e^{in\omega t},$$

and

$$\frac{1}{T}\int_0^T |(G\phi)'|^2\, dt \leqq \omega \sum |nG_n\phi_n|^2 \leqq \omega\gamma \sum |\phi_n|^2.$$

Thus, for any $t, \tau \in [0, T]$, $t < \tau$, we have

$$|G\phi(\tau) - G\phi(t)| \leqq \int_t^\tau |(G\phi)'(u)|\, du \leqq \sqrt{\tau - t}\sqrt{T\omega\gamma} M.$$

Thus the set GS is equicontinuous. Hence GS has compact closure.

The assertion that $PG = GP$ on X is trivial. □

LEMMA 2. *If* (A1)–(A4) *are true and if there are constants $\varepsilon, M > 0$ such that for all $x \in B(\hat{x}, \varepsilon)$ it is true that $\|F(x)\| \leqq M$, and if*

(1) $$[GP(F(x) - F(\hat{x})) + (I - P)GF(x)]\, \varepsilon B(0, \varepsilon),$$

then system (E) *has a solution $x \in B(\hat{x}, \varepsilon)$.*

Proof. Define $y = x - \hat{x}$ and subtract $(\hat{E})$ from (E) to see that

(2) $$y = GP(F(\hat{x} + y) - F(\hat{x})) + (I - P)GF(\hat{x} + y).$$

The hypotheses imply that the right-hand side of (2) defines a continuous map of $B(0, \varepsilon)$ into itself. By Lemma 1, the right-hand side of (2) is also completely continuous. Thus, the Schauder fixed point theorem (see, e.g., [3, p. 456, Theorem 5]) can be applied to obtain a solution of (2) with $y \in B(0, \varepsilon)$. □

Next, we consider the case where F is not necessarily continuous. We will approximate F by a sequence of nonlinear operators F_m which are assumed to satisfy the following conditions:

(A5) $F_m : X \to X$ with F_m continuous;

(A6) For any point $x \in PX$ and any sequence $\{x_m\}$ in PX, if $x_m \to x$, then $PF_m(x_m) \to PF(x)$.

We now consider the approximating equations

(E_m) $$x_m = h + GF_m(x_m),$$

and

$(\hat{E}_m)$ $$\hat{x}_m = h + GPF_m(\hat{x}_m).$$

We say that system (E) has a *weak solution* x if there exist a sequence of equations (E_m) such that (A5)–(A6) are true, solutions x_m and a function $w \in L_2(0, T)$ such that $F_m(x_m)$

tends weakly in $L_2(0, T)$ to $w(F_m(x_m) \rightharpoonup w)$, $x_m \to x$ in X, and

$$x = h + Gw.$$

We now state and prove the main result of this section.

THEOREM 1. *Suppose that* (E_m) *and* $(\hat{E}_m)$ *satisfy conditions* (A1)–(A4) *for each* $m \geqq 1$ *and that* F_m *satisfies conditions* (A5)–(A6). *Suppose there exist positive constants* M, M_1, ε, *and* ε_1 *such that* $0 < \varepsilon_1 < \varepsilon$, $\|PF_m(\hat{x}_m)\| < M_1$ *for all* $m \geqq 1$. *Then* $(\hat{E})$ *has a solution* $\hat{x}$. *If for all* $m \geqq 1$ *and for all* $x \in B(\hat{x}, \varepsilon)$ *we have* $\|F_m(x)\| \leqq M$ *and*

$$[GP(F_m(x) - F_m(\hat{x})) + (I - P)GF_m(x)] \varepsilon B(0, \varepsilon), \tag{3}$$

then (E) *has a weak solution* x *with* $\|x - \hat{x}\| < \varepsilon$.

Proof. Since the sequence $\{PF_m(\hat{x}_m)\}$ is bounded, then by possibly taking a subsequence we can assume that $PF_m(\hat{x}_m)$ tends weakly in $L_2(0, T)$ to some function z. The proof of Lemma 1 is easily extended to see that G is a completely continuous linear map from $L_2(0, T)$ into X. Thus $GPF_m(\hat{x}_m) \to Gz$ in X and $\hat{x}_m \to h + Gz$. By (A6) it follows that $\mathrm{PF}_m(\hat{x}_m) \to PF(h + Gz)$. Thus $\hat{x} = h + Gz$ solves $(\hat{E})$.

Next, by (A6), $PF_m(\hat{x}) \to PF(x)$ and so it follows from (3) that

$$[GP(F_m(x) - F(\hat{x}_m)) + (I - P)GF_m(x)] \varepsilon B(\hat{x}_m, \varepsilon_3), \tag{1'}$$

for $x \in B(\hat{x}_m, \varepsilon_3)$ and m large. Here $\varepsilon_3 \in (\varepsilon_1, \varepsilon)$. Indeed let $k = (\varepsilon - \varepsilon_1)/3$, $\varepsilon_2 = \varepsilon_1 + k$, $\varepsilon_3 = \varepsilon_2 + k$ and pick m_0 so large that for $m \geqq m_0$ we have $\|\hat{x}_m - \hat{x}\| < k$, and

$$[GP(F_m(x) - F_m(\hat{x}_m)) + (I - P)GF_m(x)] \varepsilon B(\hat{x}, \varepsilon_2) \subset B(\hat{x}_m, \varepsilon_3),$$

when $x \in B(\hat{x}_m, \varepsilon_3) \subset B(\hat{x}, \varepsilon)$. Also $\|F_m(x)\| \leqq M$ on $B(\hat{x}_m, \varepsilon_3)$. By Lemma 2 it follows that (E_m) has a solution $x_m \in B(\hat{x}_m, \varepsilon_3)$. Also $\|x_m - \hat{x}_m\| \leqq \varepsilon_3$, $x_m \in B(\hat{x}, \varepsilon)$ and $\|F_m(x_m)\| \leqq M$ for all m large.

Since $F_m(x_m)$ is bounded by M, it is weakly compact in $L_2(0, T)$. Thus by possibly taking a subsequence we can assume that $F_m(x_m) \rightharpoonup w$ for some $w \in L_2(0, T)$. Since G is completely continuous on $L_2(0, T)$ to X, then $GF_m(x_m) \to Gw$ in X. Thus $x_m = h + GF_m(x_m) \to h + Gw$. Define $x = h + Gw$. Then x is the weak solution of (E) in $B(\hat{x}, \varepsilon)$. □

The next result is a direct consequence of Theorem 1.

COROLLARY 1. *Suppose equations* (E_m) *and* $(\hat{E}_m)$ *satisfy conditions* (A1)–(A4) *for each* $m \geqq 1$ *and that assumptions* (A5)–(A6) *are true. Suppose there are constraints* M, M_1, ε, *and* ε_1 *such that* $0 < \varepsilon_1 < \varepsilon$, $\|PF_m(\hat{x}_m)\| < M_1$ *for all* $m \geqq 1$. *Then equation* $(\hat{E})$ *has a solution* $\hat{x}$. *If further* $\|F_m(x)\| \leqq M$ *and*

$$\|GP(F_m(x) - F_m(\hat{x}))\| + M\Big(\sum_{\substack{n \text{ odd} \\ |n| \geqq 3}} |G_n|^2\Big)^{1/2} \leqq \varepsilon_1, \tag{4}$$

for all $x \in B(\hat{x}, \varepsilon)$ *and all* $m \geqq 1$, *then equation* (E) *has a solution* $x \in B(\hat{x}, \varepsilon)$. □

3. Systems described by integrodifferential equations. We now consider systems described by integrodifferential equations. Let L be the linear integrodifferential operator defined by

$$(Ly)(t) = y^{(J)}(t) + \sum_{j=0}^{J-1} \left[\sum_{k=1}^{\infty} b_{jk} y^{(j)}(t - t_k) + \int_{-\infty}^{t} c_j(t - s) y^{(j)}(s)\, ds \right],$$

where $J \geqq 1$ and $\{t_n\}$ is an increasing sequence with $t_1 \geqq 0$ and $t_k \to \infty$. Let $\omega = 2\pi/T$ as above and define

$$H(s)=s^J+\sum_{j=0}^{J-1} s^j\left[\sum_{k=1}^{\infty} b_{jk}\,e^{-st_k}+c_j^*(s)\right].$$

In the following, we will require the following assumptions:

(A7) For some integer $M\geqq 0$ we have

$$\int_0^\infty t^m|c_j(t)|\,dt+\sum_{k=1}^{\infty}|b_{jk}|t_k^m<\infty,\qquad 0\leqq m\leqq 2M.$$

(A8) In the half-plane $\operatorname{Re} s\geqq 0$ the characteristic equation $H(s)=0$ has only finitely many roots $\{s_j: 1\leqq j\leqq K_1\}$. The first K_0 roots are assumed to be pure imaginary, $s_j=i\tau_j$ for $1\leqq j\leqq K_0$, while $\operatorname{Re} s_j>0$ for $K_0+1\leqq j\leqq K_1$. The multiplicity of any purely imaginary root is assumed to be at most M, where M is defined in (A7).

Let $a_j(t)$ be that function of bounded variation with continuous part $c_j(t)$, jump of height b_{jk} at t_k and no singular part. Thus we can write

$$(Ly)(t)=y^{(J)}(t)+\sum_{j=0}^{J-1}\int_0^\infty da_j\,(s)y^{(j)}(t-s).$$

Clearly L has transfer function $H(s)$ given by

$$H(s)=s^J+\sum_{j=0}^{J-1} s^j a_j^*(s).$$

As usual, we can write $L[y]=f$ as an equivalent system of first order in matrix-vector form

$$x(t)=\int_0^\infty dB(s)x(t-s)+F(t),\tag{5}$$

where the components of the vector F are $(0,\cdots,0,f)^T$ and the components of the vector $x(t)$ are $x_1(t)=y(t)$, $x_2(t)=y'(t),\cdots,x_J(t)=y^{(J-1)}(t)$. We also consider the associated matrix integrodifferential equation

$$R'(t)=\int_0^t dB(s)R(t-s),\qquad R(0)=I.\tag{6}$$

Its solution is called the resolvent matrix $R(t)$. (See, e.g., [12] for a discussion of the resolvent.) According to Theorem 2.2 in Jordan and Wheeler [8], the resolvent can be written in the form

$$R(t)=S(t)+N(t)+U(t),\tag{7}$$

where $S\in L_1(0,\infty)$ and there exist polynomials $p_j(t)$ with matrix coefficients such that

$$N(t)=\sum_{j=0}^{K_0} p_j(t)\,e^{i\tau_j t},\quad S(t)=\sum_{j=K_0+1}^{K_1} p_j(t)\,e^{s_j t}.$$

Before we state and prove the main results of this section (Theorems 2 and 3) we require the following preliminary result.

LEMMA 3. *If assumptions* (A7) *and* (A8) *are true with* $K_0=0$ (so that $N(t)\equiv 0$ *in* (7)), *then for any T-periodic continuous function F,* (5) *has the unique T-periodic solution*

$$x(t)=\int_{-\infty}^{t} S(t-s)F(s)\,ds-\int_{t}^{\infty} U(t-s)F(s)\,ds. \tag{9}$$

Proof. It is clear from (8) and from the fact that $R\in L_1(0,\infty)$ that (9) defines a T-periodic, differentiable function. It remains to be shown that $x(t)$ defined in this manner solves (5). Let us define

$$V(t)=\begin{cases} S(t) & \text{if } t\geqq 0,\\ -U(t) & \text{if } t<0.\end{cases}$$

Then (9) can be written as

$$x(t)=\int_{-\infty}^{\infty} V(t-s)F(s)\,ds.$$

By (9) we have

$$x'(t)=S(0)F(t)+U(0)F(t)+\int_{-\infty}^{t} S'(t-s)F(s)\,ds-\int_{t}^{\infty} U'(t-\tau)F(\tau)\,d\tau.$$

Using the facts that $S(0)+U(0)=R(0)=I$ is the identity matrix, that R solves (6) and (by a slight extension of Lemma 2 in Miller [13]) that U solves (5) with $F\equiv 0$, we obtain

$$\begin{aligned}
x'(t)=F(t)&+\int_{-\infty}^{t}\left[\int_{0}^{t-s} dB(t-s-u)S(u)-\int_{-\infty}^{0} dB(t-s-u)U(u)\right]F(s)\,ds\\
&-\int_{t}^{\infty}\left(\int_{-\infty}^{t-s} dB(t-s-u)U(u)\right)F(s)\,ds\\
=F(t)&+\int_{-\infty}^{t}\left(\int_{-\infty}^{t-s} dB(t-s-u)V(u)\right)F(s)\,ds\\
&+\int_{t}^{\infty}\left(\int_{-\infty}^{t-s} dB(t-s-u)V(u)\right)F(u)\,du\\
=F(t)&+\int_{-\infty}^{\infty}\left(\int_{-\infty}^{t-s} dB(t-s-u)V(u)\right)F(s)\,ds\\
=F(t)&+\int_{-\infty}^{\infty}\left(\int_{0}^{\infty} dB(u)V(t-s-u)\right)F(s)\,ds\\
=F(t)&+\int_{0}^{\infty} dB(u)\left(\int_{-\infty}^{\infty} V(t-s-u)F(s)\,ds\right)\\
=F(t)&+\int_{0}^{\infty} dB(u)x(t-u).
\end{aligned}$$

Thus $x(t)$ solves (9). □

THEOREM 2. *Assume that conditions* (A7) *and* (A8) *are true and that the nonresonance condition,*

(A9) $i\tau_j\neq in\omega$ *for all pure imaginary roots* $s_j=i\tau_j$, $1\leqq j\leqq K_0$, *and for all odd integers n,* *is satisfied. Then for any T periodic, continuous function F with Fourier series*

$$F(t) \sim \sum_{n \text{ odd}} F_n e^{in\omega t},$$

there is a unique T-periodic solution

$$y(t) = \sum_{n \text{ odd}} F_n H_n e^{in\omega t}, \tag{10}$$

where

$$H_n = S^*(in\omega) + N^*(in\omega) - \int_0^\infty U(-t)\, e^{in\omega t}\, dt = O\left(\frac{1}{n}\right),$$

S^* is the Laplace transform of $S(t)$ and $N^*(in\omega)$ is the analytic continuation of the Laplace transform of $N(t)$.

Proof. For $\varepsilon > 0$ and small consider the system

$$x'(t, \varepsilon) = \int_0^\infty dB(s)\, e^{-\varepsilon s} x(t-s, \varepsilon) - \varepsilon x(t, \varepsilon) + F(t).$$

The corresponding resolvent is

$$R(t)\, e^{-\varepsilon t} = (S(t) + N(t))\, e^{-\varepsilon t} + e^{-\varepsilon t} U(t).$$

The unique T-periodic solution is

$$x(t, \varepsilon) = \int_{-\infty}^t S(t-s)\, e^{-\varepsilon(t-s)} F(s)\, ds - \int_t^\infty U(t-s)\, e^{-\varepsilon(t-s)} F(s)\, ds$$
$$+ \int_{-\infty}^t N(t-s)\, e^{-\varepsilon(t-s)} F(s)\, ds.$$

The last term can be written in terms of its Fourier series

$$\sum_{n \text{ odd}} N^*(\varepsilon + in\omega) F_n e^{in\omega t}.$$

Since $N^*(\varepsilon + in\omega) = O(1/n)$ as $n \to \infty$ uniformly in $\varepsilon \in (0, 1]$, since $S \in L_1(0, \infty)$ and $U \in L_1(-\infty, 0)$, then we can let $\varepsilon \to 0^+$ to see that $y(t, \varepsilon)$ tends uniformly on $[0, T]$ to the function

$$y(t) = \int_{-\infty}^t S(t-s) F(s)\, ds - \int_t^\infty U(t-s) F(s)\, ds + \sum_{n \text{ odd}} N^*(in\omega) F_n e^{in\omega t}. \tag{11}$$

By continuity with respect to parameters (see, e.g., Miller [12, Chapter 2]) it follows that $y(t)$ is a T-periodic solution of (5). By the nonresonance assumption and linearity, it is the unique T-periodic solution. The formula (10) as well as the estimate $H_n = O(1/n)$ follows from (11) by elementary considerations. □

We remark that when F has a full Fourier series instead of an odd series of terms, Theorem 2 remains true under appropriate modification of the nonresonance condition (A9). Furthermore, we note that under the hypotheses of Theorem 2, if $F(t) = (0, 0, \cdots, 0, f)^T$ and if $f \in X$, then $L[y] = f$ has a unique T-periodic solution $y = Gf$ with

$$y(t) = \sum_{n \text{ odd}} f_n G_n e^{in\omega t}, \tag{12}$$

and $G_n = O(1/n)$. Indeed (12) is simply the first component of the vector equation (11).

Consider next the equation

$$Ly + f(y + a \sin(\omega t + b)) = 0, \tag{13}$$

where a and b are real constants and $a \neq 0$. We will find the following convention useful.

DEFINITION. *A map $f: R^1 \to R^1$ is said to belong to class $\mathcal{N}$ if*

(i) *it is an odd function or an odd relation*;

(ii) *if it is a function, it is required to be piecewise C^1 with discontinuities at points $\{x_j\}$*;

(iii) *if it is a relation, jump discontinuities at points $\{x_j\}$ as well as hysteresis discontinuities at ordered pairs $\{(x'_k, x''_k)\}$ are allowed* (see, e.g., Fig. 1);

(iv) *the set of points of discontinuity have no finite limit points*; *and*

(v) *there exist constants $f_0, f_1 \geqq 0$ such that $|f(y)| \leqq f_0 + f_1|y|$ for all $y \in R^1$.*

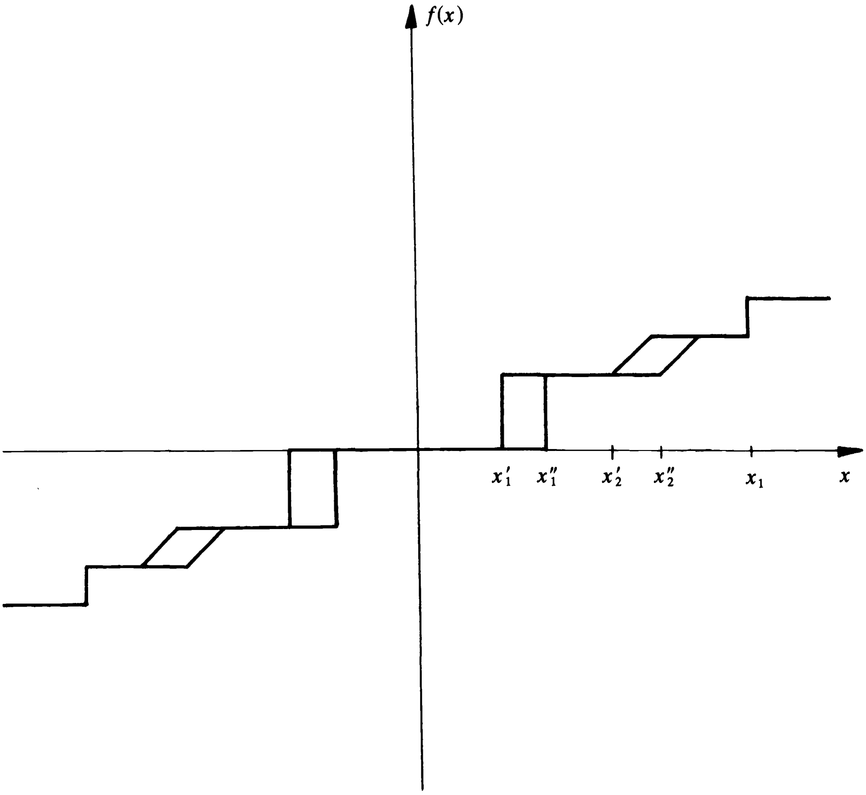


FIG. 1. *Typical member of class $\mathcal{N}$.*

In the subsequent results we will require the following assumption:

(A10) $f \in \mathcal{N}$.

Next, we define $h(t) = a \sin(\omega t + b)$, $F(\phi) = -f(\phi)$ and $x = y + h$ so that (13) assumes the form

$$L[x - h] = F(x). \tag{13'}$$

Clearly (13′) has a T-periodic solution if and only if

$$x = G(Lh) + GF(x) = h + GF(x) \tag{14}$$

has a T-periodic solution. Here G is the map determined by (12). Periodic solutions of the last equation can be obtained by using the theory in the previous section. Now we make the following additional assumption:

(A11) Let $N(A)$ be the describing function for the nonlinear term $F(x)$ and let $H(s)$ be as defined above. Assume there is a point $A = A_0 > 0$ at which the graphs of

$$y = \frac{A}{a} \quad \text{and} \quad y = |H(i\omega)|\,|H(i\omega) - N(A)|^{-1}$$

cross. (This is the usual describing function method for predicting the existence of forced periodic solutions. See [5, Chapter 3].)

Assumption (A11) ensures that the approximate system ($\hat{E}$) corresponding to system (14) has a solution of the form

$$\hat{x} = A_0 \sin(\omega t + B_0),$$

for some real constant B_0.

Next, we define a sequence $f_m(x)$ as follows. For the case where $f(x)$ is a function with discontinuities $\{x_j\}$, we can assume that $x_j < x_{j+1}$ for all j. For $m \geqq 1$ and all j we define

$$\delta_{jm} = \min\left\{\frac{(x_{j+1} - x_j)}{4}, \frac{(x_j - x_{j-1})}{4}, \frac{1}{m}\right\}.$$

Also, we define $f_m(x)$ by

$$f_m(x) = \begin{cases} f(x) & \text{if } \dfrac{(x_{j-1} + x_j)}{2} \leqq x < x_j - \delta_{jm} \\ f(x) & \text{if } (x_j + \delta_{jm}) \leqq x \leqq \dfrac{(x_{j+1} + x_j)}{2} \\ \text{linear} & \text{if } |x - x_j| \leqq \delta_{jm}. \end{cases}$$

For the case when f is a relation (with hysteresis) we modify the preceding definitions in the obvious way.

We now state and prove our last result.

THEOREM 3. *If assumptions* (A7)–(A11) *are true, if we define* $M(\varepsilon) = f_0 + f_1(\varepsilon + A_0)$ *and if there exist* ε *and* $\varepsilon_1 > 0$ *such that for all* $x \in B(\hat{x}, \varepsilon)$ *we have*

$$\|GP(F_m(x) - F(\hat{x}))\| + \Big(\sum_{\substack{n \text{ odd} \\ |n| \geqq 3}} |G_n|^2\Big)^{1/2} M(\varepsilon) \leqq \varepsilon_1 < \varepsilon, \tag{15}$$

then (13′) *has a solution* x *in the sense of Filippov such that* $x \in X$ *and such that* $\|x - \hat{x}\| \leqq \varepsilon$.

Proof. The describing function of f_m has the form

$$N_m(A) = \frac{2i}{\pi A}\int_0^{\pi} f_m(A \sin t)\, e^{-it}\, dt.$$

Thus $N_m(A) \to N(A)$ as $m \to \infty$ uniformly for A on compact subsets of $(0, \infty)$. Thus F_m satisfies condition (A11) with the point A_0 replaced by a nearby point A_m (i.e., $A_m \to A_0$

as $m \to \infty$). Thus, system $(\hat{E}_m)$ has a solution

$$\hat{x}_m(t) = \frac{A_m}{2i}[e^{i(\omega t+\theta_m)} - e^{-i(\omega t+\theta_m)}],$$

and $\hat{x}_m \to \hat{x}$ as $m \to \infty$.

Next, since the $\hat{x}_m$ are bounded, then $\{PF_m(\hat{x}_m)\}$ is a bounded sequence. Moreover, if $x \in B(\hat{x}, \varepsilon)$ we have

$$\|F_m(\phi)\| \leqq f_0 + f_1(\varepsilon + \|\hat{x}\|) \leqq f_0 + f_1(\varepsilon + A_0) = M(\varepsilon).$$

Thus, all hypotheses of Corollary 1 are satisfied and it follows that there is a subsequence $\{x_{m_j}\}$ of solutions of the approximate equations (E_{m_j}) which tend in X to a weak solution x of (E). The nature of the approximations F_m and known results for integral equations (see Filippov [4] for general background and Maeda [10] or Kiffe [9] for integral equation results) can be used to see that the weak solution is also a solution in the Filippov sense. □.

We remark that (13) can be generalized to $Ly + \int_0^\infty da_J(s) f(y(t-s)) + a \sin(\omega(t-s)+b)) = 0$ where a_J satisfies (A7) with $M = 0$. If (14) and (15) are modified to include the presence of the terms $a_J^*(in\omega)$, then Theorem 3 remains true. This extension will be needed in § IV below.

4. An Example. To demonstrate the applicability of the present results to specific cases, we consider a typical system such as the one depicted in Fig. 2 where f is a relay with deadzone as shown in Fig. 3.

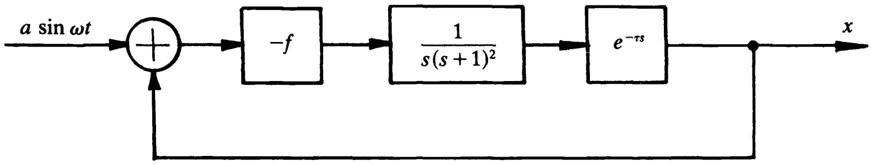


FIG. 2. *Block diagram of the example.*

The describing function for f is

$$N(A) = \frac{4D}{\pi A}\left[1 - \left(\frac{\delta}{A}\right)^2\right]^{1/2}, \quad \text{if } A > \delta.$$

This can be found in any standard table (see, e.g., the appendix in [5]). We have

$$Ly = y''' + 2y'' + y',$$

so that

$$H(s) = s(s+1)^2.$$

Since $c_j \equiv 0$, $b_{jk} = 0$ for $k > 1$ and $t_1 = 0$, assumption (A7) is satisfied. Also, assumption (A8) is satisfied with $K_1 = K_0 = 1$. The single root is $i\tau_1 = 0$ and assumption (A9) is also satisfied.

The relay depicted in Fig. 3 is clearly in the class $\mathcal{N}$. It has only two discontinuities (at the points $\eta = \pm\delta$) and it has no hysteresis. Moreover, $f_0 = D$ and $f_1 = 0$ so that $M(\varepsilon) = D$ will work.

We assume that ε and A can be chosen so that $0 < \varepsilon < \delta < \delta + \varepsilon < A$. When fixing ε and $A = A_0$ in our subsequent discussion, we will have to check that this assumption is

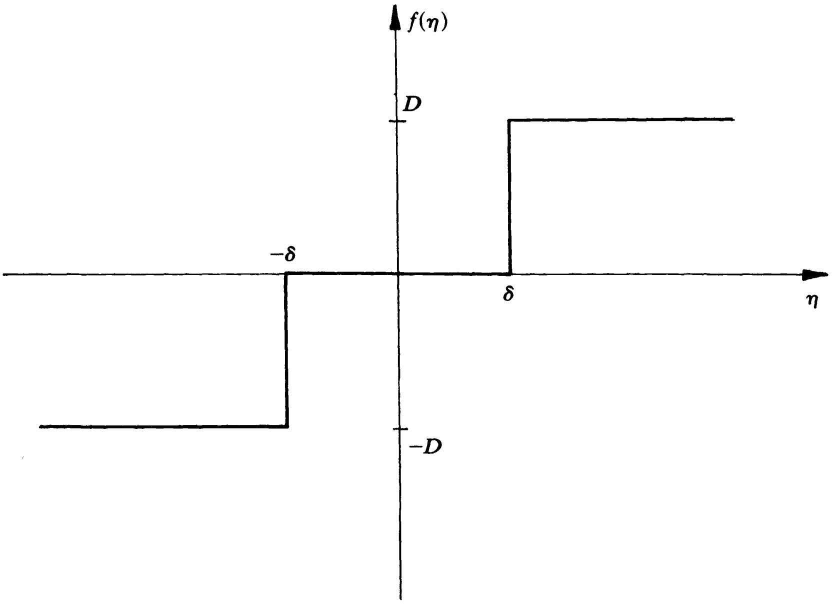


FIG. 3. *Relay with deadzone.*

indeed satisfied. Let

$$\hat{x} = A \sin \omega t, \qquad x = \hat{x} + y, \quad \text{and} \quad |y(t)| \leqq \varepsilon \text{ on } [0, T].$$

With $T = 2\pi/\omega$ we can write

$$P[F(x) - F(\hat{x})](t) = T^{-1} \int_0^T [F(x) - F(\hat{x})]\, e^{i\omega(t-\tau)}\, d\tau$$
$$+ T^{-1} \int_0^T [F(x) - F(\hat{x})]\, e^{-i\omega(t-\tau)}\, d\tau$$
$$= \alpha \cos \omega t + \beta \sin \omega t,$$

where

$$\alpha = \frac{2}{T} \int_0^T [F(x) - F(\hat{x})] \cos \omega\tau\, d\tau = \frac{1}{\pi} \int_0^{2\pi} [F(x) - F(\hat{x})] \cos s\, ds,$$

and

$$\beta = \frac{1}{\pi} \int_0^{2\pi} [F(x) - F(\hat{x})] \sin s\, ds.$$

By the change of variables used above, it is clear that without loss of generality, for the purpose of estimating the sizes of α and β, we can assume that $\omega = 1$. Let t_1, t_2 and t_3 be defined by the relations

$$A \sin t_1 = \delta - \varepsilon, \qquad A \sin t_2 = \delta, \qquad A \sin t_3 = \delta + \varepsilon.$$

Now

$$F(\hat{x})(t) = \begin{cases} 0 & \text{on } 0 < t < t_2, \\ D & \text{on } t_2 < t < \dfrac{\pi}{2}, \end{cases}$$

and $x = \hat{x} + y$, $\|y\| \leqq \varepsilon$. Thus, on $0 < t < \pi/2$, $F(x) - F(\hat{x})$ can differ from zero only in the following fashion:

(a) $[F(x) - F(\hat{x})](t) = D > 0$ for some or for all t on (t_1, t_2), and/or

(b) $[F(x) - F(\hat{x})](t) = -D < 0$ for t such that $t_2 < t < t_3$

(given the relative sizes of ε, δ, and A). Since the errors in (a) and (b) have opposite signs, they tend to cancel each other. Hence, in the worst case, only one of these errors occurs. A similar analysis applies to the intervals $j\pi/2 < t < (j+1)\pi/2$ for $j = 1, 2, 3$. In the worst case, the errors on each of these intervals will be of the same sign so that no cancellations can occur. Thus, the worst case analysis shows that the magnitude $|\beta|$ is the maximum of the two numbers

$$\begin{aligned} \frac{4}{\pi}\int_{t_1}^{t_2} D \sin t\, dt &= \frac{4D}{\pi}\left\{\sqrt{1-\left(\frac{\delta-\varepsilon}{A}\right)^2} - \sqrt{1-\left(\frac{\delta}{A}\right)^2}\right\} \\ &= \frac{4D}{\pi A}\frac{2\delta\varepsilon - \varepsilon^2}{\sqrt{A^2-(\delta-\varepsilon)^2}+\sqrt{A^2-\delta^2}}, \end{aligned} \tag{16}$$

and

$$\frac{4}{\pi}\int_{t_2}^{t_3} D \sin t\, dt = \frac{4D}{\pi A}\frac{2\varepsilon\delta+\varepsilon^2}{\sqrt{A^2-\delta^2}+\sqrt{A^2-(\delta+\varepsilon)^2}}. \tag{17}$$

Simple monotonicity arguments involving a second derivative show that the number determined by equation (17) is larger than that determined by equation (16).

In each of the four intervals $j\pi/2 < t < (j+1)\pi/2$, for $0 \leqq j \leqq 3$, the error term contributing to α is either

$$\frac{D}{\pi}\int_{t_1}^{t_2} \cos t\, dt = \frac{D\varepsilon}{\pi A},$$

or

$$\frac{-D}{\pi}\int_{t_2}^{t_3} \cos t\, dt = \frac{-D\varepsilon}{\pi A}.$$

Hence, the maximum possible error for $|\alpha|$ is $(4D\varepsilon)/(\pi A)$. Thus, if $\hat{x} = A \sin \omega t$ and $x = \hat{x} + y$, $\|y\| \leqq \varepsilon$, then $\|P(F(x) - F(\hat{x}))\|$ is at most $(\alpha^2 + \beta^2)^{1/2} \triangleq E(D, \delta, \varepsilon, A)$. By the computation above, we see that

$$E(D, \delta, \varepsilon, A) = \frac{4D}{\pi A}\left(\left[\frac{2\delta\varepsilon+\varepsilon^2}{\sqrt{A^2-\delta^2}+\sqrt{A^2-(\delta+\varepsilon)^2}}\right]^2 + \varepsilon^2\right)^{1/2}.$$

Condition (15) assumes the form

$$|G_1|E(D, \delta, \varepsilon, A) + D\left(\sum_{\substack{n \text{ odd} \\ |n| \geqq 3}} |G_n|^2\right)^{1/2} < \varepsilon. \tag{18}$$

Here $G_n = e^{i\tau n\omega}/[i\omega(i\omega+1)^2]$ so that we can compute $|G_1| = 0.5$ and $(\sum_{\substack{n \text{ odd}\\ |n|\geqq 3}} |G_n|^2)^{1/2} \cong 0.048605$.

As a specific case, let $D = a = \omega = 1$ and let $\delta = \tau = 0.1$. Then (13) assumes the form

$$y''' + 2y'' + y' + f(y(t-0.1) + \sin(t-0.1)) = 0.$$

Condition (A10) will be true if A_0 is a solution of the equation

$$\phi(A)\equiv\left(A-\frac{2\cos 0.1}{\pi}\sqrt{1-\frac{0.01}{A^2}}\right)^2+\frac{4}{\pi^2}\left(1-\frac{0.01}{A^2}\right)\sin^2 0.1-1=0.$$

This equation is easily solved numerically and yields for ϕ a unique positive root at $A_0 \cong 1.62904$. For these values of the parameters we check (18) numerically. Relation (16) was solved numerically by replacing $<$ by $=$, finding the root $\bar{\varepsilon}>0$ and then reducing $\bar{\varepsilon}$ slightly to find an ε which satisfies the inequality. (This can be accomplished by finding the root by the bisection method.) This procedure yields a nearly optimal value of $\varepsilon>0$. For example, with the present values of the parameters, $\varepsilon=0.07992$ will work. The relative error in the predicted amplitude of the periodic solution is at most

$$\frac{\varepsilon}{A_0}=4.88\%.$$

To indicate some trends, we tabulate in the following some additional cases:

$D=1$	$a=0.5$	$A_0=1.334$	$\frac{\varepsilon}{A_0}=7.01\%$
$D=1$	$a=1.0$	$A_0=1.629$	$\frac{\varepsilon}{A_0}=4.88\%$
$D=1$	$a=2.0$	$A_0=2.045$	$\frac{\varepsilon}{A_0}=3.46\%$
$D=1$	$a=3.0$	$A_0=2.632$	$\frac{\varepsilon}{A_0}=2.82\%$. □

5. Concluding remarks. At this point, some comments concerning the significance of some of the hypotheses are in order.

We first note that hypothesis (A10) could be generalized to include relations which do not satisfy the bound $|f(y)|\leqq f_0+f_1|y|$. However, a very large class of useful nonlinear elements which arise in applications may be represented by models which satisfy this inequality, frequently with $f_1=0$. We further note that hypotheses (A7) and (A8) are mild technical assumptions which are also easily satisfied in most applications.

Hypothesis (A9) constitutes the usual requirement of the describing function method that the linear part of system (13) does not admit as a solution any subharmonics. Hypothesis (A11) ensures that the approximate system $(\tilde{E})$ has a solution $\hat{x}$. The remaining hypothesis (15) can be viewed as consisting of two parts. Specifically, the first term in (15) is small when the describing function approximation is accurate for nearly sinusoidal inputs while the second term in (15) is small when the linear part of (13) determines a good low-pass filter.

All of the above hypotheses are normal, natural and expected in typical applications.

REFERENCES

[1] R. W. BASS, *Mathematical Legitimacy of Equivalent Linearization by Describing Functions*, Automatic and Remote Control, ed. J. F. Coales, London, Butterworths, 1961, pp. 895–899.

[2] A. R. BERGEN AND R. L. FRANKS, *Justification of the describing function method*, SIAM J. Control, 9, No. 4, (1971), pp. 568–589.

[3] N. DUNFORD AND J. T. SCHWARZ, *Linear Operators, Part I: General Theory*, Interscience Publishers, New York, 1958.

[4] A. F. FILIPPOV, *Differential equations with discontinuous right hand sides* (in Russian) Mat. Sb., 5, No. 1 (1960), pp. 99–128. (English Translation: American Math. Society Transl., Vol. 42 (1964), pp. 199–231.

[5] A. GELB AND W. E. VANDER VELDE, *Multiple-Input Describing Functions and Nonlinear System Design*, McGraw-Hill Book Co., New York, 1968.

[6] S. I. GROSSMAN AND R. K. MILLER, *Perturbation theory for Volterra integrodifferential systems*, J. Differential Equations, 8 (1970), pp. 457–474.

[7] J. M. HOLTZMAN, *Contraction maps and equivalent linearization*, Bell System Tech. J. 46, (1967), pp. 2405–2435.

[8] G. S. JORDAN AND R. L. WHEELER, *Structure of Volterra integral and integrodifferential systems*, SIAM J. Math. Anal., to appear (see especially Theorem 2.2).

[9] T. KIFFE, *A discontinuous Volterra integral equation*, J. Integral Equations, in press.

[10] HAJIME MAEDA, *Stability considerations for a Volterra integral equation with discontinuous nonlinearity*, SIAM J. Control, 11 (1973), pp. 202–204.

[11] A. I. MEES AND A. R. BERGEN, *Describing functions revisited*, IEEE Trans. Automatic Control, 20 (1975), pp. 473–478.

[12] R. K. MILLER, *Nonlinear Volterra Integral Equations*, W. A. Benjamin, Menlo Park, California, 1971.

[13] ———, *Structure of solutions of unstable linear Volterra integrodifferential equations*, J. Differential Equations, 15 (1974), pp. 129–157.

[14] V. M. POPOV, *Absolute stability of nonlinear systems of automatic control*, Automat. and Remote Control, 22 (1961), pp. 857–875.

[15] I. W. SANDBERG, *On the response of nonlinear control systems to periodic input signals*, Bell System Tech. J., 43 (1964), pp. 911–926.

[16] V. I. UTKIN, *Variable structure systems with sliding modes*, IEEE Trans. Automatic Control, 22, (1977), pp. 212–222.

[17] D. WILLIAMSON, *Describing function analysis and oscillation in nonlinear networks*, Internat. J. Control, 24, (1976), pp. 283–296.

[18] J. K. HALE, *Oscillation in Nonlinear Systems*, McGraw-Hill, New York, 1963.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 6, November 1980

0363-0129/80/1806-0001 $01.00/0

A CHARACTERIZATION OF THE REACHABLE SET FOR NONLINEAR CONTROL SYSTEMS*

RICHARD VINTER†

Abstract. The question of whether a set is reachable by a nonlinear control system is answered in terms of the properties of a convex optimization problem. The set is reachable or not according to whether the value of the optimization problem is zero or infinity. Our findings strengthen earlier sufficient conditions for a point not to be reachable, given in terms of Lyapunov-like functions, in that we assure that the functions exist. Our approach is to embed admissible trajectories in a space of measures, and to apply recently obtained results on the properties of measures arising in this way.

1. Introduction. In this paper we provide a characterization of the set of points reachable from a point (x_0, t_0) along solutions of the differential equation with control

$$\dot{x}(t)=f(x(t), t, u(t)). \tag{1.1}$$

Our main result is that, under very mild conditions, if no point in a closed set Γ is reachable from (x_0, t_0), then there exists a continuously differentiable function $\phi(\cdot,\cdot)$ which satisfies the partial differential inequality

$$\phi_t(x, t)+\phi_x(x, t)f(x, t, u)\leqq 0, \tag{1.2}$$

($\phi_t = \partial\phi/\partial t$ etc.), and which takes positive values on Γ and nonpositive values on points which are reachable from (x_0, t_0). A function with these properties may always be obtained from a maximizing sequence for a convex optimization problem.

Our methods are grounded in an idea originally due to L. C. Young, that "trajectories" can be embedded in a space of linear functionals [17]. We introduce a notion of "weak" reachability which turns out to be equivalent to reachability as conventionally defined; weak reachability concerns existence of a linear functional satisfying a certain "convex" constraint. We can exploit this convexity, and obtain reachability criteria using the methods of convex analysis.

Our characterization is somewhat in the spirit of a theorem of Carathéodory [3]: a Pfaffian is integrable at a point if every neighborhood of the point contains an inaccessible point.[1] The inequality (1.2) replaces the Pfaffian identified with the system of partial differential equations expressing the dynamical constraint (1.1) on the trajectories, and our function $\phi(\cdot,\cdot)$ replaces the "complete integral" (which is nonzero on locally inaccessible points and zero on locally accessible ones). Of course, our results hold under much milder conditions than would be needed to make this parallel precise [7].

There are points of contact also with the Control Theory literature. Concerning nonlinear systems, sufficient conditions for either reachability or nonreachability are known, which involve functions similar to our $\phi(\cdot,\cdot)$ function (see for example [6], [14]). But we show that such $\phi(\cdot,\cdot)$'s always exist, an apparently new result. It is well known that, for some *linear* systems, a necessary and sufficient condition that a convex set in the output space be reachable at some fixed time can be given in terms of the value

* Received by the editors June 15, 1979, and in final revised form March 17, 1980.

† Imperial College of Science and Technology, London SW7, England. (This research was carried out in part while the author was visiting the Laboratory of Information and Decision Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts, 02139.)

[1] I am indebted to R. W. Brockett for this observation.

of a convex optimization problem (over normals to hyperplanes in the output space) ([2], [4], [5], [12]). We too relate reachability to properties of a convex optimization problem, but do so in a much more general context and use very different methods from those in the standard linear theory.

Optimality conditions for nonlinear optimal control problems of a similar flavor to our reachability conditions, in that they involve $\phi(\cdot,\cdot)$'s which satisfy a partial differential inequality, are given in [8], [10], [16].

2. Preliminaries. Let D be a compact subset of $\mathbb{R}^k$. We denote by $C(D)$ the usual linear space of continuous real-valued functions on D, equipped with the supremum norm.

The topological dual of $C(D)$ is written $C^*(D)$. The dual space assumes the dual norm.

All norms are written $\|\cdot\|$; sometimes for clarity the norm in question is identified by a subscript, e.g. $\|\cdot\|_{C(D)}$.

The subset $P^{\oplus}(D)$ of $C^*(D)$ comprises bounded linear functionals which take nonnegative values on the subset

$$P(D)=\{g\in C(D)\colon g(d)\geqq 0 \text{ for all } d\in D\},$$

that is, the usual "closed positive cone" in $C(D)$.

We remark that for elements $\mu\in P^{\oplus}(\Omega)$, the dual norm of μ is given by

$$\|\mu\|_{C^*(D)}=\mu(1).$$

Here "1" denotes the function on D which takes value 1 everywhere.

When D is a (closed) cube in $\mathbb{R}^k$, we define $C^1(D)(C^\infty(D))$ to be the usual linear space of continuously differentiable (infinitely differentiable) functions on $\mathbb{R}^k$, restricted to D.

The support of an element $g\in C(D)$, supp $\{g\}$, is the closure of the subset of D on which g takes nonzero values. We shall use also the notion of the support, supp* $\{\mu\}$, of an element μ in $C^*(D)$. The set supp* $\{\mu\}$ is the complement in D of the union of all relatively open sets $\mathcal{O}$ of D such that supp $\{g\}\subset\mathcal{O}$ implies $\mu(g)=0$. This definition of supp* $\{\mu\}$ is easily shown to accord with the usual definition of the "support" of the signed Radon measure which represents μ.

3. The reachable set. Let $A\subset\mathbb{R}^{n+1}$, $\Omega\subset\mathbb{R}^m$ be compact sets. The set S is a cube in $\mathbb{R}^{n+1}$ such that $A\subset$ interior $\{S\}$. We consider the differential equation with control

$$\dot{x}(t)=f(x(t),t,u(t)). \tag{3.1}$$

The $\mathbb{R}^n$-valued function $f(x,t,u)$ is continuous in its argument $(x,t,u)\in\mathbb{R}^n\times\mathbb{R}\times\mathbb{R}^m$ and satisfies

$$\{f(x,t,u)\colon u\in\Omega\} \text{ is convex for each } (x,t)\in A. \tag{3.2}$$

An absolutely continuous $\mathbb{R}^n$-valued function $\{x(t)\colon t_0\leqq t\leqq t_1\}$ is a *trajectory* when its graph lies in A and when, for almost every $t\in[t_0,t_1]$, it satisfies (3.1) in which $\{u(t)\colon t_0\leqq t\leqq t_1\}$ is some measurable function which takes values almost everywhere in Ω. We say the trajectory *emanates* from $(x(t_0),t_0)$. The function $\{u(t)\colon t_0\leqq t\leqq t_1\}$ is called a control associated with the trajectory. We call $(x(t_1),t_1)$ the *endpoint* of the trajectory.

Given $(x_0,t_0)\in A$, we define the reachable set emanating from (x_0,t_0), $\mathcal{R}_{x_0,t_0}$, to be the subset of A:

$$\mathcal{R}_{x_0,t_0}=\{(x(t_1),t_1)\colon x(t),\ t_0\leqq t\leqq t_1 \text{ is a trajectory which emanates from } (x_0,t_0)\}.$$

Points in $\mathcal{R}_{x_0,t_0}$ are said to be *reachable from* (x_0, t_0). If a set $\Gamma \subset A$ contains a reachable point, then we say that Γ *is reachable from* (x_0, t_0).

We shall be concerned with characterizing points lying in $\mathcal{R}_{x_0,t_0}$, in a manner which does not involve examining solutions of the differential equation (3.1) for every choice of control.

The convexity assumption (3.2) has been introduced to simplify the presentation; all the results to follow apply when the convexity assumption is dropped, provided the definition of "trajectory" is modified to permit the associated controls to be "relaxed" controls [11].

4. A necessary condition for reachability. The starting point for our characterization is the following simple observation. Suppose that (x_1, t_1) is reachable from (x_0, t_0). This means that there exists a trajectory $\{x(t)\colon t_0 \leqq t \leqq t_1\}$ which emanates from (x_0, t_0) and has endpoint (x_1, t_1). Let $u(\cdot)$ be an associated control. Then for any function $\phi \in C^1(S)$, the function $t \mapsto \phi(x(t), t)$ is Lipschitz continuous and can be expressed as the "integral of its derivative,"

$$\begin{aligned}\phi(x_1, t_1) - \phi(x_0, t_0) &= \int_{t_0}^{t_1} \frac{d}{dt}\phi(x(t), t)\, dt\\ &= \int_{t_0}^{t_1} (\phi_t + \phi_x f)(x(t), t, u(t))\, dt.\end{aligned}$$

(Here, and subsequently, we use $(\phi_t + \phi_x f)(x, t, u)$ as a shorthand for $\phi_t(x, t) + \phi_x(x, t) f(x, t, u)$). If ϕ satisfies

$$(4.1) \qquad (\phi_t + \phi_x f)(x, t, u) \leqq 0$$

for all $(x, t, u) \in A \times \Omega$, then the integrand is nonpositive. It follows that $\phi\ (x_1, t_1) - \phi(x_0, t_0) \leqq 0$. We have proved the following proposition.

PROPOSITION 4.1. *If* (x_1, t_1) *is reachable from* (x_0, t_0), *then for all* $\phi \in C^1(S)$ *satisfying the inequality* (4.1), *we must have* $\phi(x_1, t_1) - \phi(x_0, t_0) \leqq 0$.

In this paper we shall be principally concerned with proving a converse of this proposition:

If (x_1, t_1) is not reachable from (x_0, t_0) then there exists $\phi \in C^1(S)$ satisfying inequality (4.1) such that $\phi(x_1, t_1) - \phi(x_0, t_0) > 0$. Equivalently, in contrapositive form, if for all $\phi \in C^1(S)$ satisfying inequality (4.1) we have $\phi(x_1, t_1) - \phi(x_0, t_0) \leqq 0$, then (x_1, t_1) is reachable from (x_0, t_0).

Actually, we shall prove something rather stronger than this in two respects. In the first place, we shall provide an analogous characterization of closed sets (and not merely points) which are not reachable from a point (x_0, t_0). In the second, we shall show that the characterization may be achieved through $C^\infty(S)$ functions rather than $C^1(S)$ functions.

5. A characterization of the reachable set. Let Γ be a closed subset of A, and let (x_0, t_0) be a point in A. Our results are conveniently stated in terms of the properties of the optimization problem, $P_{\Gamma,(x_0,t_0)}$:

Maximize $\eta(\phi)$ over $\phi \in C^\infty(S)$ which satisfy

$$(5.1) \qquad (\phi_t + \phi_x f)(x, t, u) \leqq 0 \text{ for all } (x, t, u) \in A \times \Omega.$$

Here, the function $\eta(\cdot)$ is taken to be

$$(5.2) \qquad \min_{(x,t)\in\Gamma} \phi(x_1, t_1) - \phi(x_0, t_0), \qquad \phi \in C^\infty(S).$$

Any $\phi \in C^\infty(S)$ which satisfies (5.1) will be called *feasible.* We write $\sup (P_{\Gamma,(x_0,t_0)})$ for the supremum of values assumed by $\eta(\cdot)$ on feasible ϕ's. When the supremum is achieved, $\sup (P_{\Gamma,(x_0,t_0)})$ is written $\max (P_{\Gamma,(x_0,t_0)})$.

Our main result is the following.

THEOREM 5.1. *The closed subset $\Gamma \subset A$ is reachable from (x_0, t_0) if and only if*

$$\max (P_{\Gamma,(x_0,t_0)}) = 0.$$

It is now a simple matter to show

COROLLARY 5.1. *The closed subset $\Gamma \subset A$ is not reachable from (x_0, t_0) if and only if*

$$\sup (P_{\Gamma,(x_0,t_0)}) = +\infty.$$

Proof. We write P for $P_{\Gamma,(x_0,t_0)}$. If $\sup (P) = +\infty$, then no point in Γ is reachable from (x_0, t_0), by Theorem 5.1. On the other hand, suppose that $\sup (P) < +\infty$. By positive homogeneity of the objective function $\eta(\cdot)$ and of the constraint (5.1) in problem (P), we have that $\sup (P) = 0$. But then $\max (P) = 0$, since $\eta(\cdot)$ achieves the value zero on the feasible function $\phi(\cdot, \cdot)$ which is identically zero. It follows now from Theorem 5.1 that Γ is reachable from (x_0, t_0).

The substance of this corollary is that, if the closed subset $\Gamma \subset A$ is disjoint from the reachable set, $\mathcal{R}_{x_0,t_0}$, then an infinitely differentiable $\phi(\cdot, \cdot)$, satisfying (5.1), may be chosen so that the values of $\phi(x, t) - \phi(x_0, t_0)$ are bounded below by some arbitrarily large number, as (x, t) ranges over Γ. (This is the converse to Proposition (4.1), with refinements, alluded to in § 4). Is a strengthening of our result possible, in which a single ϕ serves to characterize all points which are not reachable, rather than merely any closed subset of such points? The answer to this question is not known, but we are able to characterize all points which are not reachable through a sequence of ϕ's.

COROLLARY 5.2. *There exists a sequence $\{\phi^i(\cdot, \cdot)\}$ of feasible elements such that*

$$\lim_i \phi^i(x, t) - \phi^i(x_0, t_0) = +\infty, \tag{5.3}$$

for each $(x, t) \notin \mathcal{R}_{x_0,t_0}$, and

$$\limsup_i \phi^i(x, t) - \phi^i(x_0, t_0) \leqq 0, \tag{5.4}$$

for all $(x, t) \in \mathcal{R}_{x_0,t_0}$.

Proof. We write $\mathcal{R}$ for $\mathcal{R}_{x_0,t_0}$. Consider the family of neighborhoods $\mathcal{R}_i$ of $\mathcal{R}$ in A:

$$\mathcal{R}_i = \{x : \|x - x_1\| < i^{-1}, x_1 \in \mathcal{R}\} \cap A, \qquad i = 1, 2, \cdots.$$

(The norm here is the Euclidean norm.) The set $\mathcal{R}_i$ is relatively open, whence $\mathcal{R}_i^C$, the complement of $\mathcal{R}_i$ in A, is closed. Clearly $\mathcal{R}$ and $\mathcal{R}_i^C$ are disjoint for each i. By Corollary 5.1, then, we can choose some feasible $\phi^i \in C^\infty(S)$ such that

$$\phi^i(x, t) - \phi^i(x_0, t_0) > i \quad \text{for all } (x, t) \in \mathcal{R}_i^C, \quad i = 1, 2 \cdots.$$

The sequence of sets $\{\mathcal{R}_i^C\}$ is monotone; it follows that $\phi^i(x, t) - \phi^i(x_0, t_0) \to \infty$, as $i \to \infty$, for any $(x, t) \in \bigcup_i \mathcal{R}_i^C$. But the set $\mathcal{R}$ is closed (see, e.g., [11]), whence $\bigcup_i \mathcal{R}_i^C = \mathcal{R}^C$, where $\mathcal{R}^C$ is the complement of $\mathcal{R}$ in A. We have shown that the sequence $\{\phi^i(\cdot, \cdot)\}$ satisfies (5.3) on the complement of $\mathcal{R}$. The sequence satisfies (5.4) on $\mathcal{R}$, by Theorem 5.1, since each ϕ^i is feasible.

6. Weak reachability. In the proof of Theorem 5.1, to follow in § 7, "weak reachability" will have a crucial role. In this section we introduce this concept and relate it to reachability.

Suppose that the closed subset $\Gamma \subset A$ is reachable from (x_0, t_0). Then there exists a trajectory $\{x(t): t_0 \leqq t \leqq t_1\}$ emanating from (x_0, t_0) and with endpoint (x_1, t_1) in Γ. Let $\{u(t): t_0 \leqq t \leqq t_1\}$ be an associated control.

The pair of functions $\{x(t), u(t): t_0 \leqq t \leqq t_1\}$ defines a linear functional μ on $C(S \times \Omega)$:

$$\mu(g) = \int_{t_0}^{t_1} g(x(t), t, u(t))\, dt, \qquad g \in C(S \times \Omega).$$

Notice that $\mu \in P^{\oplus}(S \times \Omega)$, since $\mu(g)$ is bounded by $(t_1 - t_0) \cdot \|g\|_{C(s \times \Omega)}$ for $g \in C(S \times \Omega)$, and since $\mu(g)$ is nonpositive valued for $g \in P(S \times \Omega)$.

We also have that

$$\mu(\phi_t + \phi_x f) = \phi(x_1, t_1) - \phi(x_0, t_0), \tag{6.1}$$

for all $\phi \in C^{\infty}(S)$. This is true since

$$\mu(\phi_t + \phi_x f) = \int_{t_0}^{t_1} (\phi_t + \phi_x f)(x(t), t, u(t))\, dt$$

$$= \int_{t_0}^{t_1} \frac{d}{dt}[\phi(x(t), t)]\, dt = \phi(x_1, t_1) - \phi(x_0, t_0).$$

Equation (6.1) may be written

$$\mu(\phi_t + \phi_x f) = \beta(\phi) - \phi(x_0, t_0), \tag{6.2}$$

for all $\phi \in C^{\infty}(S)$. In (6.2), β is the evaluation map at the point (x_1, t_1); that is, β is defined by

$$\beta(\phi) = \phi(x_1, t_1), \qquad \phi \in C(S).$$

It is clear that β is an element in the subset $P^{\oplus}(S)$. We note also that

$$\|\beta\| = 1, \tag{6.3}$$

and

$$\text{supp}^* \{\beta\} \subset \Gamma, \tag{6.4}$$

(the support of β, supp* $\{\beta\}$, is defined in § 2).

Now consider the function $d_A(x, t, u)$ on $A \times \Omega$,

$$d_A(x, t, u) = \min_{(y,s) \in A} \|(x, t) - (y, s)\|.$$

The function d_A is continuous, and so may be viewed as an element in $C(A \times \Omega)$. We observe that

$$\mu(d_A) = 0. \tag{6.5}$$

This is true, since $(x(t), t) \in A$, $t_0 \leqq t \leqq t_1$, from which it follows that $\mu(d_A) = \int_{t_0}^{t_1} d_A(x(t), t, u(t)\, dt = 0$.

The properties (6.2)–(6.5) prompt the following definition: The closed subset $\Gamma \subset A$ is *weakly reachable from* (x_0, t_0) if there exist $\mu \in P^{\oplus}(S \times \Omega)$ and $\beta \in P^{\oplus}(S)$ such that (6.2)–(6.5) are satisfied.

What do the linear functionals μ which "reach" Γ (that is, μ's which satisfy (6.2)–(6.5) for some β) look like? We have already exhibited one. But such examples do not exhaust the possibilities. To illustrate this we consider $\mu = \sum_{i=1}^{k} \alpha_i \mu_i$, a convex combination of μ's arising from trajectories $x_i(\cdot)$ emanating from (x_0, t_0), with associated controls $u_i(\cdot)$ and having endpoints (x_i, t_i) in Γ:

$$\mu(g) = \sum_{i=1}^{k} \alpha_i \int_{t_0}^{t_1} g(x_i(t), t, u_i(t)\, dt.$$

The functional μ lies in $P^{\oplus}(S \times \Omega)$ and satisfies (6.2)–(6.5) when β is taken as

$$\beta = \sum_i \alpha_i \cdot \delta(x_i, t_i),$$

in which $\delta(x_i, t_i)$ means the evaluation map at the point (x_i, t_i). However μ does not correspond to any single trajectory (unless all the μ_i's with nonzero weights, α_i, are the same). Thus μ's which satisfy (6.2)–(6.5), for some β, can be rather complicated. The following result is therefore somewhat unexpected:

LEMMA. *The closed subset $\Gamma \subset A$ is weakly reachable from (x_0, t_0) if and only if Γ is reachable from (x_0, t_0).*

Proof. We have already shown that, if the set Γ is reachable from (x_0, t_0), then Γ is weakly reachable from (x_0, t_0).

Now let Γ be weakly reachable from (x_0, t_0). Then there exist $\mu \in P^{\oplus}(S \times \Omega)$ and $\beta \in P^{\oplus}(S)$ which satisfy (6.2)–(6.5). Suppose that (6.2) holds for all $\phi \in C^1(S)$ (not merely $C^{\infty}(S)$). With this assumption, the reachability of Γ follows as a byproduct of the proof of [15, Theorem 5.1]. Indeed μ is feasible for the "weak" problem introduced in [15]. As such, μ defines a feasible element $\tilde{\mu}$ for the "parametric" problem of [15]. We may associate with $\tilde{\mu}$ an "admissible" trajectory $\{x(t): t_0 \leqq t \leqq t_1\}$ corresponding to some "relaxed" control. But we have assumed convexity of the velocity set $\{f(x, t, u): u \in \Omega\}$ for each $(x, t) \in A$; $x(\cdot)$ is therefore also associated with an "ordinary" control, and defines a trajectory (in the sense understood here) emanating from (x_0, t_0) and having endpoint in Γ (see [15, p. 511]).

We remark that here A is permitted to be a general compact subset of $\mathbb{R}^{n+1}$, whereas in [15] it was taken as a cylinder set. However all the results and proofs in [15] carry over to admit our greater generality with only trivial modifications.

It remains to show that (6.2) holds for all $\phi \in C^1(S)$, if it holds for all $\phi \in C^{\infty}(S)$. Take any $\bar{\phi} \in C^1(S)$. The function $\bar{\phi}$ is the restriction of some continuously differentiable function ψ on $\mathbb{R}^{n+1}$ to S (this property is built in to the definition of $C^1(S)$). Now apply Theorems 1.7 and 1.8 of [1, p. 5 et seq.] to the function ψ; we obtain a sequence of functions $\{\phi^i\}$ in $C^{\infty}(S)$ with the properties

$$(\phi_t^i + \phi_x^i f)(x, t, u) \to (\psi_t + \psi_x f)(x, t, u),$$

$$\phi^i(x, t) \to \psi(x, t),$$

as $i \to \infty$, uniformly over $(x, t, u) \in S \times \Omega$.

But (6.2) is satisfied along the sequence $\{\phi^i\}$ of infinitely differentiable functions. Since the functions μ and β are continuous, and since the functions ψ, $\psi_t + \phi_x f$ coincide with the functions $\bar{\phi}$, $\bar{\phi}_t + \bar{\phi}_x f$ on $A \times \Omega$, we have that

$$\mu(\bar{\phi}_t + \bar{\phi}_x f) = \beta(\bar{\phi}) - \bar{\phi}(x_0, t_0).$$

We have shown that we may indeed suppose that (6.2) holds for all $\phi \in C^1(S)$.

7. Proof of the main result.

Proof of Theorem 5.1. Let Γ be a compact subset of A, and let (x_0, t_0) be a point in A. We shall abbreviate problem $(P_{\Gamma,(x_0,t_0)})$ to (P).

Suppose that the point (x_1, t_1) in Γ is reachable from (x_0, t_0). We have already shown (Proposition 4.1) that, for any feasible ϕ,

$$\phi(x_1, t_1) - \phi(x_0, t_0) \leqq 0.$$

Certainly, then, we have

$$\min_{(x,t)\in\Gamma} \phi(x, t) - \phi(x_0, t_0) \leqq \phi(x_1, t_1) - \phi(x_0, t_0) \leqq 0.$$

Thus sup $(P) \leqq 0$. But then max $(P) = 0$, because $\phi = 0$ is feasible.

We now prove the converse statement. Suppose that

$$\max(P) = 0.$$

Define the transformation from the linear space $C^\infty(S)$ to the normed space $C(A \times \Omega)$ as follows:

$$(G(\phi))(x, t, u) = (\phi_t + \phi_x t)(x, t, u) \quad \text{for all } (x, t, u) \in A \times \Omega.$$

The functional $\eta(\cdot)$ on $C^\infty(S)$ is defined (as before):

$$\eta(\phi) = \min_{(x,t)\in\Gamma} \phi(x, t) - \phi(x_0, t_0).$$

Problem (P) may be expressed: Maximize $\eta(\phi)$ over $\phi \in C^\infty(S)$ which satisfy the constraint $G(\phi) \in P(A \times \Omega)$, ($P(A \times \Omega)$ is the "positive cone" defined in § 2). Now

(i) the function G is linear, and the function η is concave on $C^1(S)$;

(ii) there exists $\phi \in C^1(S)$ such that

$$-G(\phi) \in \text{interior}\{P(A \times \Omega)\}, \qquad (\phi\ (x, t) = -t \text{ will do});$$

(iii) max $(P) < \infty$.

It is well known (see, for example, [13, p. 47]) that, in consequence of properties (i)–(iii), there exists

$$\bar{\Lambda} \in P^{\oplus}(A \times \Omega)$$

such that

$$\max_{\phi} \{\eta(\phi) - \bar{\Lambda}(G(\phi))\} = \max(P)(=0). \tag{7.1}$$

We may associate with $\bar{\Lambda}$ an element $\Lambda \subset P^{\oplus}(S \times \Omega)$ defined as follows:

$$\Lambda(g) = \bar{\Lambda}(\tilde{g}) \quad \text{for } g \in C(S \times \Omega), \tag{7.2}$$

where $\tilde{g}$ is the restriction of g to $A \times \Omega$.

It follows from the definition of Λ that

$$\bar{\Lambda}(G(\phi)) = \Lambda(\phi_t + \phi_x t) \quad \text{for } \phi \in C^1(S).$$

Equation (7.1) implies, then,

$$\min_{(x,t)\in\Gamma} \phi(x, t) - \phi(x_0, t_0) - \Lambda(\phi_t + \phi_x t) \leqq 0 \tag{7.3}$$

for all $\phi \in C^\infty(S)$.

We proceed to show that existence of $\Lambda \subset P^{\oplus}(S \times \Omega)$ which satisfies (7.3) implies that Γ is weakly reachable from (x_0, t_0). We shall require therefore (see § 6) that

$$\Lambda(d_A) = 0, \tag{7.4}$$

and that there exists some $\beta \in P^{\oplus}(A)$ with support in Γ and of unit norm, such that

$$\Lambda(\phi_t+\phi_x f)=\beta(\phi)-\phi(x_0, t_0) \tag{7.5}$$

for all $\phi \in C^{\infty}(S)$. Equation (7.4) follows from (7.2) since the restriction of d_A to $A\times\Omega$ is the zero function.

It remains then to construct the functional β. Equation (7.5) suggests that we should examine the linear functional

$$\phi \mapsto \Lambda(\phi_t+\phi_x f)+\phi(x_0, t_0) \tag{7.6}$$

on $C^{\infty}(S)$, and show that it extends to a *continuous* linear functional on $C(S)$; the extension may then be taken as defining β. Since $C^{\infty}(S)$ is dense in $C(S)$, this amounts to finding a number K such that

$$|\Lambda(\phi_t+\phi_x f)+\phi(x_0, t_0)|\leqq K\cdot \max_{(x,t)\in S}|\phi(x, t)| \tag{7.7}$$

for all $\phi \in C^{\infty}(S)$.

By inequality (7.3), we have for any $\phi \in C^{\infty}(S)$,

$$\begin{aligned}\Lambda(\phi_t+\phi_x f)+\phi(x_0, t_0)&=-\Lambda(-\phi_t-\phi_x f)-(-\phi(x_0, t_0))\\ &\leqq -\min_{(x,t)\in\Gamma}\{-\phi(x, t)\}.\end{aligned}$$

For all $\phi \in C^{\infty}(S)$ then,

$$\Lambda(\phi_t+\phi_x f)+\phi(x_0, t_0)\leqq \max_{(x,t)\in\Gamma}\phi(x, t). \tag{7.8}$$

It follows that

$$\Lambda(\phi_t+\phi_x f)+\phi(x_0, t_0)\leqq \|\phi\|_{C(S)},$$

for all $\phi \in C^{\infty}(S)$. But the functional (7.6) is linear; we have therefore

$$|\Lambda(\phi_t+\phi_x f)+\phi(x_0, t_0)|\leqq \|\phi\|_{C(S)},$$

for all $\phi \in C^{\infty}(S)$. Inequality (7.7) has been verified when $K=1$.

We have shown that the continuous linear functional $\beta \in C^*(S)$ is indeed well defined as the continuous extension to $C(S)$ of the linear functional (7.6).

We now check that β has the required properties. The functional β of course satisfies (7.5) since it is an extension of the map (7.6).

Inequality (7.3) implies that

$$\beta(\phi)\geqq \min_{(x,t)\in\Gamma}\phi(x, t), \tag{7.9}$$

for all $\phi \in C^{\infty}(S)$ and therefore for all $\phi \in C(S)$, by density of $C^{\infty}(S)$ in $C(S)$ and by continuity of the functions involved. We conclude that β lies in the subspace $P^{\oplus}(S)$ of $C(S)$. Likewise, (7.8) implies that

$$\|\beta\|\leqq 1. \tag{7.10}$$

However taking $\phi=1$ in (7.9) we see that

$$\|\beta\|=\beta(1)\geqq 1.$$

It follows now from (7.10) that

$$\|\beta\|=1.$$

We deduce from (7.8) that

$$\beta(\phi) \leqq \max_{(x,t)\in\Gamma} \phi(x, t),$$

for all $\phi \in C(S)$. If then $\phi \in C(S)$ has support in the open set $S\backslash\Gamma$, this inequality, together with (7.9), implies that $\beta(\phi) = 0$. In other words, the function β has support in Γ. This completes verification that β has the necessary properties which ensure that Γ is weakly reachable from (x_0, t_0). Γ then is reachable from (x_0, t_0) by Lemma (6.1).

8. A related characterization of the reachable set. Suppose that the point (x_1, t_1) is not reachable from (x_0, t_0). It has been shown by abstract methods that, in this case, there exists some $\phi \in C^\infty(S)$ which satisfies

$$(\phi_t + \phi_x t)(x, t, u) \leqq 0 \tag{8.1}$$

for all $(x, t, u) \in A \times \Omega$, and

$$\phi(x_1, t_1) - \phi(x_0, t_0) > 0. \tag{8.2}$$

It is tempting to try to exhibit such a function ϕ directly. It is far from clear how this should be done in a general context. However we can always find a function $\bar{\phi}$ which satisfies rather weaker properties. Notice that if ϕ satisfies (8.1), then it also satisfies the condition:

(8.3) For any trajectory $x(\cdot)$, $t \to \phi(x(t), t)$ is a monotone nonincreasing function.

The following construction of a function $\bar{\phi}$ which satisfies (8.2) and (8.3) (rather than (8.2) and the stronger condition (8.1)) is due to David Allwright.

Corresponding to $(x, t) \in A$, we define the set

$$\mathcal{B}_{x,t} = \{(\bar{x}, \bar{t}) \in A: (x, t) \in \mathcal{R}_{\bar{x},\bar{t}}\}.$$

Thus, $\mathcal{B}_{x,t}$ is the set of points reachable from (x, t) in "reverse time." Now take

$$\bar{\phi}(x, t) = \min \{\|(x_0, t_0) - y\|: y \in \mathcal{B}_{x,t}\}, \tag{8.4}$$

for all $(x, t) \in A$.

Obviously, if $x(s)$, $\bar{t} \leqq s \leqq t$, is a trajectory, then $\mathcal{B}_{x(\bar{t}),\bar{t}} \subset \mathcal{B}_{x(t),t}$. It follows from (8.4) then that $\bar{\phi}$ satisfies condition (8.3). Now (x_1, t_1) is not reachable from (x_0, t_0). This means that $(x_0, t_0) \notin \mathcal{B}_{x_1,t_1}$. But under our assumptions $\mathcal{B}_{x_1,t_1}$ is closed; it follows that $\bar{\phi}(x_1, t_1) > 0$. Note also that $(x_0, t_0) \in \mathcal{B}_{x_0,t_0}$, so that $\bar{\phi}(x_0, t_0) = 0$. We have shown that

$$\bar{\phi}(x_1, t_1) - \bar{\phi}(x_0, t_0) > 0.$$

On the other hand, if there exists a function $\bar{\phi}$ on A which satisfies condition (8.3) and also

$$\bar{\phi}(\bar{x}, \bar{t}) - \bar{\phi}(x_0, t_0) > 0,$$

for some $(\bar{x}, \bar{t}) \in A$, then $(\bar{x}, \bar{t})$ is obviously not reachable. We have proved:

PROPOSITION 8.1. *The point (x_1, t_1) is not reachable if and only if there exists a function $\bar{\phi}$ on A satisfying condition* (8.3), *and such that*

$$\bar{\phi}(x_1, t_1) - \bar{\phi}(x_0, t_0) > 0.$$

The function $\bar{\phi}$ may always be chosen as (8.4).

We remark that the function $\bar{\phi}$ given by (8.4) need not be continuous, since there is no guarantee that the multifunction which carries points (x, t) into the subsets $\mathcal{B}_{x,t}$ (equipped with the Hausdorff metric) is continuous.

In one respect, Proposition 8.1 is superior to Theorem 5.1: in Proposition 8.1 a single $\bar{\phi}$ serves to characterize all points which are not reachable from (x_0, t_0). Theorem 5.1 associates a ϕ function with merely every closed subset of points which are not reachable from (x_0, t_0).

Theorem 5.1, however, provides a far tighter test for nonreachability than Proposition 8.1. The theorem tells us that we can confine the search for ϕ's which satisfy (8.2) to C^∞ functions subject to a simple pointwise constraint on the derivatives (8.1). By contrast, application of Proposition 8.1 involves a search over general functions, and ones which satisfy an awkward constraint, condition (8.3), involving as it does certain properties of the totality of trajectories and therefore the solution to the differential equation for all controls and all initial conditions.

The fact that Theorem 5.1 does not associate a single ϕ with all unreachable points we may view as the price paid for these advantages.

In the present context, we see the function $\bar{\phi}$ is of interest principally for suggesting possible condidates for the function ϕ, whose existence is guaranteed by Theorem 5.1.

9. An autonomous version of Theorem 5.1: a counterexample. Consider now the situation when the right-hand side of the differential equation does not depend on time: thus

$$\dot{x}(t)=f(x(t), u(t)),$$

and when A is the cylinder set $Q \times [0, T]$, with Q some compact subset of $\mathbb{R}^n$. The set-up is, otherwise, the same as in § 3. We shall denote the set of points reachable from $(x_0, 0)$ by $\mathcal{R}^T_{x_0,0}$, rather than $\mathcal{R}_{x_0,0}$, to emphasize its dependence on the parameter T.

We now address the problem of characterizing the set $\mathcal{R}^\infty_{x_0}$ comprising all points in $\mathbb{R}^n$ reachable from $(x_0, 0)$ at some nonnegative time:

$$\mathcal{R}^\infty_{x_0} = \bigcup_{T \geqq 0} \{x : (x, T) \in \mathcal{R}^T_{x_0}\}$$

(we do not use the word "reachable" in the precise sense of § 3). Note that $\mathcal{R}^\infty_{x_0}$ is a subset of $\mathbb{R}^n$, and not of $\mathbb{R}^{n+1}$

Let x_1, x_0 be points in Q. Let K be a closed cube in $\mathbb{R}^n$ such that $Q \subset$ interior $\{K\}$, and let ϕ be an element in $C^\infty(K)$. Suppose that $\phi(x_1)-\phi(x_0)>0$ and $\phi_x f(x, u) \leqq 0$ for all $(x, u) \in Q \times \Omega$. Then the kind of elementary reasoning that led to Proposition 4.1 gives us that $x_1 \notin \mathcal{R}^\infty_{x_0}$. The sufficient condition that results, under which the point x_1 does not lie in $\mathcal{R}^\infty_{x_0}$, is an autonomous version of Proposition 4.1.

It is natural to conjecture an autonomous version of our main result, Theorem 5.1, say,

"Suppose that $x_1 \notin$ closure $\{\mathcal{R}^\infty_{x_0}\}$; then there exists some $\phi \in C^\infty(K)$ such that $\phi_x f(x, u) \leqq 0$ for all $(x, u) \in Q \times \Omega$, and $\phi(x_1)-\phi(x_0)>0$."

Unfortunately the conjecture as stated is false. This is illustrated by the following, in which there does not exist even a C^1 function (let alone a C^∞ function) with the required properties.

Example 9.1. Consider

$$\dot{x}(t)=x^2(t), \qquad t \geqq 0,$$
$$Q=[-1, +1], \qquad \Omega=[-1, +1].$$

This is a control system in which the control action is trivial. The set $\mathcal{R}^\infty_0$ is the single point $\{0\}$. Yet we shall show that there is no point $x_1 \neq 0$ and no element $\phi \in C^1(\mathbb{R})$ such that $\phi(x_1)-\phi(0)>0$, and

$$\phi_x(x) \cdot x^2 \leqq 0 \tag{9.1}$$

for all $x \in [-1, +1]$. Suppose to the contrary that such a ϕ exists. Suppose further that x_1 is positive (negative x_1's are treated similarly). Then, for $\varepsilon \in (0, x_1]$,

$$\begin{aligned}\phi(x_1)-\phi(\varepsilon) &= \int_{\varepsilon}^{x_1} \phi_x(x)\,dx \\ &= \int_{\varepsilon}^{x_1} (\phi_x(x)\cdot x^2)\cdot x^{-2}\,dx.\end{aligned}$$

But the integrand is always nonpositive by inequality (9.1). It follows that $\phi(x_1)-\phi(\varepsilon) \leqq 0$. However, ϕ is continuous, whence $\phi(x_1)-\phi(x_0) \leqq 0$, a contradiction.

It is instructive to examine the difficulties which we run into when we try to adapt our methods to obtain an autonomous version of Theorem 5.1. Suppose that $x_1, x_0 \in Q$ are such that $\phi(x_1)-\phi(x_0) \leqq 0$ for all $\phi \in C^\infty(K)$ constrained by $\phi_x f(x, u) \leqq 0$ for all $(x, u) \in Q \times \Omega$. We should like to conclude (under appropriate conditions of course) that $x_1 \in \mathcal{R}_{x_0}^\infty$. Mimicking our previous arguments, we might consider under what conditions we have

(i) there exists a functional $\Lambda \in P^{\oplus}(Q \times \Omega)$ such that

$$\Lambda(\phi_x f) = \phi(x_1) - \phi(x_0) \quad \text{for all } \phi \in C^\infty(Q),$$

and

(ii) the existence of such a Λ implies $x_1 \in \mathcal{R}_{x_0}^\infty$.

The difficulties we now encounter are twofold. Firstly the map $\bar{G}: C^\infty(K) \to C(Q \times \Omega)$ defined by

$$\bar{G}(\phi) = \phi_x f$$

does not automatically satisfy the property,

$$\bar{G}(\bar{\phi}) \in \text{interior}\,\{P(Q \times \Omega)\},$$

for some $\bar{\phi} \in C^\infty(K)$, as is required to deduce (i) from standard theorems. Secondly, an autonomous version of Lemma (6.1) (this is what (ii) amounts to) is not available.

A possible approach to the problem of obtaining conditions under which an autonomous version of the theorem applies is to examine a priori assumptions under which these difficulties may be overcome.

REFERENCES

[1] S. Agmon, *Lectures on Elliptic Boundary Value Problems*, Van Nostrand Mathematical Studies, Van Nostrand, New York, 1965.

[2] H. A. Antosiewicz, *Linear control systems*, Arch. Rational Mech. Anal., 12 (1963), pp. 313–324.

[3] C. Carathéodory, *Untersuchungen über die grundlagen der Thermodynamic*, Math. Ann. (1909), pp. 355–386.

[4] R. Conti, *Contributions to linear control theory*, J. Differential Equations, 1 (1965), pp. 427–445.

[5] M. C. Delfour and S. K. Mitter, *Reachability of perturbed systems and min sup problems*, SIAM J. Control, 7 (1969), pp. 521–533.

[6] S. B. Gerschwin and D. H. Jacobson, *A controllability theory for nonlinear systems*, IEEE Trans. Automat. Control, AC-16 (1971), pp. 36–47.

[7] H. Hermes, *Controllability and the singular problem*, SIAM J. Control, 2 (1965), pp. 241–260.

[8] A. D. Ioffe, *Convex functions occurring in variational problems and the absolute minimum problems*, Math. USSR-Sb., 17 (1972), pp. 191–208.

[9] R. M. Lewis and R. B. Vinter, *New representation theorems for consistent flows*, Proc. London Math. Soc. (to appear).

[10] ———, *Relaxation of optimal control problems to equivalent convex programs*, J. Math. Anal. Applic. (to appear).

[11] E. J. McShane, *Relaxed controls and variational problems*, SIAM J. Control, 5 (1967), pp. 438–485.

[12] S. K. Mitter, *Theory of inequalities and the controllability of linear systems*, in Mathematical Theory of Control, A. V. Balakrishnan and L. Neustadt, eds., Academic Press, New York, 1967.

[13] R. T. Rockafellar, *Conjugate Duality and Optimization*, Regional Conference Series in Applied Mathematics, Society for Industrial and Applied Mathematics, Philadelphia, 1974.

[14] Y. Sunahara, T. Kabeuchi, Y. Asada, S. Aihara and K. Kishino, *On stochastic controllability for non-linear systems*, IEEE Trans. Automat. Control, AC-19 (1974), pp. 49–53.

[15] R. B. Vinter and R. M. Lewis, *The equivalence of strong and weak formulations for certain problems in optimal control*, this Journal, 16 (1978), pp. 546–570.

[16] ———, *A necessary and sufficient condition for optimality of dynamic programming type, making no a priori assumptions on the controls*, this Journal, 16 (1978), pp. 571–583.

[17] L. C. Young, *Lectures on the Calculus of Variations and Optimal Control Theory*, W. B. Saunders, Philadelphia, 1969.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 6, November 1980

THE POLYNOMIAL EQUATION $QQ_c+RP_c=\Phi$ WITH APPLICATION TO DYNAMIC FEEDBACK*

E. EMRE†

Abstract. Based on some recent results in algebraic system theory, necessary and sufficient conditions are given to achieve a given nonsingular matrix in the denominator of a matrix fraction description of a dynamic feedback system. A characterization of the required dynamic feedback laws is presented.

To achieve this, first, a complete solution is given to a problem of algebra on polynomial matrix equations.

1. Introduction. It is well known that if Q, R are $p\times p$ and $p\times m$ polynomial matrices where Q is nonsingular and if Φ is a $p\times r$ polynomial matrix, then there exist polynomial matrices Q_c, P_c such that $QQ_c+RP_c=\Phi$ iff every common left divisor of Q and P is a left divisor of Φ. However, in general, there does not exist a satisfactory characterization of the polynomial matrices Q_c and P_c to achieve this, although one can find some partial results (see Rosenbrock (1970, Chaps. 2, 5), Wolovich (1974, Chap. 7), Rosenbrock and Hayton (1977), and the references therein). One can obtain a characterization, the only complete one known, by obtaining a minimal basis for the kernel of $[Q, R]$ (see Forney (1977) and Rosenbrock and Hayton (1977)). This requires first obtaining a particular pair Q_c, P_c such that $QQ_c+RP_c=\Phi$. Then any other pair $\bar{Q}_c$, $\bar{P}_c$ can be obtained as

$$\begin{bmatrix}\bar{Q}_c\\ \bar{P}_c\end{bmatrix}=\begin{bmatrix}Q_c\\ P_c\end{bmatrix}+ML,$$

where M is a minimal basis matrix for the kernel of $[Q, R]$, and L is some polynomial matrix. Such an approach does not provide much insight into the problem because this characterization is in terms of a particular solution which is not unique, and of M whose relation to $[Q, R]$ is only partially known.

In § 2 of this paper, based on a realization given by Fuhrmann (1976), we give a system theoretic criterion for the existence, and a characterization (a parametrization) of Q_c, P_c such that $\Phi=QQ_c+RP_c$. This provides a method to construct all such Q_c, P_c, and more insight into this problem. A nice feature of this characterization is that it directly involves Q and R; these are fixed quantities, and any pair Q_c, P_c can be obtained directly, without first requiring the knowledge of a particular one. We then show that these results also lead to a characterization of solutions of the general polynomial equation $AX=B$.

A more difficult problem is to achieve $QQ_c+RP_c=\Phi$, with the constraint that Q_c be nonsingular and $P_cQ_c^{-1}$ proper. This problem arises when we consider dynamic feedback, especially in regards to stabilizing a system. For this, the reader is referred to Rosenbrock and Hayton (1977), Wolovich (1974, Chap. 7), and Emre (1978a).

In § 3, based on the ideas of § 2, we develop an existence criterion as well as a characterization (which also provides a method of construction) of all such Q_c, P_c under certain conditions. Namely, we consider the set of $p\times p$ polynomial matrices Φ, for which there exist two sets of nonnegative integers, $\{\alpha_i\}_{i=1}^{p}$ and $\{\beta_i\}_{i=1}^{p}$, and a $p\times p$

* Received by the editors April 18, 1979, and in final revised form March 3, 1980. This research was supported in part by US Army Research Grant DAAG-29-77-G-0225 and US Air Force Grant AFOSR 76-3034 Mod. B through the Center for Mathematical System Theory, University of Florida, Gainesville FL 32611.

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

nonsingular constant matrix T, such that

$$\lim_{z\to\infty}\begin{bmatrix} z^{-\alpha_1} & & 0 \\ & \ddots & \\ 0 & & z^{-\alpha_p}\end{bmatrix}\Phi\begin{bmatrix} z^{-\beta_1} & & 0 \\ & \ddots & \\ 0 & & z^{-\beta_p}\end{bmatrix}=T. \tag{1.1}$$

This last condition, which was first used in Rosenbrock and Hayton (1977) where T was assumed to be a unit matrix, is a symmetric generalization of row and column properness of polynomial matrices. It theoretically does not cause a loss of generality. This issue is discussed in detail in § 3. Furthermore, the results of the main theorems in §§ 2 and 3 are valid for the case of an arbitrary commutativity ring, provided that Q is row proper and T is invertible over the ring.

Finally, in § 4, we apply the results of § 3 to the problem of pole assignment. In particular, we give another derivation of a general theorem of Rosenbrock and Hayton (1977, Thm. 6) and discuss a conjecture given in that paper, which, in the author's belief, give more insight into these problems.

2. The equation $QQ_c+RP_c=\Phi$. In this section, we consider the following problem: Given Q, R, Φ, polynomial matrices with Q nonsingular, do there exist polynomial matrices Q_c, P_c such that $QQ_c+RP_c=\Phi$? If so, how can one obtain such possible pairs (Q_c, P_c)? For this, we first introduce some notation and preliminaries, and then present a system theoretic solution to the problem.

In the following, K is an arbitrary field. For an integer $p\geqq 1$, $K^p[z]$ denotes the set of polynomials in z with coefficients in K^p. $K^p((z^{-1}))$ denotes the formal power series of the form $\sum_{i=k}^{\infty} a_i z^{-i}$, where k is an integer and a_i is in K^p.

A formal power series $a\in K^p((z^{-1}))$ is called proper (strictly proper) if $k\geqq 0 (k>0)$. $z^{-1}K^p[[z^{-1}]]$ denotes the set of strictly proper formal K^p-power series. For a $p\times p$ nonsingular polynomial matrix Q, K_Q is defined to be the K-linear space of polynomial vectors x in $K^p[z]$ such that $Q^{-1}x$ is strictly proper.

The K-linear maps Π and Π_Q are defined as follows:

$$\text{II}: K^p((z^{-1}))\to z^{-1}K^p[[z^{-1}]];\quad x\mapsto \text{the strictly proper part of } x,$$

$$\Pi_Q: K^p[z]\to K_Q; x\mapsto Q\Pi(Q^{-1}x).$$

For a $p\times r$ polynomial matrix Φ with the ith column φ_i, we define $\Pi_Q(\Phi)$ to be the $p\times r$ matrix whose ith column is $\Pi_Q(\varphi_i)$. It is easy to see that to each $p\times r$ polynomial matrix Φ and $p\times p$ nonsingular polynomial matrix Q, there corresponds a unique $p\times r$ polynomial matrix Q_1 such that

$$\Phi=QQ_1+\Pi_Q(\Phi). \tag{2.1}$$

For a K-linear map $A: X_1\to X_2$ where X_1, X_2 are K-linear spaces, Im A denotes the image of X_1 under A as a K-linear space. For a matrix B, $\text{Sp}_K B$ denotes the K-linear subspace spanned by the columns of B. For a polynomial matrix P, $\boldsymbol{\delta}_{ci}(P)$ denotes the degree of the ith column of P.

If P is a $p\times m$ polynomial matrix with its ith column expressed as $p_i=\sum_{j=0}^{v_i} a_{ij}z^j$, where $a_{iv_i}\neq 0$, we say that P is *column proper* iff $a_{1v_1},\cdots, a_{mv_m}$ is a linearly independent set. P is said to be *row proper* iff its transpose is column proper (Wolovich (1974, Chap. 2)).

We will frequently use the well-known result that if Q_1 is a $p\times p$ column proper polynomial matrix, then

$$P_1Q_1^{-1} \text{ is proper iff } \delta_{ci}(Q_1)\geqq\delta_{ci}(P_1), \qquad i=1,\cdots,p.$$

(See Wolovich (1974, Chap. 5).)

Throughout this paper, Q, R are given $p\times p$ and $p\times m$ polynomial matrices with Q nonsingular such that

$$Z := Q^{-1}R$$

is a strictly proper rational matrix. In what follows, the following lemma and the remark play a fundamental role.

LEMMA 2.2a (Fuhrmann (1976, § 6)). *Let $\Sigma=(F,G,H)$ be defined as follows:*

$$G:K^m\to K_Q;\quad u\mapsto Ru,$$

$$F:K_Q\to K_Q;\quad x\mapsto \Pi_Q(zx),$$

$$H:K_Q\to K^p;\quad x\mapsto (Q^{-1}x)_{-1},$$

where for $a\in K^p((z^{-1}))$, $(a)_{-1}$ denotes the coefficient of z^{-1}. Then Σ, with the state space K_Q, is an observable realization of Z. Furthermore, Σ is reachable iff Q, R are relatively left prime.

Remark 2.2b. In this realization, viewed as a discrete time system, an input sequence $u=(\cdots,0,\cdots,0,u(-q),\cdots,u(-1),u(0);0,0,\cdots)$, where $q\geqq 0$ is an integer, is denoted by the polynomial vector $u(z)=u(-q)z^q+\cdots+u(0)$. Then, $\Pi_Q(u(z))$ gives the state reached by Σ at time $t=1$, with u applied to Σ which was at zero-state at $t=-q$. If $y_1z^{-1}+y_2z^{-2}+\cdots$ is the power series expansion of $Q^{-1}\Pi_Q(Ru(z))$ in z^{-1}, then y_i is the output of Σ at time $t=i$. In particular, if $u(z)$ is an input sequence as above and if $Zu(z)=P+S$ where P is the polynomial part and S is the strictly proper part (as a power series in z^{-1}), then the coefficients of P denote the outputs of Σ for $t\leqq 0$, and those of S denote the outputs of Σ for $t>0$. For details of this approach to realization theory, the reader is referred to Kalman, Falb, and Arbib (1969, Chap. 10), Fuhrmann (1976), and Emre (1980).

Let

$$n := \dim \Sigma$$

and

$$W_i := \operatorname{Im} G+\operatorname{Im} FG+\cdots+\operatorname{Im} F^iG, \qquad i=0,1,\cdots.$$

Now we have the following theorem, the main result of this section:

THEOREM 2.3. *Let Q be a $p\times p$ nonsingular polynomial matrix, and let R be a $p\times m$ polynomial matrix such that $Q^{-1}R$ is strictly proper. Let Φ be a $p\times r$ polynomial matrix. Then there exist $p\times r$ and $m\times r$ polynomial matrices Q_c and P_c such that*

$$QQc+RP_c=\Phi \tag{2.4}$$

iff

$$\mathrm{Sp}_K\,\Pi_Q(\Phi)\subset W_{n-1}. \tag{2.5}$$

Proof. Suppose that (2.4) holds. Then using the representation (2.1), we have

$$ZP_c=Q^{-1}\Pi_Q(\Phi)+Q_1-Q_c,$$

and hence

$$\Pi_Q(RP_c)=\Pi_Q(\Phi).$$

But, then, from Lemma 2.2a and Remark 2.2b it follows that the ith column of P_c is an input driving Σ from the zero-state to the state which is the ith column of $\Pi_Q(\Phi)$. Also, the ith columns of $Q^{-1}\Pi_Q(\Phi)$ and $Q_1 - Q_c$ are, respectively, the future and past-and-present outputs caused by the application of this input. Hence (2.5) must hold.

Conversely, suppose that (2.5) holds. Then to the ith column of $\Pi_Q(\Phi)$ there corresponds a polynomial vector p_{ci} which is an input driving Σ from the zero-state to that state. Define P_c to be the matrix whose ith column is p_{ci}, $i = 1, \cdots, r$. Let Q_p be the polynomial matrix whose ith column is the past-and-present outputs caused by p_{ci}. Then, by (2.2a) and (2.2b), we have

$$ZP_c = Q^{-1}\Pi_Q(\Phi) + Q_p.$$

If we define

$$Q_c := Q_1 - Q_p,$$

we obtain

$$QQ_c + RP_c = QQ_1 - QQ_p + \Pi_Q(\Phi) + QQ_p = \Phi. \qquad \square$$

COROLLARY 2.6. *For any $p \times r$ polynomial matrix Φ, there exist polynomial matrices P_c, Q_c such that* (2.4) *holds iff Q and R are relatively left prime.*

Proof. From Lemma 2.2, if Q, R are relatively left prime, then Σ is reachable. Thus, for any $p \times r$ polynomial matrix Φ, we have

$$\mathrm{Sp}_K\, \Pi_Q(\Phi) \subset W_{n-1}.$$

Hence, by Theorem 2.3, (2.4) can be achieved. Since these arguments are reversible the converse also holds. $\square$

Remark 2.7. From the proof of Theorem 2.3, we characterize pairs of polynomial matrices (P_c, Q_c) to achieve (2.4) for a given polynomial matrix Φ, and obtain a computational method to find them. We first compute a concrete realization from the given Q and R, by calculating matrix representations $\hat{F}$, $\hat{G}$, $\hat{H}$ of the F, G, H defined in Lemma 2.2, relative to some bases of K^m, K_Q and K^p. We also obtain the matrix representation of $\Pi_Q(\Phi)$, which will be a matrix over K, say $\hat{\Phi}$. Then the condition (2.5) is clearly equivalent to the condition that

$$\operatorname{rank}[\hat{G}, \hat{F}\hat{G}, \cdots, \hat{F}^{n-1}\hat{G}] = \operatorname{rank}[\hat{G}, \cdots, \hat{F}^{n-1}\hat{G}, \hat{\Phi}].$$

Then the set of all possible P_c to achieve (2.4) corresponds to the set of all possible solutions of the linear matrix equations in X,

$$\hat{\Phi} = [\hat{G}, \cdots, \hat{F}^i\hat{G}]X, \tag{2.8}$$

for all i for which the resulting equation has a solution.

Of course, it would be more convenient to solve (2.8) for each column of $\hat{\Phi}$ separately, thereby obtaining the corresponding columns of possible P_c to achieve (2.4). Thus, (2.8) provides a characterization of such P_c's. Once a P_c is obtained, we get the corresponding Q_p (past-and-present outputs caused by P_c), as the polynomial part of $Q^{-1}RP_c$. Then Q_c is simply equal to $Q_1 - Q_p$. Note here that to every P_c there corresponds a unique Q_c.

Remark 2.9. It also follows that the least possible value for $\boldsymbol{\delta}_{ci}(P_c)$ is the least integer k for which the ith column of $\Pi_Q(\Phi)$ is included in W_k. $\boldsymbol{\delta}_{ci}(Q_c)$ is determined by the ith columns of Q_p and Q_1.

Remark 2.10. The results of this section provide a characterization (parametrization) of solutions of the general polynomial equation $AX = B$, where A, B are given $p \times m$ and $p \times r$ polynomial matrices, and X is an $m \times r$ unknown polynomial matrix. This can be seen as follows. If A is not of full row rank, then there exists a $p \times p$

unimodular polynomial matrix M such that $MA = \left[\begin{smallmatrix} A_1 \\ 0 \end{smallmatrix}\right]$, where A_1 is a $\bar{p}(\leqq p) \times m$ row proper polynomial matrix. (See Wolovich (1974, Chapter 2).) Let $MB = \left[\begin{smallmatrix} B_1 \\ B_2 \end{smallmatrix}\right]$ be partitioned accordingly. Then clearly a polynomial solution exists iff $B_2 = 0$ and $A_1X = B_1$ has a solution. Let $\bar{A}_1$ be the highest row coefficient matrix of A_1. As A_1 is row proper, $\bar{A}_1$ has full row rank. Hence there exists an $m \times m$ nonsingular constant matrix T_1 such that $\bar{A}_1T_1 = [I_{\bar{p}}, 0]$. Let $A_1T_1 = [Q, R]$ be partitioned accordingly. Then Q is row proper and $\delta_{ri}(Q) > \delta_{ri}(R)$. Thus $Q^{-1}R$ is strictly proper. Letting $\bar{X} := T_1^{-1}X$, we obtain the equation

$$[Q, R]\bar{X} = B_1,$$

to which all the results of this section can be applied. Thus we can obtain a characterization of all possible X.

3. Linear dynamic feedback. In this section we apply the ideas of § 2 to the problem of modifying a system by dynamic output feedback.

Let Z be a $p \times m$ strictly proper rational transfer matrix with the dynamic interpretation

$$y = Zu.$$

We apply the dynamic feedback law

$$u = -Z_cy + v,$$

where Z_c is an $m \times p$ proper rational matrix and v is the external input. Then the closed loop transfer matrix Z_f can be written as

$$Z_f = [I_p + ZZ_c]^{-1}Z.$$

We assume that Z is given in a left matrix fraction form as

$$Z = Q^{-1}R.$$

We express Z_c in a right matrix fraction form as

$$Z_c = P_cQ_c^{-1}.$$

Then we obtain a matrix fraction description of Z_f as

$$Z_f = Q_c[QQ_c + RP_c]^{-1}R.$$

Now we consider the following questions: Suppose Φ is a given $p \times p$ nonsingular matrix. Can we find a proper rational matrix Z_c with some matrix fraction description $Z_c = P_cQ_c^{-1}$, such that $QQ_c + RP_c = \Phi$? What are necessary and sufficient conditions on Φ to achieve this? How can we characterize such a matrix Z_c when one exists? What can be said about the dynamic orders of such Z_c's?

The following theorem gives an answer to these questions for the case where (1.1) holds and a characterization of possible Q_c, P_c can be obtained from its constructive proof. In the following, whenever Q is referred to as row proper, it will be assumed that its highest degree coefficient matrix is I_p. Since this can always be achieved by a constant nonsingular output transformation, this does not cause a loss of generality.

THEOREM 3.1. *Let Φ be a given $p \times p$ polynomial matrix. Let Q be a given row proper $p \times p$ polynomial matrix with highest degree row coefficient matrix I_p and row degrees $\mu_1 \geqq \cdots \geqq \mu_p$, and let R be a given $p \times m$ polynomial matrix such that $Q^{-1}R$ is strictly proper. Let P_c be a given $m \times p$ polynomial matrix. Then there exists a column proper polynomial matrix Q_c, such that $P_cQ_c^{-1}$ is proper and*

$$QQ_c + RP_c = \Phi, \tag{3.2}$$

iff there exist integers $\gamma_1 \geqq \cdots \geqq \gamma_p \geqq 0$ *such that the following conditions hold.*

$$\text{(i)} \qquad \lim_{z\to\infty}\left\{\begin{bmatrix} z^{-\mu_1} & & 0 \\ & \cdot & \\ 0 & & z^{-\mu_p}\end{bmatrix}\Phi\begin{bmatrix} z^{-\gamma_1} & & 0 \\ & \cdot & \\ 0 & & z^{-\gamma_p}\end{bmatrix}\right\} = T$$

exists and is nonsingular, and

$$\text{(ii)} \qquad \gamma_i \geqq \delta_{ci}(P_c) \geqq r_i - 1, \qquad i = 1, \cdots, p,$$

where r_i *is the least integer* k *for which the* ith *column of* $\Pi_Q(\Phi)$, φ_i, *is in* W_{k-1}, *and* P_c *is a polynomial matrix whose* ith *column is an input which drives* Σ *from zero-state to* φ_i. *Further if* (i) *and* (ii) *holds, then* T *is the highest degree column coefficient matrix of* Q_c, *and* γ_i *is the* ith *column degree of* Q_c.

Proof. Suppose that $Z_c = P_cQ_c^{-1}$ is as in the hypothesis, and that (3.2) holds. Then by Theorem 2.3 we have $\mathrm{Sp}_K\,\Pi_Q(\Phi) \subset W_{n-1}$. Also, the ith column of P_c is an input which leads Σ from zero-state to the state φ_i. Since $r_i - 1$ is the degree of a minimum length input driving Σ to φ_i from the zero-state, we must have $\delta_{ci}(P_c) \geqq r_i - 1$. Then, since $P_cQ_c^{-1}$ is proper, we must have $\gamma_i = \delta_{ci}(Q_c) \geqq \delta_{ci}(P_c) \geqq r_i - 1$, which proves (ii). Furthermore, we have

$$\begin{aligned}
&\lim_{z\to\infty}\left\{\begin{bmatrix} z^{-\mu_1} & & 0 \\ & \cdot & \\ 0 & & z^{-\mu_p}\end{bmatrix}\Phi\begin{bmatrix} z^{-\gamma_1} & & 0 \\ & \cdot & \\ 0 & & z^{-\gamma_p}\end{bmatrix}\right\} \\
&= \lim_{z\to\infty}\left\{\begin{bmatrix} z^{-\mu_1} & & 0 \\ & \cdot & \\ 0 & & z^{-\mu_p}\end{bmatrix}(QQ_c + RP_c)\begin{bmatrix} z^{-\gamma_1} & & 0 \\ & \cdot & \\ 0 & & z^{-\gamma_p}\end{bmatrix}\right\} \\
&= \lim_{z\to\infty}\left\{\begin{bmatrix} z^{-\mu_1} & & 0 \\ & \cdot & \\ 0 & & z^{-\mu_p}\end{bmatrix}(Q)\right\}\lim_{z\to\infty}\left\{Q_c\begin{bmatrix} z^{-\gamma_1} & & 0 \\ & \cdot & \\ 0 & & z^{-\gamma_p}\end{bmatrix}\right\} \\
&\qquad + \lim_{z\to\infty}\left\{\begin{bmatrix} z^{-\mu_1} & & 0 \\ & \cdot & \\ 0 & & z^{-\mu_p}\end{bmatrix}R\right\}\lim_{z\to\infty}\left\{P_c\begin{bmatrix} z^{-\gamma_1} & & 0 \\ & \cdot & \\ 0 & & z^{-\gamma_p}\end{bmatrix}\right\} \\
&= I_p \cdot T + 0 = T,
\end{aligned}$$

which proves (i).

Conversely, suppose that (i) and (ii) hold for the given set of integers $\gamma_1 \geqq \cdots \geqq \gamma_p \geqq 0$. Now express Φ as

$$\Phi = QQ_1 + \Pi_Q(\Phi).$$

Then, by (i), we have

$$\text{(3.3)} \qquad T = \lim_{z\to\infty}\left\{\begin{bmatrix} z^{-\mu_1} & & 0 \\ & \cdot & \\ 0 & & z^{--\mu_p}\end{bmatrix}QQ_1\begin{bmatrix} z^{-\gamma_1} & & 0 \\ & \cdot & \\ 0 & & z^{-\gamma_p}\end{bmatrix}\right\} + 0.$$

Suppose that

$$\lim_{z\to\infty}\left\{Q_1\begin{bmatrix} z^{-\gamma_1} & & 0 \\ & \cdot & \\ 0 & & z^{-\gamma_p}\end{bmatrix}\right\}$$

does not exist. Then

$$Q_1\begin{bmatrix} z^{-\gamma_1} & & 0 \\ & \ddots & \\ 0 & & z^{-\gamma_p}\end{bmatrix}$$

is not a proper rational matrix. Since

$$\begin{bmatrix} z^{-\mu_1} & & 0 \\ & \ddots & \\ 0 & & z^{-\mu_p}\end{bmatrix} Q = I_p + \cdots,$$

neither is the product of the two rational matrices, which contradicts (3.3). Hence

$$\lim_{z\to\infty}\left\{Q_1\begin{bmatrix} z^{-\gamma_1} & & 0 \\ & \ddots & \\ 0 & & z^{-\gamma_p}\end{bmatrix}\right\}$$

is a matrix with finite entries. Thus, the limit of the product is equal to the product of the limits, i.e.,

$$T = I_p \cdot \lim_{z\to\infty}\left\{Q_1\begin{bmatrix} z^{-\gamma_1} & & 0 \\ & \ddots & \\ 0 & & z^{-\gamma_p}\end{bmatrix}\right\},$$

which proves that Q_1 is a column proper matrix with the highest degree column coefficient matrix T, and the column degrees $\gamma_1, \cdots, \gamma_p$.

Now choose P_c such that the ith column of P_c drives Σ from the zero-state to φ_i and such that $\gamma_i \geqq \delta_{ci}(P_c) \geqq r_i - 1$. Let Q_p be the polynomial matrix whose columns represent the past-and-present outputs due to P_c. Define $Q_c := Q_1 - Q_p$. Then $\Phi = QQ_c + RP_c$. As Z is strictly proper, $\delta_{ci}(Q_p) < \delta_{ci}(P_c)$, $\gamma_i = \delta_{ci}(Q_1) \geqq \gamma_{ci}(P_c) > \delta_{ci}(Q_p)$, which implies that Q_c is column proper with the highest degree column coefficient matrix T and $\delta_{ci}(Q_c) = \gamma_i$. As $\gamma_i \geqq r_i - 1$, $Z_c := P_cQ_c^{-1}$ is proper. This completes the proof. □

Remark 3.4. The second part of the proof of Theorem 3.1 provides a characterization and a method of construction of the possible pairs (Q_c, P_c), to achieve (3.2) with a proper Z_c, along the lines given in Remark 2.7.

Remark 3.5. Equation (3.3) proves that for a given Φ satisfying conditions (3.1i–ii), the upper bound of the orders of the compensators to achieve (3.2) is $\sum_{i=1}^{r} \gamma_i$. If the resulting Q_c, P_c obtained by the method suggested in Remark 3.5 are not relatively right prime, one can achieve the same Z_f with a compensator of lower order. Therefore, a nontrivial problem of interest is to obtain Q_c, P_c, satisfying (3.2), with nonunimodular common right divisors, thereby obtaining simpler compensators. This indicates the necessity of deeper research on the method of Remark 3.5.

Remark 3.6. Another problem of interest is to find conditions for relative right primeness of Q_c and $QQ_c + RP_c$, and relative left primeness of R and $QQ_c + RP_c$, for a better understanding of the resulting system Z_f.

Remark 3.7. In Theorem 3.1 we have assumed the condition (3.1i) on Φ. Theoretically, this is not a constraint on Φ because whenever $\Phi = QQ_c + RP_c$ for some $Z_c := P_cQ_c^{-1}$ which is proper, there exist unimodular polynomial matrices M_1, M_2 such that $M_1\Phi M_2$ satisfies (3.1i) for some integers $\gamma_1 \geqq \cdots \geqq \gamma_p$. However, there is not a known general procedure to determine (the existence of) such M_1, M_2 and $\{\gamma_i\}_{i=1}^{p}$ for a general $p \times p$ polynomial matrix Φ.

First of all, for a given $p \times p$ nonsingular polynomial matrix Φ, the possible sets of integers $\{\alpha_i\}_{i=1}^{p}$ and $\{\beta_i\}_{i=1}^{p}$, $\alpha_1 \geqq \cdots \geqq \alpha \geqq 0$ and $\beta_1 \geqq \cdots \geqq \beta_p \geqq 0$, are not in general unique such that the condition

$$\lim_{z\to\infty}\left\{\begin{bmatrix} z^{-\alpha_1} & & 0 \\ & \cdot & \\ 0 & & z^{-\alpha_p}\end{bmatrix} M_1\Phi M_2 \begin{bmatrix} z^{-\beta_1} & & 0 \\ & \cdot & \\ 0 & & z^{-\beta_p}\end{bmatrix}\right\} = T, \tag{3.8}$$

where T is a nonsingular constant matrix, holds for some unimodular polynomial matrices M_1 and M_2, if any exist. This fact is clear by considering the case where Φ is a scalar polynomial.

However, condition (3.1i) dictates that for (3.2) to be achieved, it is necessary to have $\alpha_i = \mu_i$, $i = 1, \cdots, p$.

The question which arises now is the following. If we constrain α_i to be equal to μ_i for $1 \leqq i \leqq p$, is the set of integers $\beta_1 \geqq \cdots \beta_p \geqq 0$ unique for which there exist unimodular polynomial matrices M_1, M_2 such that (3.8) holds?

The answer to this question is not known at the present. It can be approached by considering those nonsingular $p \times p$ Φ's which can be written in the form $\Phi = \Phi_1 T_1 T_2 \Phi_2$, where Φ_1 and Φ_2 are $p \times \gamma$ and $\gamma \times p (\gamma \geqq p)$ polynomial matrices of rank p. Then M_1 (or M_2) can be chosen such that $M_1\Phi_1(\Phi_2 M_2)$ is row (column) proper. T_1, T_2 are constant nonsingular matrices such that (3.9) holds for some nonsingular constant matrix T. In this case, (3.1i) requires that the row degrees of $M_1\Phi_1$ be $\{\mu_i\}_{i=1}^{p}$. In general, this approach requires considering such factorizations of a polynomial matrix Φ. For partial results the reader is referred to Emre (1978b), Fuhrmann (1977) and the references given there.

Another possible approach is to choose Φ_1 and Φ_2 in a suitable way a priori. The simplest case, of course, is to choose $\gamma = p$. In this case, we will have $Z_f = Q_c\Phi_2^{-1}T_2^{-1}T_1^{-1}\Phi_1^{-1}R$. For a treatment of the pole assignment problem for this case, the reader is referred to Emre (1978a).

Remark 3.9. In Remark 3.7 we have considered the general case of the existence of $M_1, M_2, \{\alpha_i\}_{i=1}^{p}, \{\beta_i\}_{i=1}^{p}$. If we consider the problem of pole assignment only, there is, as we will see in the next section, a technique given in Rosenbrock and Hayton (1977, Lemma 1), which makes the application of Theorem 3.1 to derive previous techniques of pole assignment possible.

Remark 3.10. Clearly, the results of Theorem 2.1 and 3.1 are also valid in the case where K is an arbitrary commutativity ring, provided that Q is row proper (i.e., the highest coefficient row matrix of Q is invertible over K), and T is invertible over K.

4. Pole assignment. In this section, we will apply the results of § 3 to the problem of pole assignment. In particular we derive a theorem of Rosenbrock and Hayton (1977, Thm. 6) and discuss a conjecture given there. Our derivation provides more insight into this theorem, and also a characterization of the dynamic feedback systems to achieve the desired pole assignment.[1]

In what follows, for $\psi \in K[z]$, $\delta(\psi)$ denotes the degree of ψ.

THEOREM 4.1. *Let $\{\psi_i\}_{i=1}^{p}$ be a sequence of monic polynomials with ψ_i dividing ψ_{i-1}, $i = 2, \cdots, p$. Assume that Q and R are left prime. Let v_1 be the largest reachability index of Z. Then a sufficient condition for the existence of a proper $m \times p$ rational matrix Z_c with a matrix fraction description $Z_c = P_cQ_c^{-1}$ such that the invariant factors of $QQ_c + RP_c$*

[1] As it is shown in that paper, previous pole assignment results, such as those of Brasch and Pearson (1970) and Rosenbrock (1970, pp. 190–193), can be obtained from this theorem.

are $\psi_1, \cdots, \psi_p$, is

$$\sum_{i=1}^{k} \delta(\psi_i) \geqq \sum_{i-1}^{k} (\mu_i+v_1-1), \qquad k=1,\cdots,p, \tag{4.2}$$

with equality holding when $k=p$.

Proof. First we need the following lemma.

LEMMA 4.3 (Rosenbrock and Hayton (1977, Lemma 1)). *Let $\{\alpha_i\}_{i=1}^{p}$, $\{\beta_i\}_{i=1}^{p}$ be two sequences of integers such that $\alpha_1 \geqq \cdots \geqq \alpha_p \geqq 0$ and $\beta_1 \geqq \cdots \geqq \beta_p \geqq 0$. Let $\{\psi_i\}_{i=1}^{p}$ be a sequence of given monic polynomials such that ψ_i divides ψ_{i-1}, $i=2,\cdots p$. Then there exists a $p \times p$ polynomial matrix Φ whose invariant factors are $\{\psi_i\}_{i=1}^{p}$ and which satisfies*

$$\lim_{z\to\infty} [\operatorname{diag}(z^{-\alpha_i})\Phi \operatorname{diag}(z^{-\beta_i})] = I_p$$

iff

$$\sum_{i=1}^{k} \delta(\psi_i) \geqq \sum_{i=1} (\alpha_i+\beta_i), \qquad k=1,\cdots,p,$$

with equality holding when $k=p$.

Now suppose we are given a set of polynomials $\{\psi_i\}_{i=1}^{p}$ as in the hypothesis. Then from Lemma 4.3 with $\alpha_i=\mu_i$ and $\beta_i=v_1-1$, $i=1,\cdots,p$, there exists a $p\times p$ polynomial matrix Φ which satisfies (3.1i) with $\gamma_i=v_1-1$, $i=1,\cdots,p$. As Q, R are relatively left prime, Σ is reachable. Since every state can be reached in v_1 steps, r_i in (3.1ii) is always less than or equal to v_1. Hence, by Theorem 3.1, the proof follows. □

Remark 4.4. A method to obtain a polynomial matrix Φ as in Lemma 4.3 is given in the proof in Rosenbrock and Hayton (1977). Results of § 3 not only yield insight into why Theorem 4.1 is true, but they also give a characterization (a parametrization) of the possible compensators $Z_c=P_cQ_c^{-1}$ to achieve $\Phi=QQ_c+RP_c$.

Remark 4.5. There is a conjecture in Rosenbrock and Hayton (1977, § 6) which considers whether, assuming that $p \leqq m$, (4.2) can be replaced by the more symmetric and sharper condition

$$\sum_{i=1}^{k} \delta(\psi_i) \leqq \sum_{i=1}^{k} (\mu_i+v_i-1), \qquad k=1,\cdots,p, \tag{4.6}$$

with equality holding when $k=p$.

The validity of this conjecture would imply that for each such set of $\{\psi_i\}_{i=1}^{p}$ satisfying (4.6), there exists at least one set of integers $\{t_i\}_{i=1}^{p}$, $t_1 \geqq \cdots \geqq t_p \geqq 0$, not necessarily $t_i=v_i$, such that

(i)
$$\sum_{i=1}^{k} \delta(\psi_i) \leqq \sum_{i=1}^{k} (\mu_i+t_i-1), \qquad k=1,\cdots,p,$$

with equality holding when $k=p$,

(ii) there exists a polynomial matrix Φ whose invariant factors are $\psi_1,\cdots,\psi_p$, as in Lemma 4.3, with $\alpha_i=\mu_i$ and $\beta_i=t_i-1$, $i=1,\cdots,p$, such that the ith column of $\Pi_Q(\Phi)$ is reachable in t_i steps.

However, the problem of checking whether this implication is true does not seem to be tractable at this point.

Remark 4.7. In § 3 and in this section we have assumed factorizations of the form $Z=Q^{-1}R$ and $Z_c=P_cQ_c^{-1}$. Dually, one can use factorizations of the form $Z=PQ^{-1}$ and $Z_c=Q_c^{-1}R_c$, use the expression $Z_f=Z(I_m+Z_cZ)^{-1}$, and apply the same results after transpositions. In this case, the roles of the reachability and observability indices and p and m change.

REFERENCES

F. M. BRASCH AND J. B. PEARSON (1970), *Pole placement using dynamic compensators*, IEEE Transactions on Automatic Control, AC-15, pp. 34–43.

E. EMRE (1978a), *Pole assignment by dynamic feedback*, Control Systems Center Report No. 413, University of Manchester, England; Int. J. Control, to appear.

——— (1978b), *Nonsingular factors of polynomial matrices and* (A, B)*-invariant subspaces*, Memorandum COSOR 78-12, Dept. of Mathematics, Eindhoven University of Technology, The Netherlands; this Journal, 18, pp. 288–296.

—— (1980), *On a natural realization of matrix fraction descriptions*, IEEE Trans. Autom. Contr., Correspondence, AC-25, pp. 288–289.

G. D. FORNEY (1975), *Minimal bases of rational vector spaces, with applications to multivariable linear systems*, SIAM J. Control, 13, pp. 493–520.

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

——— (1977), *Simulation of linear systems and factorizations of matrix polynomials*, Int. J. Control, to appear.

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

H. H. ROSENBROCK (1970), *State Space and Multivariable Theory*, John Wiley, New York.

H. H. ROSENBROCK AND G. E. HAYTON (1977), *The general problem of pole assignment*, Control Systems Center Report no. 288, University of Manchester, England.

W. A. WOLOVICH (1974), *Linear Multivariable Systems*, Springer, New York.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 6, November 1980

0363-0129/80/1806-0003 $01.00/0

NONLINEAR COMPLEMENTARITY PROBLEMS IN A FUNCTION SPACE*

TAKAO FUJIMOTO†

Abstract. A class of nonlinear complementarity problems defined by compact operators is considered in the space of continuous functions on a compact Hausdorff space, and an existence theorem established. The method of proof is based on Schauder's fixed-point theorem. Related topics such as a least element and indifferent optimization problems are also discussed.

1. Introduction. Many contributions have been made concerning the existence and the uniqueness of solutions for complementarity problems. In the case of nonlinear complementarity problems (NLCP), the reader is referred to Karamardian [6], [7], Eaves [3], Moré [11], [12], Kojima [8] etc. (For the linear case, see, e.g., Kaneko [5]). In recent issues of this journal, Fisher and Tolle [4] and Watson [16] have provided constructive algorithms for finding solutions to NLCP problems together with general existence theorems (see also the literature in [4] and [16]). These algorithms are closely related to those for finding fixed-points of nonlinear mappings or solving nonlinear equation systems.

On the other hand, in 1972, Cottle and Veinott, Jr. [2] characterized the convex polyhedral sets having a least element. Following this work, Tamir [15] showed complementarity properties associated with z-functions and m-functions. Simple proofs of these properties are given in Bod [1]. The purpose of the present paper is to extend some of the results by Tamir and Bod to a function space. Our proof is similar to that in the finite dimensional case and is based on a fixed-point theorem by Schauder.

2. Notation and complementarity problems. Let S be a compact Hausdorff space and $C(S)$ be the set of continuous functions on S. The symbol 0 stands for the null element of the Banach space $C(S)$. K denotes the cone of nonnegative continuous functions on S. A partial order is introduced by the cone K; i.e., $x \geqq y$ $(x, y \in C(S))$ if $x - y \in K$. A function $x \equiv \min(x^1, \cdots, x^n)$ for $x^1, \cdots, x^n \in C(S)$ means that defined by $x(s) \equiv \min(x^1(s), \cdots, x^n(s))$ at each point s in S; $\max(x^1, \cdots, x^n)$ is similarly defined. Let T be an operator from K into $C(S)$.

In this paper we consider the following class of nonlinear complementarity problems.

(NLCP): Find $x \in K$ such that $x - Tx - b \geqq 0$ and $\min(x, x - Tx - b) = 0$, where b is a given element in $C(S)$. We make the following assumptions.

A.1. T is compact (or completely continuous).

A.2. T is monotone with respect to K; i.e., $Tx \geqq Ty$ if $x \geqq y$ for $x, y \in K$.

A.3. There exists an $x^0 \in K$ such that $x^0 - Tx^0 - b \geqq 0$.

3. Existence. First we prove an existence theorem.

THEOREM 3.1. *Given* A.1–A.3, *there exists a solution to the* NLCP.

Proof. Let us consider the mapping F defined as

$$F: x \in K \to Fx \equiv \max(Tx + b, 0).$$

F is also compact (Yosida [17, p. 278]). Next define the set $D \equiv \{x | x \in K, x \leqq x^0\}$, where x^0 is the element given in the assumption A.3. D is bounded, closed and convex. By A.2

* Received by the editors September 10, 1979, and in final form March 10, 1980.

† Department of Economics, University of Kagawa, Takamatsu, Kagawa 760, Japan.

it is clear that F maps D into itself, i.e.,

$$FD \equiv \{Fx | x \in D\} \subset D.$$

By A.1, the set FD is compact. Thus Schauder's theorem [14, p. 175] insures the existence of a fixed-point x^*:

$$x^* = Fx^* = \max (Tx^* + b, 0).$$

It is not difficult to see that x^* is a solution to the NLCP. □

Remark 3.1. x^* can be approximated by an iterative method. Define $x^1 \equiv \max (Tx^0 + b, 0)$ and successively $x^{n+1} \equiv \max (Tx^n + b, 0)$. The sequence $\{x^n\}$ is monotone and convergent by A.1.

Remark 3.2. Note that a fixed-point x^* found in Theorem 3.1 satisfies the inequality $x^* \leqq x^0$.

4. Least element. Now define the set $E \equiv \{x | x \in K, x - Tx - b \geqq 0\}$. An element x^{**} of the set E is called a *least element* if $x^{**} \leqq x$ for all $x \in E$. We have

THEOREM 4.1. *Given* A.1–A.3, *there exists a least element in E.*

Proof. Denote by E^* the set of fixed-points of the mapping F from D into D. Clearly $E^* \subset E$. E^* is closed and contained in the compact set FD. So, E^* is also compact. Now define the set E_s for each $s \in S$:

$$E_s \equiv \{x^* | x^* \in E^*, x^*(s) \leqq y^*(s) \text{ for all } y^* \in E^*\}.$$

That is, E_s is the set of functions in E^* which attain the minimum value at s among those functions in E^*. E_s is not empty, since E^* is compact, and moreover E_s is closed. For any finite set $I \equiv \{s^1, \cdots, s^n\} \subset S$, $\bigcap_{s \in I} E_s \neq \emptyset$. To prove this, take $x_s^* \in E_s$ for each $s \in I$. Then, the function $x^{00} \equiv \min_{s \in I} \{x_s^*\}$ is in E by A.2. Noting Remark 3.2, we know the existence of an element $z^* \in E^*$ such that $z^* \leqq x^{00}$. This implies that $z^* \leqq x_s^*$ for all $s \in I$, and thus $z^* \in \bigcap_{s \in I} E_s$. This finite intersection property means $\bigcap_{s \in S} E_s \neq \emptyset$ by the compactness of E^*. Denote by x^{**} an element in $\bigcap_{s \in S} E_s$. Evidently $x^{**} \in E^*$ and $x^{**} \leqq x^*$ for all $x^* \in E^*$.

Now take any x in E. Define $x^{00} \equiv \min (x, x^0) \in E \cap D$. Theorem 3.1 and Remark 3.2 ensure the existence of an element $x^* \in E^*$ such that $x^* \leqq x^{00} \leqq x$. Therefore, $x^{**} \leqq x$ for all $x \in E$. □

Remark 4.1. A least element of E is clearly unique.

Remark 4.2. A least element of E is a solution to the NLCP because x^{**} is in E^*.

5. Indifferent optimization problems. A scalar function $f(x)$ defined on K is said to be *isotonic* if $x \geqq y$ for $x, y \in K$ implies $f(x) \geqq f(y)$. Now consider the following programming problem:

(P): minimize $f(x)$ subject to $x \in K$ and $x - Tx - b \geqq 0$.

THEOREM 5.1. *Given* A.1–A.3, *a least element* x^{**} *of E is a solution to the problem* (P) *whatever isotonic function is specified as an objective function.*

This theorem is easily proven by the definitions of a least element and the set E.

6. Uniqueness of a solution to the NLCP. We make one more assumption (see Yun [18]):

A.4. $x \leq y$ for $x, y \in K$ implies $x - Tx \not\geqq y - Ty$.

Here $x \leq y$ means $y - x \in K - \{0\}$. Almost needless to say, if T is contractive, i.e., $\|Tx - Ty\| < \|x - y\|$ for $x, y \in K$, then A.4 is satisfied ($\|\cdot\|$ denotes the maximum-norm).

THEOREM 6.1. *Given* A.1–A.4, *the* NLCP *has a unique solution.*

Proof. Let x^* be a solution and suppose z^* is another solution different from x^*.

Define a set $U \equiv \{s \mid s \in S, x^*(s) < z^*(s)\}$. Without loss of generality, we assume U is not empty. First note that $z^*(s) > 0$ for $s \in U$, and so

$$(z^* - Tz^* - b)(s) = 0 \quad \text{for } s \in U, \tag{6.1}$$

because z^* is a solution to the NLCP. Next define a function $y \equiv \max(x^*, z^*)$. From A.2 and A.4, we have

$$(y - Ty)(s) > (x^* - Tx^*)(s) \geqq b(s) \quad \text{for some } s \in U. \tag{6.2}$$

By A.2, on the other hand, we have

$$(z^* - Tz^*)(s) \geqq (y - Ty)(s) \quad \text{for all } s \in U. \tag{6.3}$$

From (6.2) and (6.3), it follows $(z^* - Tz^*)(s) > b(s)$ for some $s \in U$, contradicting (6.1). Thus, x^* must be a unique solution. □

7. Additional remarks. When the operator T is a linear integral operator on $C([\alpha, \beta])$, a necessary and sufficient condition for T to be compact is known (see Radon [13]). Closely related to our analysis is the memoir by Krein and Rutman [10] in which nonlinear compact operators were dealt with in relation to the eigenvalue problem of positive operators (see also Krasnoselskii [9]).

REFERENCES

[1] P. Bod, *On closed sets having a least element*, in Optimization and Operations Research, W. Oettli and K. Ritter, eds., Springer, Berlin, 1976, pp. 23–34.

[2] R. W. Cottle and A. F. Veinott, Jr., *Polyhedral sets having a least element*, Math. Programming, 3 (1972), pp. 238–249.

[3] B. C. Eaves, *On the basic theorem of complementarity*, Math. Programming, 1 (1971), pp. 68–75.

[4] M. L. Fisher and J. W. Tolle, *The nonlinear complementarity problem: existence and determination of solutions*, this Journal, 15 (1977), pp. 612–624.

[5] I. Kaneko, *Linear complementarity problems and characterizations of Minkowski matrices*, Linear Algebra and Appl., 5 (1978), pp. 111–129.

[6] S. Karamardian, *The nonlinear complementarity problem with applications, Part I and Part II*, J. Optim. Theory Appl., 4 (1969), pp. 87–88 and pp. 167–181.

[7] ———, *The complementarity problem*, Math. Programming, 2 (1972), pp. 107–129.

[8] M. Kojima, *A unification of the existence theorems of the nonlinear complementarity problem*, Math. Programming, 9 (1975), pp. 257–277.

[9] M. A. Krasnoselskii, *Positive Solutions of Operator Equations*, P. Noordhoff, Gronigen, The Netherlands, 1964.

[10] M. G. Krein and M. A. Rutman, *Linear operators leaving invariant a cone in a Banach space*, Uspehi Mat. Nauk, 3 (1948), pp. 3–95. American Mathematical Society Translation no. 26 (1950).

[11] J. J. Moré, *Classes of functions and feasibility conditions in nonlinear complementarity problems*, Math. Programming, 6 (1974), pp. 327–338.

[12] ———, *Coercivity conditions in nonlinear complementarity problems*, SIAM Rev., 16 (1974), pp. 1–16.

[13] J. Radon, *Über lineare Funktionaltransformationen und Funktionalgleichungen*, S.-B. Öster, Akad. Wiss., 128 (1919), pp. 1083–1121.

[14] J. Schauder, *Der Fispunktsatz in Funktionalräumen*, Studia Math., 2 (1930), pp. 171–180.

[15] A. Tamir, *Minimality and complementarity properties associated with z-functions and m-functions*, Math. Programming, 7 (1974), pp. 17–31.

[16] L. T. Watson, *Solving the nonlinear complementarity problem by a homotopy method*, this Journal, 17 (1979), pp. 36–46.

[17] K. Yosida, *Functional Analysis*, Springer, Berlin, 1965.

[18] K. K. Yun, *On the existence of a unique and stable market equilibrium*, J. Econom. Theory, 20 (1979), pp. 118–123.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 6, November 1980

0363-0129/80/1806-0004 $01.00/0

ON KALMAN'S PROCEDURE FOR THE COMPUTATION OF THE CONTROLLABLE/OBSERVABLE CANONICAL FORM*

DANIEL L. BOLEY†

Abstract. An example is given in which the algorithm given in [Kalman, SIAM J. Control Ser. A, 1 (1963), pp. 152–192] to compute the joint controllability-observability decomposition fails.

In § 7 of Kalman [1], there is described a procedure to compute the *joint* controllability-observability decomposition of a linear time-invariant dynamic system,

$$\mathbf{x}=F\mathbf{x}+G\mathbf{u},\qquad \mathbf{y}=H\mathbf{x}, \tag{1}$$

where I follow the notation in Kalman [1]. In its basic outline, the decomposition proceeds in four steps which are sketched as follows:

(2.a) *Controllability decomposition.*

Compute a transformation to transform the system (1) to the form:

$$\dot{\bar{\mathbf{x}}}=\begin{bmatrix}F_{11} & F_{12}\\ 0 & F_{22}\end{bmatrix}\bar{\mathbf{x}}+\begin{bmatrix}G_1\\ 0\end{bmatrix}\mathbf{u},$$

$$\mathbf{y}=H\bar{\mathbf{x}}.$$

(2.b) *Provisional observability decomposition.*

Treating each subsystem

$$\dot{\bar{\mathbf{x}}}_1=F_{11}\bar{\mathbf{x}}_1+G_1\mathbf{u},$$

$$\dot{\bar{\mathbf{x}}}_2=F_{22}\bar{\mathbf{x}}_2,$$

in isolation, compute the observability decomposition of each of those two subsystems. The result has the form

$$\dot{\bar{\bar{\mathbf{x}}}}=\begin{bmatrix}F'_{11} & F'_{12} & F'_{13} & F'_{14}\\ 0 & F'_{22} & F'_{23} & F'_{24}\\ 0 & 0 & F'_{33} & F'_{34}\\ 0 & 0 & 0 & F'_{44}\end{bmatrix}\bar{\bar{\mathbf{x}}}+\begin{bmatrix}G'_1\\ G'_2\\ 0\\ 0\end{bmatrix}\mathbf{u},$$

$$\mathbf{y}=[0\quad H'_2\quad 0\quad H'_4]\bar{\bar{\mathbf{x}}}.$$

At this point,

F'_{11} represents the controllable unobservable part $C\bar{O}$,
F'_{22} represents the controllable observable part $C\,O$,
F'_{33} represents the provisional uncontrollable unobservable part $\bar{C}\bar{O}$,
F'_{44} represents the provisional uncontrollable observable part $\bar{C}O$.

(2.c) *Decoupling.*

Compute a transformation to zero out F'_{23}, the coupling between the controllable observable ($C\,O$) part and the uncontrollable unobservable ($\bar{C}\bar{O}$) part, without filling in the zeros already computed.

* Received by the editors October 16, 1979, and in final form February 29, 1980. This research was supported in part by the U.S. Army Research Office under grant no. DAAG29-78-G0179.

† Department of Computer Science, Stanford University, Stanford, California 94305.

(2.d) *Final observability decomposition.*

Compute a transformation that will adjust the observability split in the uncontrollable half $(\bar{C}\bar{O}+\bar{C}O)$ to give a final result that has the form

$$\dot{\mathbf{z}}=\begin{bmatrix} F''_{11} & F''_{12} & F''_{13} & F''_{14} \\ 0 & F''_{22} & 0 & F''_{24} \\ 0 & 0 & F''_{33} & F''_{34} \\ 0 & 0 & 0 & F''_{44} \end{bmatrix}\mathbf{z}+\begin{bmatrix} G''_1 \\ G''_2 \\ 0 \\ 0 \end{bmatrix}\mathbf{u},$$

$$\mathbf{y}=[0 \quad H''_2 \quad 0 \quad H''_4]\mathbf{z}.$$

In Kalman [1], it is stated that the dimension of the $\bar{C}\bar{O}$ part, F'_{33}, should be at least as big as the same part, F''_{33}, in the complete decomposition after step (d); and conversely, the $\bar{C}O$ part, F'_{44}, should be no bigger than the same part, F''_{44}, in the complete decomposition, since the two parts together will have the same dimension. However, an example is presented here which does not have this property.

In this example, part $\bar{C}\bar{O}$, which should be of dimension 1 in the complete decomposition, comes out empty. We start with the final decomposition, and transform it into a form in which part $\bar{C}\bar{O}$ will be empty by applying only similarity transformations. We start with

$$(3) \qquad F=\begin{bmatrix} -2 & -1 & 1 & 1 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & 2 & 4 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad G=\begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad H=[0 \quad 1 \quad 0 \quad 1].$$

Note that this is in a fully decomposed state, in the same order as is described in step (d). Note also that this system has distinct eigenvalues, and thus is diagonalizable. We apply the transformation:

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

by computing $F \leftarrow TFT^{-1}$, $G \leftarrow TG$, $H \leftarrow HT^{-1}$, to get the equivalent system represented by:

$$(4) \qquad F=\begin{bmatrix} -2 & -1 & 2 & 1 \\ 0 & -1 & 3 & 5 \\ 0 & 0 & 2 & 4 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad G=\begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad H=[0 \quad 1 \quad -1 \quad 1].$$

In the same way, we then apply the orthogonal transformation

$$T=\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1/\sqrt{2} & 1/\sqrt{2} \\ 0 & 0 & 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix},$$

to return H to its original zero structure $[0 \quad X \quad 0 \quad X]$. We obtain:

$$(5) \qquad F=\begin{bmatrix} -2 & -1 & 3/\sqrt{2} & 1/\sqrt{2} \\ 0 & -1 & 8/\sqrt{2} & -2/\sqrt{2} \\ 0 & 0 & 7/2 & -3/2 \\ 0 & 0 & 5/2 & -1/2 \end{bmatrix}, \quad G=\begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad H=[0 \quad 1 \quad 0 \quad -\sqrt{2}].$$

Note that the controllable part (upper left 2×2 block) has been left unchanged.

If we treat each half, controllable and uncontrollable, in isolation, and compute the observability decomposition for each half, we get a (correct) $C\bar{O}/CO$ split into two parts of dimension 1 and 1 respectively for the controllable half, but we get a (misleading) $\bar{C}\bar{O}/\bar{C}O$ split into two parts of dimension 0 and 2 respectively for the uncontrollable half. In short, it looks as if the entire uncontrollable part is observable. Since part $\bar{C}\bar{O}$ (block F'_{33} in (2.b)) is empty, steps (c) and (d) are empty steps—there is no coupling to adjust for and we have what the procedure would give as the final decomposition. When part CO of dimension 1 is added in, it would appear that the dimension of the entire observable part is 3, whereas it should be 2. This should be clear both from system (3) and from the fact that rank $[H, HF, HF^2, \cdots]=2$, for one can easily compute that $HF^2=H$.

Conclusion. An example has been exhibited which fails to have a property essential to the commonly used procedure for computing the joint controllability-observability canonical form of a linear dynamic system.

REFERENCE

[1] R. E. KALMAN, *Mathematical description of linear dynamic systems*, SIAM J. Control Ser. A, 1 (1963), pp. 152–192.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 6, November 1980

0363-0129/80/1806-0005 $01.00/0

NONUNIQUENESS OF SOLUTIONS IN THE CALCULUS OF VARIATIONS: A GEOMETRIC APPROACH*

DOMINIQUE HENRI†

Abstract. Differential topology is the latest of the many mathematical tools which have been used in the study of optimization problems. In this paper we apply it to the fundamental problem of the calculus of variations in $\mathbb{R}^n$:

$$\mathcal{P}_{\xi,T}\begin{cases}\operatorname{Inf}\displaystyle\int_0^T f(x(t),\dot{x}(t))\,dt,\\ x(0)=\xi_0, \qquad x(T)=\xi.\end{cases}$$

We show that if f is smooth, coercive (i.e., grows quickly at infinity), and convex with respect to $\dot{x}$, this problem has exactly one solution for almost every end condition (ξ, T). Then, using Thom's transversality theorems, we classify those points (ξ, T) in $\mathbb{R}^n \times \mathbb{R}$ where there is more than one solution (the singularity set). If the dimension is low ($n \leqq 4$), there are but a finite number of singularity types which will fit almost all functions f.

Preliminaries.

0.1. Value function. Let $f:\mathbb{R}^n\times\mathbb{R}^n\to\mathbb{R}$ be a C^∞ mapping satisfying the three following assumptions.

(i) For every $\zeta\in\mathbb{R}^n$, the mapping $\eta\to f(\zeta,\eta)$ is convex.

(ii) There exists a convex, monotone function Φ, bounded from below, from $\mathbb{R}^+$ to $\mathbb{R}$, such that

$$\lim_{t\to+\infty}\frac{\Phi(t)}{t}=+\infty$$

and

$$f(\zeta,\eta)\geqq\Phi(\|\eta\|)\quad\forall(\zeta,\eta)\in\mathbb{R}^n\times\mathbb{R}^n.$$

(iii) For every $(\zeta,\eta)\in\mathbb{R}^n\times\mathbb{R}^n$, the matrix of the second partial derivatives of f with respect to η,

$$f''_{\eta^2}=\left(\frac{\partial^2 f}{\partial\eta_i\,\partial\eta_j}\right),$$

is positive definite.

We associate with f the problem

$$\inf\int_0^T f(x(t),\dot{x}(t))\,dt,$$

where the infimum is taken over all absolutely continuous functions x, from $[0, T]$ to $\mathbb{R}^n$ (with derivative $\dot{x}$ almost everywhere) satisfying

$$x(0)=\xi_0,\qquad x(T)=\xi,$$

ξ_0 being a fixed point in $\mathbb{R}^n$ and (ξ, T) lying in $\mathbb{R}^n\times]0,+\infty[$.

Let $\mathcal{P}_{\xi,T}$ denote this problem and $V(\xi, T)$ the value of this infimum.

DEFINITION 0.1. *The function $(\xi, T)\to V(\xi, T)$ defined on $\mathbb{R}^n\times]0,+\infty[$ is called the value function (or the Hamilton-Jacobi-Bellman function) of the problem.*

* Received by the editors December 12, 1978, and in revised form November 23, 1979.

† Centre de Recherche de Mathématiques de la Décision, Université Paris-Dauphine, 75775-Paris Cedex 16.

It is a well-known fact (see [8]) that, under the hypotheses (i) and (ii), the problem $\mathcal{P}_{\xi,T}$ admits, for every (ξ, T), at least one optimal solution satisfying the Euler Lagrange (E. L.) equation

$$f'_\zeta(x(t), \dot{x}(t)) = \frac{d}{dt} f'_\eta(x(t), \dot{x}(t)).$$

It follows from condition (iii) that this last equation can be solved with respect to $\ddot{x}(t)$, yielding

$$\text{(E. L.)} \qquad \ddot{x}(t) = E(x(t), \dot{x}(t)),$$

which shows that the optimal solutions are smooth.

But, in general, the optimal solution is not unique for every (ξ, T), and it is intuitively clear that one of the reasons for which V may fail to be differentiable lies in this fact. This point will be dealt with in Theorem 1.

We first state the simplest result concerning the value function.

THEOREM 0.1. *V is locally Lipschitz on* $\mathbb{R}^n \times]0, +\infty[$.

LEMMA 0.1. *Let K be a compact subset of $\mathbb{R}^n \times]0, +\infty[$. Then, there exist non-negative constants A_K and B_K such that, for every optimal solution x of $\mathcal{P}_{\xi,T}$ with $(\xi, T) \in K$, we have*

$$\|x(t)\| \leqq A_K \qquad \|\dot{x}(t)\| \leqq B_K \quad \forall t \in [0, T].$$

Proof. See [9]. (This is an unpublished result of I. Ekeland.)

COROLLARY. *Let K be a compact subset of $\mathbb{R}^n \times]0, +\infty[$. Then, the family of all optimal solutions of $\mathcal{P}_{\xi,T}$, with $(\xi, T) \in K$, as well as the family of their derivatives, are equi-Lipschitz.*

Proof. Immediate consequence of Lemma 0.1 and the (E. L.) equation.

Proof of Theorem 0.1. Let (ξ, T) and $(\xi'_, T')$ be two points in K, x and x' optimal solutions of $\mathcal{P}_{\xi,T}$ and $\mathcal{P}_{\xi,T'}$, respectively.

We associate with x and x' the path y defined on $[0, T]$ by

$$y(t) = x'\left(\frac{t}{T} T'\right) + \frac{t}{T}(\xi - \xi').$$

Since $y(0) = \xi_0$ and $y(T) = \xi$, we have

$$(*) \qquad \int_0^T f(y(t), \dot{y}(t))\, dt \geqq V(\xi, T).$$

Using Lemma 0.1 and the fact that f is locally Lipschitz, we can easily find constants M and N (depending only on K) such that

$$\int_0^T f(y, \dot{y})\, dt \leqq \int_0^{T'} f(x', \dot{x}')\, dt + M\|\xi - \xi'\| + N|T - T'|,$$

which, taking $(*)$ into account, yields

$$V(\xi, T) - V(\xi', T') \leqq M\|\xi - \xi'\| + N|T - T'|.$$

We complete the proof by introducing

$$y'(t) = x\left(\frac{t}{T'} T\right) + \frac{t}{T'}(\xi' - \xi),$$

which leads to

$$V(\xi', T') - V(\xi, T) \leqq M\|\xi - \xi'\| + N|T - T'|. \qquad \text{Q.E.D.}$$

0.2. Local ε-supports. Let E be a Banach space, E^* its topological dual. Let $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ denote respectively the bracket: $E^*\times E\to\mathbb{R}$ and the norm on E.

DEFINITION 0.2. *Let f be a real valued function on E, and let $\varepsilon>0$. We shall say that f is locally ε-supported at $x_0\in E$, if there exists an open neighborhood $\mathcal{U}$ of x_0 and a bounded linear functional u_0^*, such that*

$$f(x)-f(x_0)\geqq\langle u_0^*, x-x_0\rangle-\varepsilon\|x-x_0\| \quad \forall x\in\mathcal{U}.$$

This definition can be looked upon as a weak differentiability property. For more details, see [6]. Its main interest lies in the following theorem:

THEOREM 0.2. *Suppose that the mapping $x\to\|x\|$: $E\to\mathbb{R}^+$ is Fréchet differentiable on $E-\{0\}$. Then, for every lower semicontinuous function $f:E\to\mathbb{R}$, and for every $\varepsilon>0$, the set of points where f is locally ε supported, is dense in E.*

Remark. The assumption made upon E covers the case of spaces $\mathbb{R}^n$ and more generally, of all Hilbert spaces, but also of many standard spaces of analysis, for example L^p, with $1<p<+\infty$.

We shall apply Theorem 0.2 to the value function V in order to derive a uniqueness result about the optimal solutions.

1. Generic uniqueness and differentiability of the value function. We keep the notations and hypotheses of the preliminaries and state now a first result about generic uniqueness.

PROPOSITION 1. *The set of points $(\xi, T)\in\mathbb{R}^n\times]0,+\infty[$, such that there exists a unique solution to the problem $\mathcal{P}_{\xi,T}$, contains a dense G_δ.*

Remark. If $f(x,\dot{x})=\langle a(x)\dot{x},\dot{x}\rangle$, where $a(x)$ is a positive definite matrix, Proposition 1 is nothing else but a well-known result of Riemannian geometry, which holds in infinite dimension (see [4]).

The proof of Proposition 1 will involve two lemmas.

LEMMA 1.1. *Let K be a compact subset of $\mathbb{R}^n\times]0,+\infty[$. Then there exists a nonnegative constant M_K such that, for every $\varepsilon>0$, if V is locally ε-supported at $(\xi, T)\in K$, the following inequality holds for every pair (x_1,x_2) of optimal solutions of $\mathcal{P}_{\xi,T}$:*

$$\|\dot{x}_1(T)-\dot{x}_2(T)\|\leqq M_K\varepsilon.$$

Proof. Suppose V is locally ε-supported at $(\xi, T)\in K$. There exist $\alpha>0$, $u\in\mathbb{R}^n$, $\theta\in\mathbb{R}$, such that

$$\|\xi-\xi'\|+|T-T'|\leqq\alpha\Rightarrow V(\xi',T')-V(\xi,T)\geqq\langle u,\xi'-\xi\rangle$$
$$+\theta(T'-T)-\varepsilon(\|\xi-\xi'\|+|T-T'|).$$

Let x be an optimal solution of $\mathcal{P}_{\xi,T}$. With every ξ' such that $\|\xi'-\xi\|\leqq\alpha$, let us associate the path $y_{\xi'}$ defined on $[0,T]$ by

$$y_{\xi'}(t)=x(t)+\frac{t}{T}(\xi'-\xi).$$

We have

$$y_{\xi'}(0)=x(0)=\xi_0,$$
$$y_{\xi'}(T)=x(T)+(\xi'-\xi)=\xi';$$

[1] In other words, a countable intersection of open dense subsets.

therefore,

$$V(\xi', T') \leqq \int_0^T f(y_{\xi'}, \dot{y}_{\xi'})\,dt,$$

which gives, using the first inequality,

$$\int_0^T \left\{ f\left(x(t)+\frac{t}{T}(\xi'-\xi),\, \dot{x}(t)+\frac{\xi'-\xi}{T}\right) - f(x(t), \dot{x}(t)) \right\} dt \geqq \langle u, \xi'-\xi\rangle - \varepsilon\|\xi-\xi'\|.$$

Now, by Lemma 0.1, there is a $\beta > 0$, depending only on K such that, if $\|\xi-\xi'\| \leqq \beta$,

$$\int_0^T \left\{ f\left(x(t)+\frac{t}{T}(\xi'-\xi),\, \dot{x}(t)+\frac{\xi'-\xi}{T}\right) - f(x(t), \dot{x}(t)) \right\} dt$$

$$\leqq \int_0^T \left\langle f'_\zeta(x(t), \dot{x}(t)), \frac{t}{T}(\xi'-\xi)\right\rangle dt + \int_0^T \left\langle f'_\eta(x(t), x(t)), \frac{\xi'-\xi}{T}\right\rangle dt + \varepsilon\|\xi'-\xi\|.$$

Using the Euler equation, we can express the right-hand side integrals in the form

$$\left[f'_\eta(x(t), \dot{x}(t)), \frac{t}{T}(\xi'-\xi)\right]_0^T = \langle f'_\eta(\xi, \dot{x}(T)), \xi'-\xi\rangle,$$

so that, for every ξ' in a suitable neighborhood of ξ,

$$\langle u - f'_\eta(x(T), \dot{x}(T)), \xi'-\xi\rangle \leqq 2\varepsilon\|\xi'-\xi\|.$$

which implies

$$\|u - f'_\eta(x(T), \dot{x}(T))\| \leqq 2\varepsilon.$$

In particular, we have, for any pair (x_1, x_2) of optimal solutions of $\mathcal{P}_{\xi,T}$:

$$\|f'_\eta(\xi, \dot{x}_1(T)) - f'_\eta(\xi, \dot{x}_2(T))\| \leqq 4\varepsilon.$$

Now, the existence of M_K follows easily from the assumption (iii) made upon f, and from Lemma 0.1.

COROLLARY. *Suppose V is differentiable at the point (ξ, T). Then there exists a unique solution to the problem $\mathcal{P}_{\xi,T}$.*

Proof. V is, in this case, locally ε-supported at (ξ, T), for every $\varepsilon > 0$. Therefore, we have $\dot{x}_1(T) = \dot{x}_2(T)$, for any pair of optimal solutions. Since, on the other hand, $x_1(T) = x_2(T) = \xi$, and x_1 and x_2 both satisfy the (E. L.) equation, we have $x_1 = x_2$ by the uniqueness theorem for the Cauchy problem.

We now introduce the following

Notation. Let K be a compact subset of $\mathbb{R}^n \times]0, +\infty[$ and θ a constant >0. We denote by $R_{\theta,K}$ the set of points $(\xi, T) \in K$ such that, for every pair (x_1, x_2) of optimal solutions of $\mathcal{P}_{\xi,T}$, the inequality $\|\dot{x}_1(T) - \dot{x}_2(T)\| < \theta$ holds.

By Lemma 1.1 and Theorem 0.2., $R_{\theta,K}$ is a dense subset of K, for every $\theta > 0$. Furthermore, consider the following lemma:

LEMMA 1.2. *For every $\theta > 0$ $R_{\theta,K}$ is an open subset of K.*

Proof. Suppose the contrary holds. Then there exist a $\theta > 0$, $(\xi, T) \in R_{\theta,K}$, a sequence (ξ_n, T_n) in K, converging to (ξ, T), and, for every integer n, two optimal solutions x_n^1 and x_n^2 of ξ_{ξ_n,T_n} such that

$$\|\dot{x}_n^1(T_n) - \dot{x}_n^2(T_n)\| \geqq \theta.$$

By Lemma 0.1, there exist constants A and $B>0$ such that

$$\|x_n^i(t)\| \leqq A, \qquad \|\dot{x}_n^i(t)\| \leqq B \qquad \forall n \in N, \quad \forall t \in [0, T_n], \quad i=1,2,$$

so that we may assume $(\dot{x}_n^i(T_n))$ converging to some limit v_i.

Let y_1 and y_2 denote respectively the maximal solutions of the (E.L.) equation with initial conditions:

$$y_1(T)=\xi, \qquad \dot{y}_1(T)=v_1,$$

$$y_2(T)=\xi, \qquad \dot{y}_2(T)=v_2,$$

and $]T-a_i, T+b_i[$ the interval where y_i is defined.

We shall prove that $T-a_i<0$ and the restriction of y_i to $[0, T]$ is optimal, which will contradict the fact that (ξ, T) belongs to $R_{\theta,K}$, since

$$\|v_1-v_2\| = \lim \|\dot{x}_n^1(T_n)-\dot{x}_n^2(T_n)\| \geqq \theta.$$

First of all, set

$$z_n^i(t)=x_n^i(t+(T_n-T)) \quad \text{for } t \in [T-T_n, T].$$

The z_n^i are solutions of the (E. L.) equation and

$$\lim z_n^i(T) = \lim x_n^i\ (T_n)=\xi,$$

$$\lim \dot{z}_n^i(T)=\lim \dot{x}_n^i(T_n)=v_i.$$

Therefore, for every $\varepsilon>0$, z_n^i may be extended to a solution of the (E. L.) equation on $[T-T_n, T] \cup [T-a_i+\varepsilon,\ T+b_i-\varepsilon]$ when n is big enough, and the sequence z_n^i (respectively $\dot{z}_n^i$) converges to y_i (respectively $\dot{y}_i$) uniformly on $[T-a_i+\varepsilon, T+b_i-\varepsilon]$.

Recall now that, by the corollary of Lemma 0.1, x_n^i and $\dot{x}_n^i$ are equi-Lipschitz.

We derive the following from the preceding assertions and Ascoli's theorem:

(1) For every $\varepsilon>0$, x_n^i may be extended to a solution of the (E. L.) equation on $[0, T_n] \cup [T-a_i+\varepsilon,\ T+b_i-\varepsilon]$ when n is big enough.

(2) The sequence x_n^i (respectively $\dot{x}_n^i$) converges to y_i (respectively $\dot{y}_i$) uniformly on $[T-a_i+\varepsilon, T+b_i-\varepsilon]$.

(3) There exists subsequences of x_n^i and $\dot{x}_n^i$ converging uniformly on $[0, T]$.

This last assertion shows that $T-a_i<0$ (recall that y_i is a *maximal* solution of the (E. L.) equation).

The last point to be proved is that y_i is optimal. We have

$$y_i(T)=\xi, \quad \text{by assumption,}$$

$$y_i(0)=\lim x_n^i(0)=\xi_0,$$

and

$$\int_0^T f(y_i, \dot{y}_i)\, dt = \lim \int_0^{T_n} f(x_n^i, \dot{x}_n^i)\, dt = \lim\ V(\xi_n, T_n) = V(\xi, T),$$

because of the uniform convergence on $[0, T]$ and the continuity of V. Q.E.D.

Proof of Proposition 1. Let K be a compact subset of $\mathbb{R}^n \times]0, +\infty[$. By Lemmas 1.1 and 1.2, $R_{1/n,K}$ is an open dense subset of K, for every integer n. Therefore, $G=\bigcap_{n \in \mathbb{N}^*} R_{1/n,K}$ is a dense G_δ of K. If $(\xi, T) \in G$ and (x_1, x_2) is a pair of optimal solutions of $\mathscr{P}_{\xi,T}$, we have $\dot{x}_1(T)=\dot{x}_2(T)$ and consequently $x_1=x_2$.

What follows now is standard. The relationship between the differentiability of V and the uniqueness of optimal solutions being very close, Proposition 1 will imply that V is differentiable on an open dense subset.

First recall:

DEFINITION 1.1. *Let* $(\xi, T) \in \mathbb{R}^n \times]0, +\infty[$, *$x$ an optimal solution of* $\mathcal{P}_{\xi,T}$ *and* $v = \dot{x}(0)$. *We can define a mapping ϕ from a suitable neighborhood of* (v, T) *to* $\mathbb{R}^n \times]0, +\infty[$ *by setting*

$$\phi(w, S) = (x_w(S), S),$$

where x_w is the solution of the (E. L.) equation with initial conditions $x_w(0) = \xi_0$, $\dot{x}_w(0) = w$. (This makes sense when (w, S) is sufficiently close to (v, T), x_w being then well defined on the whole interval $[0, S]$.)

We shall say that (ξ, T) is conjugate to $(\xi_0, 0)$ along x (or x is a degenerate solution of $\mathcal{P}_{\xi,T}$) if

$$\operatorname{Jac} \phi(v, T) = 0.$$

We are now able to prove the following:

THEOREM 1. *There exists an open dense subset Ω of* $\mathbb{R}^n \times]0, +\infty[$ *such that*:

(a) *If* $(\xi, T) \in \Omega$, $\mathcal{P}_{\xi,T}$ *admits a unique optimal solution and this optimal solution is nondegenerate.*

(b) *The restriction of V to Ω is* C^∞.

Proof. Let Γ_1 be the set of all points (ξ, T) which are not conjugate to $(\xi_0, 0)$ along any optimal solution. By Sard's theorem, Γ_1 contains a dense G_δ of $\mathbb{R}^n \times]0, +\infty[$ (see [1]).

On the other hand, if Γ_2 denotes the set of all points (ξ, T) such that $\mathcal{P}_{\xi,T}$ admits a unique optimal solution, Γ_2 contains a dense G_δ by Proposition 1 and, therefore, so does $\Gamma_1 \cap \Gamma_2$.

We shall prove that (a) and (b) hold when Ω is a suitable neighborhood of $\Gamma_1 \cap \Gamma_2$.

Take $(\bar{\xi}, \bar{T}) \in \Gamma_1 \cap \Gamma_2$ and let $\bar{x}$ be the unique optimal solution of $\mathcal{P}_{\bar{\xi}\bar{T}}$. Set $\bar{v} = \dot{\bar{x}}(0)$.

There exists an open neighborhood $\mathcal{U}$ of $\bar{v}$ and an $\varepsilon > 0$ such that the solution x_w of the (E. L.) equation with initial conditions

$$x_w(0) = \xi_0, \qquad \dot{x}_w(0) = w,$$

is well defined on $[0, S]$ when $(w, S) \in \mathcal{U} \times]\bar{T} - \varepsilon, \bar{T} + \varepsilon[$.

Furthermore, we can choose $\mathcal{U}$ and ε so that ϕ (defined in Definition 1.1) is a C^∞ diffeomorphism from $\mathcal{U} \times]\bar{T} - \varepsilon, \bar{T} + \varepsilon[$ onto some neighborhood $\mathcal{V} \times]\bar{T} - \varepsilon, \bar{T} + \varepsilon[$ of $(\bar{\xi}, \bar{T})$, with inverse ψ.

We claim that the proof is reduced to the following:

SUBLEMMA. *There exists an open neighborhood $\mathcal{W}$ of $(\bar{\xi}, \bar{T})$ such that, for every point (ξ, S) in $\mathcal{W}$ and every optimal solution y of $\mathcal{P}_{\xi,S}$, we have*

$$(\dot{y}(0), S) \in \mathcal{U} \times]\bar{T} - \varepsilon, \bar{T} + \varepsilon[.$$

For, if this is proved, we shall have by definition of ψ

$$V(\xi, S) = J \circ \psi(\xi, S) \quad \forall (\xi, S) \in \mathcal{W},$$

where J assigns to each (w, S) the scalar $\int_0^S f(x_w, \dot{x}_w)\, dt$. This will yield (a) and (b).

Proof of the Sublemma. Suppose the contrary holds. Then, there would exist a sequence (ξ_n, S_n) converging to $(\bar{\xi}, \bar{T})$ and, for every integer n, an optimal solution z_n of $\mathcal{P}_{\xi_n, S_n}$ such that

$$\dot{z}_n(0) \notin \mathcal{U}.$$

By the same kind of argument as in Lemma 1.2, this would define an optimal solution $\bar{z}$ of $\mathcal{P}_{\bar{\xi},\bar{T}}$, such that $\dot{\bar{z}}(0) \notin \mathcal{U}$ and therefore $\bar{z} \neq \bar{x}$, which is impossible since $(\bar{\xi}, \bar{T})$ lies in Γ_2.

Notice at last that at any point where V is differentiable (it results from several steps of this chapter, but it is a well-known fact) we have

$$\nabla_\xi V(\xi, T) = f'_\eta(\xi, \dot{x}(T)),$$

$$\frac{\partial V}{\partial T}(\xi, T) = f(\xi, \dot{x}(T)) - \langle f'_\eta(\xi, \dot{x}(T)), \dot{x}(T)\rangle,$$

where x is the unique optimal solution to $\mathcal{P}_{\xi,T}$, so that V is a solution of the Hamilton-Jacobi-Bellman equation:

$$\frac{\partial V}{\partial T}(\xi, T) + \operatorname*{Max}_{u \in \mathbb{R}^n} \{\langle \nabla_\xi V(\xi, T), u\rangle - f(\xi, u)\} = 0.$$

2. Generic singularities. We are now going to study the complement of the open dense subset Ω defined in Theorem 1. Let us give first a precise definition.

DEFINITION 2.1. *We shall say that a point* $(\bar{\xi}, \bar{T}) \in \mathbb{R}^n]0, +\infty[$ *is a singularity* (*or a singular point*) *of the problem* $\mathcal{P}$ *in one of the following cases*:

— *either the problem* $\mathcal{P}_{\bar{\xi},\bar{T}}$ *admits at least two distinct optimal solutions,*

— *or* $(\bar{\xi}, \bar{T})$ *is conjugate to* $(\xi_0, 0)$ *along an optimal solution of* $\mathcal{P}_{\bar{\xi},\bar{T}}$.

A nonsingular point will be of course called a regular point. The set of all singular points is included in a nowhere dense closed subset of $\mathbb{R}^n \times]0, +\infty[$ (Theorem 1), and, as we shall see now, this set cannot be crossed by an optimal solution.

PROPOSITION 2.1. *Let* (ξ, T) *be a singular point of the problem* $\mathcal{P}$ *and* x *an optimal solution of* $\mathcal{P}_{\xi,T}$. *Suppose* x' *is a solution of the* (E. L.) *equation on the interval* $[0, T+\varepsilon]$ (*where* $\varepsilon > 0$), *whose restriction to* $[0, T]$ *coincides with* x. *Then* x' *is not an optimal solution of the problem* $\mathcal{P}_{x'(T+\varepsilon),T+\varepsilon}$.

Proof. In the second case of Definition 2.1, Proposition 2.1 is a well-known theorem of Jacobi (see [8]).

Suppose $\mathcal{P}_{\xi,T}$ admits an optimal solution y distinct from x. Set

$$z(S) = y(S) \quad \text{if } s \in [0, T],$$
$$z(S) = x'(S) \quad \text{if } s \in [T, T+\varepsilon].$$

If x' was an optimal solution of $\mathcal{P}_{x'(T+\varepsilon),T+\varepsilon}$, so would be z, which is impossible since $\dot{z}$ is discontinuous at T.

COROLLARY 1. *If* $\mathcal{P}_{\xi,T}$ *admits two distinct optimal solutions* x *and* y, *we have* $x(t) \neq y(t)$, $\forall t \in]0, T[$.

COROLLARY 2. *The union of the set of singular points and of the point* $(\xi_0, 0)$ *is pathwise connected.*

For, if (ξ, T) is regular, the unique solution x of $\mathcal{P}_{\xi,T}$ leads $(\xi_0, 0)$ to (ξ, T) without crossing any singularity.

Our aim is now to give a description of generic singularities and, if possible, to classify them. Before stating the main theorem, we must introduce a notation and recall a definition.

Let Φ be a convex, monotone function $\mathbb{R}^+ \to \mathbb{R}^+_*$ such that

$$\lim_{t \to +\infty} \frac{\Phi(t)}{t} = +\infty.$$

Let α be a nonnegative real number and n a nonnegative integer. We denote by $\mathscr{A}_\Phi^\alpha(n)$ the set of C^∞ functions $f:\mathbb{R}^n\times\mathbb{R}^n\to\mathbb{R}$ satisfying:

$$\text{(i)}\qquad f(\zeta,\eta)\geqq\Phi(\|\eta\|)\quad\forall(\zeta,\eta)\in\mathbb{R}^n\times\mathbb{R}^n,$$

$$\text{(ii)}\qquad \langle f''_{\eta^2}(\zeta,\eta)\cdot z, z\rangle\geqq\alpha\|z\|^2\quad\forall(\zeta,\eta)\in\mathbb{R}^n\times\mathbb{R}^n,\quad\forall z\in\mathbb{R}^n.$$

Condition (ii) implies the convexity of $\eta\to f(\zeta,\eta)$ and we can apply the results of § 1 to every $f\in\mathscr{A}_\Phi^\alpha(n)$.

$\mathscr{A}_\Phi^\alpha(n)$ is provided with the Whitney C^∞ topology. (A sequence (f_n) is said to converge in $\mathscr{A}_\Phi^\alpha(n)$ if it converges uniformly on any compact subset, and so does each of its derivatives.) It is a complete metric space, therefore a Baire space. A property will be said to hold almost everywhere (or generically) in $\mathscr{A}_\Phi^\alpha(n)$, if it holds on a dense G_δ of $\mathscr{A}_\Phi^\alpha(n)$.

DEFINITION 2.2. (See [11]) *By a Whitney stratification of a p-dimensional C^∞ manifold M, we mean a partition of M into submanifolds Z_i, of codimension i (called the strata) such that*:

(1) $\bar{Z}_i\supset Z_{i+1}$.

(2) *Let* $1\leqq i<j\leqq p$.

For any sequences (x_n) of points in Z_i and (y_n) in Z_j converging to some point x in Z_j, such that the secants x_ny_n converge (in the projective space) and the tangent spaces $T_{x_n}Z_i$ converge (in the Grassmannian), we have

$$D\subset\pi,\quad \textit{where } D=\lim x_ny_n,\quad \pi=\lim T_{x_n}Z_i.$$

THEOREM 2. *Suppose $n\leqq 4$. Then, for almost every f in $\mathscr{A}_\Phi^\alpha(n)$ we have*:

(1) *For every $(\xi, T)\in\mathbb{R}^n\times]0,+\infty[$, the problem $\mathscr{P}_{\xi,T}$ admits at most a finite number of optimal solutions.*

(2) *The space $\mathbb{R}^n\times]0,+\infty[$ is provided with a Whitney stratification whose stratum of codimension 0 is exactly the set of regular points of the problem associated with f.*

We shall specify later the significance of the higher codimension strata. Let us be explicit only for the case $n=1$, illustrated in Fig. 1.

The stratum Z_0 corresponds to the open dense subset Ω of Theorem 1.

The stratum Z_1 consists of points (ξ, T) such that $\mathscr{P}_{\xi,T}$ admits two distinct nondegenerate optimal solutions.

The stratum Z_2 consists of isolated points corresponding to one of the two following cases:

Z_2^c: three distinct nondegenerate optimal solutions.

Z_2^b: one degenerate optimal solution.

Notice that, because of Proposition 2.1, no schema of the type in Fig. 2 may occur.

The proof of Theorem 2 is based on the concepts of codimension and unfolding. The main idea consists of the choice of a suitable gradient model which allows us to interpret $\mathbb{R}^n\times]0,+\infty[-\Omega$ as a catastrophic set. Then, applying Thom's tranvsersality theorem, it is proved that generically only elementary catastrophes may occur. For the basic concepts of unfolding, catastrophe, codimension, transversality, see [1], [2], [12], [13].

The construction which will follow now is, to a great extent, inspired by K. Jänisch [10] who relates the caustics of a wave front to the bifurcation set of a gradient model.

The proof will consist of five steps.

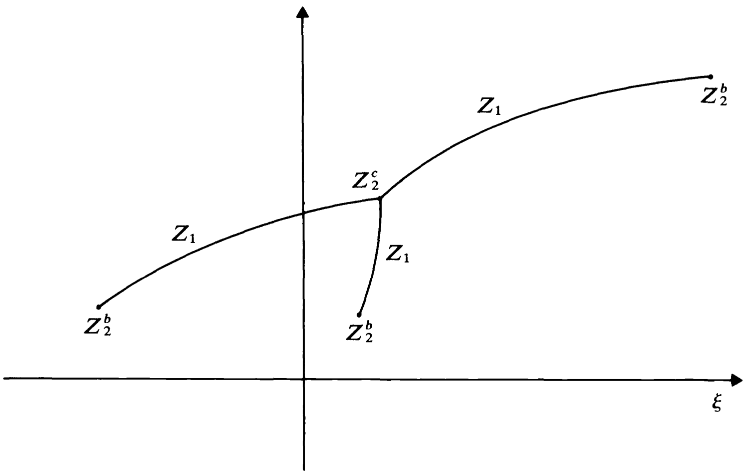


FIG. 1.

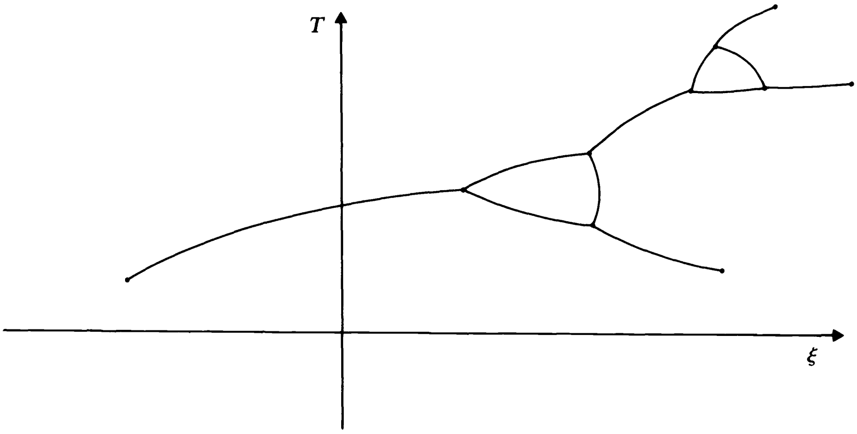


FIG. 2

(A) We first fix $f \in \mathscr{A}^\alpha_\Phi(n)$. Let $(\bar\xi, \bar T) \in \mathbb{R}^n \times]0, +\infty[$ and $\bar x$ be an optimal solution of $\mathscr{P}_{\bar\xi,\bar T}$. Fix $s_0 \in]0, T[$ and set

$$\Sigma_0 = \{(\zeta, S) \in \mathbb{R}^n \times]0, +\infty[,\ V(\zeta, S) = V(\bar x(s_0), s_0)\}.$$

By Proposition 2.1, V is differentiable on a neighborhood of $(\bar x(s_0), s_0)$ and $\bar x$ is the unique solution of $\mathscr{P}_{\bar x(s_0), s_0}$. (Note that $(\bar\xi, \bar T)$ need not be a regular point.) We have:

$$\nabla_\xi V(\bar x(s_0), s_0) = f'_\eta(\bar x(s_0), \dot{\bar x}(s_0)),$$

$$\frac{\partial V}{\partial T}(\bar x(s_0), s_0) = f(\bar x(s_0), \dot{\bar x}(s_0)) - \langle \dot{\bar x}(s_0), f'_\eta(\bar x(s_0), \dot{\bar x}(s_0))\rangle,$$

so that, under the assumptions made upon f

$$\left(\nabla_\xi V(\bar x(s_0), s_0), \frac{\partial V}{\partial T}(\bar x(s_0), s_0)\right) \neq 0.$$

Consequently, there exists a neighborhood $\mathcal{U}_0$ of $(\bar{x}(s_0), s_0)$ in $\mathbb{R}^n \times]0, +\infty[$ such that $\Sigma_0 \cap \mathcal{U}_0$ is a 1-codimensional submanifold of $\mathbb{R}^n \times]0, +\infty[$.

(B) Note now that, if $\mathcal{U}_0$ is sufficiently small, we can find a neighborhood $\mathcal{V}_0$ of $(\bar{\xi}, \bar{T})$ such that, for every $(\xi, T) \in \mathcal{V}_0$ and $(\zeta, S) \in \Sigma_0 \cap \mathcal{U}_0$, the problem

$$\begin{cases} \inf \int_S^T f(x, \dot{x})\, dt, \\ x(S) = \zeta, \qquad x(T) = \xi \end{cases}$$

admits a unique optimal solution.

Furthermore, the value function $W((\zeta, S), (\xi, T))$ of this problem is smooth on the product $(\Sigma_0 \cap \mathcal{U}_0) \times \mathcal{V}_0$. (Since the initial condition is not fixed, this is not exactly the result of § 1, but it may be proved easily by the same kind of arguments.) See Fig. 3.

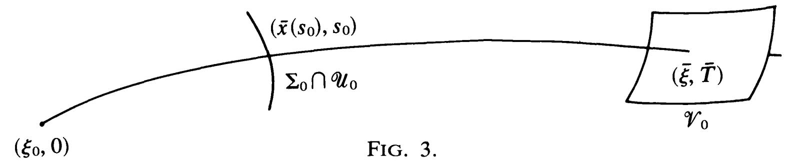


FIG. 3.

The interest of making the initial condition run over $\Sigma_0 \cap \mathcal{U}_0$ lies in the following fact.

Let $(\xi, T) \in \mathcal{V}_0$. Suppose x is an optimal solution of $\mathcal{P}_{\xi,T}$ such that there exists $\bar{t} \in]0, T[$, with $(x(\bar{t}), \bar{t}) \in \Sigma_0 \cap \mathcal{U}_0$ (we shall say that x crosses $\Sigma_0 \cap \mathcal{U}_0$).

Then, by Bellman's optimality principle,

$$W((x(\bar{t}), \bar{t}), (\xi, T)) = \operatorname*{Min}_{(\zeta, S) \in \Sigma_0 \cap \mathcal{U}_0} W((\zeta, S), (\xi, T)).$$

(Recall V is constant on $\Sigma_0 \cap \mathcal{U}_0$.)

So, when $\mathcal{P}_{\xi,T}$ admits two distinct optimal solutions crossing $\Sigma_0 \cap \mathcal{U}_0$ (with $(\xi, T) \in \mathcal{V}_0$), the function $W((\cdot, \cdot), (\xi, T))$ attains its minimum at two distinct points.

Likewise, when a point $(\xi, T) \in \mathcal{V}_0$ is conjugate to $(\xi_0, 0)$ along an optimal solution crossing $\Sigma_0 \cap \mathcal{U}_0$, $W((\cdot, \cdot), (\xi, T))$ has a degenerate minimum on $\Sigma_0 \cap \mathcal{U}_0$.

Recall now that, if $W((\cdot, \cdot), (\xi, T))$ attains its minimum at k distinct points (ζ_i, S_i) and if r_i denotes respectively the codimension of the germ of $W((\cdot, \cdot), (\xi, T))$ at (ζ_i, S_i), we have by definition,

$$\operatorname{codim} W((\cdot, \cdot), (\xi, T)) = k - 1 + \sum_{i=1}^{k} r_i.$$

We can specify the singularities which occur when $\operatorname{codim} W((\cdot, \cdot), (\xi, T)) = p \leqq 5$. The case $p = 0$ corresponds to one nondegenerate minimum, $p = 1$ to two nondegenerate minima, $p = 2$ to three nondegenerate minima or to a cusp, $\cdots$.

We have only to consider here singularities associated with local minima: nondegenerate, cusp, butterfly. Swallow tail, mushroom and wave are excluded (see [2], [12] and [13]).

(C) LEMMA 2.2.1. *We can suppose $\mathcal{U}_0$ and $\mathcal{V}_0$ sufficiently small so that, for every $(\xi, T) \in \mathcal{V}_0$, every (ζ_1, S_1) and $(\zeta_2, S_2) \in \Sigma_0 \cap \mathcal{U}_0$, and every x_1 and x_2 such that*

$$\begin{matrix} x_i(S_i) = \zeta_i, \\ x_i(T) = \xi, \end{matrix} \qquad \int_{S_i}^T f(x_i, \dot{x}_i)\, dt = W((\zeta_i, S_i), (\xi, T)),$$

we have, whenever $(\zeta_1, S_1) \neq (\zeta_2, S_2)$,

$$(x_1(t_1), \dot{x}_1(t_1)) \neq (x_2(t_2), \dot{x}_2(t_2)) \quad \forall t_1, t_2.$$

Proof. Suppose there exist t_1 and t_2 with

$$x_1(t_1) = x_2(t_2), \qquad \dot{x}_1(t_1) = \dot{x}_2(t_2).$$

Then, since x_1 and x_2 are both solutions of the (E. L.) equation,

$$x_1(t + (t_1 - t_2)) = x_2(t),$$

which shows, since $x_1(T) = \xi$, $x_2(T) = \xi$,

$$t_1 = t_2, \qquad x_1(t) = x_2(t) \quad \forall t \geqq \max(S_1, S_2).$$

But, since the mapping $t \to (\bar{x}(t), t)$ is transversal to $\Sigma_0 \cap \mathcal{U}_0$ at $(\bar{x}(s_0), s_0)$, this cannot occur, whenever $\mathcal{U}_0$ and $\mathcal{V}_0$ are small enough.

(D) Before stating the chief lemma, let us introduce some convenient notations. Let J be the ring of polynomials in n variables of degree $\leqq 8$. The product space J^7 is provided with a stratification whose ith stratum Q_1 is nothing but the space of multijets of codimension i. We denote by $\mathcal{S}_i$ the closed subset of Q_i consisting of the singular multijets associated with local minima, for $1 \leqq i \leqq 5$, and R the set of all multisingularities of codimension $\geqq 6$. (For a precise study of these sets, see [13].)

LEMMA 2.2.2. *For every* $(\xi, T) \in \mathcal{V}_0$ *and* $(\zeta, S) \in \mathcal{U}_0 \cap \Sigma_0$, *let* $jW((\zeta, S), (\xi, T))$ *denote the* 8*-jet of* W *with respect to* (ζ, S). *Then, for almost every* $f \in \mathcal{A}^{\alpha}_{\Phi}(n)$, *the mapping*

$$((\xi, T), (\zeta_1, S_1), \cdots, (\zeta_7, S_7)) \to (jW((\zeta_1, S_1), (\xi, T)), \cdots, jW(\zeta_7, S_7), (\xi, T)),$$

$$\mathcal{V}_0 \times (\Sigma_0 \cap \mathcal{U}_0)^7 \to J^7,$$

is transversal to the $\mathcal{S}_i$*'s and to* R, *whenever the* (ζ_i, S_i)*'s are pairwise distinct.*

Proof. We shall show that, if (ξ, T) is fixed, the mapping

$$(f, (\zeta_i, S_i)_{1 \leqq i \leqq 7}) \to (JW((\zeta_i, S_i), (\xi, T)))_{1 \geqq i \leqq 7}$$

is a submersion. For it will follow that the mapping

$$(f, (\xi, T), (\zeta_i, S_i)) \to (JW((\zeta_i, S_i), (\xi, T)))$$

is a submersion and the lemma will then be an easy consequence of Thom's transversality theorem.

Take $(\xi, T) \in \mathcal{V}_0$. For every $(\zeta, S) \in \Sigma_0 \cap \mathcal{U}_0$, let $x_{\zeta,S}$ be the unique path satisfying

$$x_{\zeta,S}(S) = \zeta, \qquad x_{\zeta,S}(T) = \xi,$$

and

$$\int_S^T f(x_{\zeta,S}, \dot{x}_{\zeta,S})\, dt = W((\zeta, S), (\xi, T)).$$

Let us perturb f by δf and let $x_{\zeta,S} + \delta x_{\zeta,S}$ be the new path associated with (ζ, S). We have, omitting the second order terms,

$$\delta W((\zeta, S), (\xi, T)) = \int_S^T \delta f(x_{\zeta,S}, \dot{x}_{\zeta,S})\, dt + \int_S^T \langle f'_x(x_{\zeta,S}, \dot{x}_{\zeta,S}) \delta x_{\zeta,S} \rangle\, dt$$

$$+ \int_S^T \langle f'_{\dot{x}}(x_{\zeta,S}, \dot{x}_{\zeta,}\ {}_S), \delta \dot{x}_{\zeta,S} \rangle\, dt$$

$$= \int_S^T \delta f(x_{\zeta,S}, \dot{x}_{\zeta,S})\, dt + [\langle f'_x(x_{\zeta,S}, \dot{x}_{\zeta,S}), \delta x_{\zeta,S} \rangle]_S^T.$$

But, since, for every f, $x_{\zeta,S}(S) = \zeta$, $x_{\zeta,S}(T) = \xi$,

$$\delta x_{\zeta,S}(S) = \delta x_{\zeta,S}(T) = 0,$$

and, hence

$$\delta W((\zeta, S), (\xi, T)) = \int_S^T \delta f(x_{\zeta,S}, \dot{x}_{\zeta,S})\, dt.$$

Finally, we only have to show that the mapping, which assigns to g the 8 order jet of the function

$$\psi(\zeta, S) = \int_S^T g(x_{\zeta,S}, \dot{x}_{\zeta,S})\, dt,$$

is surjective. By part (C), there exists a scalar $t_0 \in]s_0, T[$ and an $\varepsilon > 0$, such that the set

$$\{(x_{\zeta,S}(t), \dot{x}_{\zeta,S}(t)), (\zeta, S) \in \Sigma_0 \cap \mathcal{U}_0, t \in]t_0, t_0 + \varepsilon[\}$$

is a submanifold diffeomorphic to $(\Sigma_0 \cap \mathcal{U}_0) \times]t_0, t_0 + \varepsilon[$.

If ψ is fixed, we can choose a C^∞ function g such that

$$g(x_{\zeta,S}(t), \dot{x}_{\zeta,S}(t)) = \psi(\zeta, S)\phi(t),$$

where ϕ is a function with compact support included in $]t_0, t_0 + \varepsilon[$, satisfying

$$\int_{t_0}^{t_0+\varepsilon} \phi(u)\, du = 1.$$

We have then

$$\int_S^T g(x_{\zeta,S}, \dot{x}_{\zeta,s})\, dt = \psi(\zeta, S) \quad \forall (\zeta, S) \in \Sigma_0 \cap \mathcal{U}_0. \qquad \text{Q.E.D.}$$

COROLLARY. *For almost every $f \in \mathcal{A}_\Phi^\alpha(n)$, the number of solutions of $\mathcal{P}_{\xi,T}$ is finite, for every $(\xi, T) \in \mathbb{R}^n \times]0, +\infty[$.*

Proof. By (B) and Lemma 2.2.2, for almost every f, there is a finite number of solutions to $\mathcal{P}_{\xi,T}$ crossing $\Sigma_0 \cap \mathcal{U}_0$, for every $(\xi, T) \in \mathcal{V}_0$. But, since $\Sigma_0 \cap \{(\zeta, S), s_0 \leqq S \leqq \bar{T}\}$ is compact, and since $\mathbb{R}^n \times]0, +\infty[$ can be covered by a countable infinity of neighborhoods like $\mathcal{V}_0$, the corollary results from Baire's theorem.

(E) *End of the proof.* The first part of Theorem 2 has just been demonstrated. The last difficulty lies in the fact that the optimal solutions of $\mathcal{P}_{\xi,T}$, where $(\xi, T) \in \mathcal{V}_0$, need not cross $\Sigma_0 \cap \mathcal{U}_0$; hence, the function W, as defined in (B), doesn't provide us exactly with the gradient model needed.

Anyway, let $(\bar{\xi}, \bar{T}) \in \mathbb{R}^n \times]0, +\infty[$. We may suppose that there are p and only p solutions to $\mathcal{P}_{\bar{\xi},\bar{T}}$: $\bar{x}_i, \cdots, \bar{x}_p$.

Let us associate with each of the $\bar{x}_i$'s a submanifold $\Sigma_0 \cap \mathcal{U}_i$ defined as in (A), and a mapping W_i,

$$(\Sigma_0 \cap \mathcal{U}_i) \times \mathcal{V}_0 \to \mathbb{R}, \quad \text{as in (B).}$$

Whenever $\mathcal{V}_0$ is sufficiently small, the optimal solutions of $\mathcal{P}_{\xi,T}$, with $(\xi, T) \in \mathcal{V}_0$, cross necessarily one of the submanifolds $\Sigma_0 \cap \mathcal{U}_i$.

Set $\Sigma = \bigcup_{i=1}^p (\Sigma_0 \cap \mathcal{U}_i)$ and define $W: \Sigma \times \mathcal{V}_0 \to \mathbb{R}$ by

$$W|_{(\Sigma_0 \cap \mathcal{U}_i) \times \mathcal{V}_0} = W_i.$$

(Suppose the $\Sigma_0 \cap \mathcal{U}_i$'s are pairwise disjoint.)

Now, the set of singularities belonging to $\mathcal{V}_0$ is exactly composed of the points $(\xi, T) \in \mathcal{V}_0$ such that

$$\operatorname{codim} W((\cdot, \cdot), (\xi, T)) \geqq 1.$$

Applying again Thom's transversality theorem, just as in Lemma 2.2.2, we can show that, for almost every f, the mapping

$$((\xi, T), (\zeta_i, S_i)_{1\leqq i\leqq 7}) \to (jW((\xi, T), (\zeta_i, S_i))_{1\leqq i\leqq 7}$$

is transversal to the $\mathcal{S}_i$'s and to R, whenever the $(\zeta_i, S_{i,})$'s are pairwise disjoint.

This concludes the proof of Theorem 2.

The meaning of the strata of codimension $\geqq 1$ is now clear. There are three types of generic optimal solutions:

Type I: associated with a nondegenerate minimum of $W((\cdot, \cdot), (\xi, T))$.

Type II: associated with a cusp.

Type III: associated with a butterfly.

We thus obtain Table 1.

TABLE 1

Strata	Z_0	Z_1	Z_2		Z_3		Z_4			Z_5			
Number of solutions													
Type I	1	2	3	0	4	1	5	2	0	6	3	1	0
Type II	0	0	0	1	0	1	0	1	0	0	1	0	2
Type III	0	0	0	0	0	0	0	0	1	0	0	1	0

Acknowledgment. This work has been prepared under the helpful guidance of Professor Ivar Ekeland. At the 1978 ICM Congress in Helsinki, the author found that H. Sussman (unpublished) and P. Brunovsky [3] had independently obtained related results, in a C^ω (analytic) setting, instead of a C^∞ (indefinitely differentiable) one. The approach followed here is more in the spirit of [5].

REFERENCES

[1] R. ABRAHAM AND J. ROBBIN, *Transversal Mappings and Flows*, W. A. Benjamin, New York, 1967.

[2] T. BROCKER, *Differentiable Germs and Catastrophes*, Cambridge University Press, Cambridge, 1975.

[3] P. BRUNOVSKY, *Every normal linear system has a regular time-optimal synthesis*, Math. Slovaca, 28, (1978), pp. 81–100.

[4] I. EKELAND: *The Hopf Rinow theorem in infinite dimension*, J. Differential Geometry, to appear.

[5] ———, *Discontinuités de champs hamiltoniens et existence de solutions optimales en calcul des variations*, Publications Mathématiques de l'I.H.E.S., No. 47 (1978), p. 5.32.

[6] I. EKELAND AND G. LEBOURG, *Generic Fréchet differentiability and perturbed optimization problems in Banach spaces*, Trans. Amer. Math. Soc., (1976), p. 223.

[7] I. EKELAND AND R. TEMAM, *Analyse convexe et problèmes variationnels*, Dunod, Gauthier Villars Paris, 1973.

[8] I. M. GELFAND AND S. V. FOMIN, *Calculus of Variations*, Prentice-Hall, Englewood Cliffs, NJ, 1963.

[9] D. HENRI, *Sur l'unicité des solutions optimales en calcul des variations*, Thése de 3e cycle, Université Paris IX - Dauphine, 1978.

[10] K. JANISCH, *Caustics and catastrophes*, Math. Ann., 209 (1974), pp. 161–180.

[11] J. N. MATHER, *Stratifications and mappings*, in Dynamical Systems, Peixoto, Academic Press, New York,

[12] G. WASSERMANN, *Stability of Unfoldings*, Springer-Verlag, Berlin, 1974.

[13] C. ZEEMAN, *Classification of the elementary catastrophes of codimension* $\leqq 5$, in Structural Stability: The Theory of Catastrophes and its Applications, Lecture Notes, 525, pp. 263–327.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 6, November 1980

A SHIFT OPERATOR APPROACH TO BILINEAR SYSTEM THEORY*

ARTHUR E. FRAZHO†

Abstract. Using a transform representation, we present a bilinear realization theory for a Volterra series input–output map. The approach involves the definition of appropriate shift operators on linear spaces associated with the transforms of the kernals in the Volterra series. This approach yields in a very simple manner a theory of minimality and connections with the concepts of span reachability and observability. It also leads to a characterization of finite dimensional realizability in terms of rationality properties of the transforms.

1. Introduction. Shift operators have played an important role in linear operator theory [23], [32], and linear system theory [3], [19], [24]. Our purpose here is to extend these ideas to the study of discrete-time nonlinear dynamical systems of the bilinear type. This extension is carried out by introducing an appropriate abstract state-space and certain nonlinear shift operators defined on this space. The consequence is an elegant and illuminating treatment of the key theoretical issues. New results are obtained, and many prior results are derived in a simple way.

The presentation can be divided into four main parts. The first concerns a transfer function representation for nonlinear input–output maps. This requires a special transform (called the Λ-transform). An advantage of our transfer function is that it is simply related to input–output maps of bilinear and state-affine systems. In the second part, certain shift operators are applied to the transfer function and this results in an abstract shift realization of the corresponding input–output map. The importance of the shift concept in nonlinear realization questions has been noted by [2], [6], [17], [31], and others. The shift operators used here are believed to be new. They are basically linear and nonlinear transformations on a "Fock space." The third part shows how the abstract shift realization can be applied to obtain a theory of minimality, reachability, and observability for bilinear systems. Simple proofs for many of the standard results on minimality, reachability and observability for bilinear systems [7], [11], [26] are included. The last part is devoted to the question of finite dimensional realizations of the bilinear type. It includes regularity conditions on the transfer function, a test that determines whether or not finite dimensional realizations exist, and the exploration of special rational forms that guarantee the existence of a finite dimensional realization.

To make our discussion more explicit, we now introduce some notation and terminology. It forms the basis for our treatment of bilinear systems. Throughout, $\mathcal{U}$, $\mathcal{V}$, $\mathcal{X}$, $\mathcal{Y}$ are linear spaces over the field $\mathcal{K}$. The dimension of these spaces can be finite or infinite, and no topological structure is assumed. The infinite dimensional spaces are handled in the purely algebraic setting [22]. The set of nonnegative integers is denoted by $I \doteq \{0, 1, 2, \cdots\}$, and $I^n \doteq I \times I \times I \times \cdots \times I$ is the n-fold Cartesian product of I.

The input–output maps we consider are formally given by

$$
\begin{aligned}
y_n &= \sum_{i_1=0}^{n} \tilde{\theta}_1(n-i_1)[u_{i_1}] + \sum_{i_2=0,i_1=0}^{n,i_2} \tilde{\theta}_2(i_2-i_1, n-i_2)[u_{i_1}, u_{i_2}] + \cdots \\
&= \sum_{m=1}^{\infty} \sum_{i_m=0,i_{m-1}=0,\cdots,i_1=0}^{n,i_m,\cdots,i_2} \tilde{\theta}_m(i_2-i_1, i_3-i_2, \cdots, n-i_m)[u_{i_1}, u_{i_2}, \cdots, u_{i_m}],
\end{aligned}
\tag{1.1}
$$

* Received by the editors January 11, 1979, and in final revised form March 17, 1980. This research was supported in part by the U.S. Air Force Office of Scientific Research, Air Force Systems Command, under grant AFOSR-77-3158.

† School of Aeronautics and Astronautics, Purdue University, W. Lafayette, Indiana 47907.

where $u_n \in \mathscr{U}$, $y_n \in \mathscr{Y}$ for $n \in I$ and $\tilde{\theta}_m(k_1, k_2, \cdots, k_m)$ is a m-linear operator for all $(k_1, k_2, \cdots, k_m) \in I^m$, $m > 0$. We call $\{\tilde{\theta}_m\}_{m=1}^{\infty}$ the *kernel sequence* for system (1.1). The summation convention on $i_1, \cdots, i_m$ puts the kernels in lower triangular form. As in [6], [25], [26], [29], this turns out to be especially convenient in our subsequent formulation. One can also express input–output maps in symmetric form [7], [10]. Throughout this paper, strictly lower triangular (SLT) kernel sequences will be used. The kernel sequence $\{\tilde{\theta}_m\}_{m=1}^{\infty}$ is *strictly lower triangular* if

$$\tilde{\theta}_m(k_1, \cdots, k_m) = 0 \quad \text{if one or more of the } k_i\text{'s are zero and } m > 0. \tag{1.2}$$

Since we have not defined a topology on $\mathscr{Y}$, the infinite sum in (1.1) does not make any sense. To get around this difficulty it is always assumed that the input sequence $\{u_i\}_{i=0}^{\infty}$ has finite support. If $\{u_i\}$ has support in $[0, k]$ and (1.1) is SLT, then

$$\tilde{\theta}_m(i_2 - i_1, \cdots, n - i_m)[u_{i_1}, \cdots, u_{i_m}] = 0, \quad \text{if } m > k. \tag{1.3}$$

This implies that (1.1) contains only a finite number of nonzero terms. Thus SLT input–output maps are legitimate input–output maps whenever the input sequence has finite support.

The bilinear system we study is given by

$$\begin{aligned} x_{n+1} &= Ax_n + N(u_n)x_n + Bu_n, \\ y_n &= Cx_n, \end{aligned} \tag{1.4}$$

where $u_n \in \mathscr{U}$, $x_n \in \mathscr{X}$, $y_n \in \mathscr{Y}$ for $n \in I$, A, B, C are linear operators on the appropriate spaces, and $N: \mathscr{U} \times \mathscr{X} \to \mathscr{X}$ is a bilinear operator; i.e., for fixed $x \in \mathscr{X}$, $N(u)x$ is a linear operator in u; and for fixed $u \in \mathscr{U}$, $N(u)x$ is a linear operator in x. (For convenience the bilinear operator N for (1.4) is always written as $N(\cdot)$.) System (1.4) is denoted by $\{A, B, C, N, \mathscr{X}\}$. For (1.4) it is always assumed that the initial condition is zero; i.e., $x_0 = 0$.

By recursively computing the solution for $\Sigma = \{A, B, C, N, \mathscr{X}\}$ (with $x_0 = 0$), it is easy to show that Σ generates a SLT input–output map whose first two kernels are given by

$$\begin{aligned} \tilde{\theta}_{1\Sigma}(i+1) &= CA^iB, && \text{if } i \in I, \\ \tilde{\theta}_{1\Sigma}(i) &= 0, && \text{if } i = 0; \end{aligned} \tag{1.5}$$

$$\begin{aligned} &\tilde{\theta}_{2\Sigma}(i+1, j+1)[u_1, u_2] = CA^jN(u_2)A^iBu_1, \\ &\tilde{\theta}_{2\Sigma}(i, j) = 0, \quad \text{if i or } j \text{ is zero.} \end{aligned} \tag{1.6}$$

The general term is

$$\begin{aligned} &\tilde{\theta}_{n\Sigma}(i_1 + 1, \cdots, i_n + 1)[u_1, \cdots, u_n] = CA^{i_n}N(u_n)A^{i_{n-1}}N(u_{n-1}) \cdots N(u_2)A^{i_1}Bu_1, \\ &\tilde{\theta}_{n\Sigma}(k_1, \cdots, k_n) = 0, \quad \text{if one or more of the } k_i\text{'s are zero,} \end{aligned} \tag{1.7}$$

where $(i_1, i_2, \cdots, i_n) \in I^n$, $u_i \in \mathscr{U}$, $n > 0$. In (1.7) the bilinear operator N appears $(n-1)$ times. Thus each $\tilde{\theta}_{n\Sigma}(i_1, \cdots, i_n)[\cdot, \cdot, \cdots, \cdot]$ is a n-linear operator.

Within this framework, we pose the following problem of bilinear realization: given any SLT input–output map (1.1) with kernel sequence $\{\tilde{\theta}_n\}_{n=1}^{\infty}$, find a system $\Sigma = \{A, B, C, N, \mathscr{X}\}$ such that (1.1) is the input–output map for Σ. Specifically, if for Σ

$$\tilde{\theta}_n(i_1, \cdots, i_n) = \theta_{n\Sigma}(i_1, \cdots, i_n) \quad \text{for all } n \geqq 1, \quad (i_1, \cdots, i_n) \in I^n, \tag{1.8}$$

then Σ is called a *realization* of $\{\tilde{\theta}_n\}$. The following additional terminology will be useful.

System $\Sigma = \{A, B, C, N, \mathscr{X}\}$ is *finite dimensional* if dimension $(\mathscr{X}) < \infty$. System Σ is a *minimal realization of* $\{\tilde{\theta}_n\}$ if dimension $(\mathscr{H}) \leqq$ dimension $(\bar{\mathscr{H}})$ for all other realizations $\{\bar{A}, \bar{B}, \bar{C}, \bar{N}, \bar{\mathscr{X}}\}$ of $\{\tilde{\theta}_n\}$. If $\Sigma_i = \{A_i, B_i, C_i, N_i, \mathscr{X}_i\}$, $i = 1, 2$ are two realizations of $\{\tilde{\theta}_n\}$, and there exists a linear operator $H: \mathscr{X}_1 \to \mathscr{X}_2$ such that $HA_1 = A_2H$, $HB_1 = B_2$, $C_2H = C_1$, and $HN_1(u) = N_2(u)H$ for all $u \in \mathscr{U}$, then H *sends* Σ_1 *into* Σ_2. If H is an isomorphism that sends Σ_1 into Σ_2 then Σ_1 is *equivalent* to Σ_2.

To complete this section we summarize the organization of the paper. Section 2 is devoted entirely to notation and the definition of the Λ-transform. In § 3 the Λ-transform is used to introduce the transfer function corresponding to the input–output map (1.1). This leads to the transfer function θ_Σ for the bilinear system $\Sigma = \{A, B, C, N, \mathscr{X}\}$. In § 4 shift operators are applied to a transfer function to obtain a restricted backward shift realization (RBSR) which (abstractly) solves the bilinear realization problem. In § 5 the RBSR is shown to be a minimal realization which is equivalent to all minimal realizations. In § 6 shift operators are used to treat questions of reachability, observability, and minimality. In § 7 necessary and sufficient conditions for a transfer function to admit a finite dimensional bilinear realization are given. Concluding remarks and further references to the literature are given in § 8. In a future paper (Part II) we discuss state-affine systems and realization algorithms.

2. Transform notation and operators. First we introduce some general notation. $\mathscr{K}$ is a field and 1 is the identity element on $\mathscr{K}$. If B is a Hamel basis for the linear space $\mathscr{X}$, then dim $(\mathscr{X})$ is the cardinality of the set B. Let Q be a subset of $\mathscr{X}$; then $\bigvee Q$ denotes the (finite) linear span of the set Q. The linear space of all n-linear operators $T: \mathscr{U}^n \to \mathscr{Y}$ is denoted by $\mathscr{L}(\mathscr{U}^n; \mathscr{Y})$. By convention we set $\mathscr{L}(\mathscr{U}^0; \mathscr{Y}) = \mathscr{Y}$. If $H \in \mathscr{L}(\mathscr{U}; \mathscr{Y})$ then $\mathscr{R}(H)$ is the range of the linear operator H. If $\mathscr{U}_1$ is a linear subspace of $\mathscr{U}$ then $H|\mathscr{U}_1$ is the linear operator H restricted to $\mathscr{U}_1$. The identity operator on $\mathscr{U}$ is denoted by $I_{\mathscr{U}}$. A linear operator is an *isomorphism* if and only if it is one to one and onto.

If $\mathscr{V}_i$ is an infinite set of linear spaces then $\bigoplus_1^\infty \mathscr{V}_i \doteq \mathscr{V}_1 \oplus \mathscr{V}_2 \oplus \mathscr{V}_3 \oplus \cdots$ is the direct sum of the linear spaces $\mathscr{V}_i$, $i \geqq 1$. Clearly $\bigoplus_1^\infty v_i \in \bigoplus_1^\infty \mathscr{V}_i$ if and only if $v_i \in \mathscr{V}_i$ for all $i \geqq 1$. Whenever we write $\bigoplus v_i$, it is understood that the index starts at one; i.e., $\bigoplus v_i = \bigoplus_1^\infty v_i$, or $\bigoplus \mathscr{V}_{n+3} = \mathscr{V}_4 \oplus \mathscr{V}_5 \oplus \mathscr{V}_6 \oplus \cdots$. Sometimes elements in $\bigoplus \mathscr{V}_i$ are represented by an infinite tuple; i.e., $\bigoplus v_i = \{v_1, v_2, v_3, \cdots\} \in \bigoplus \mathscr{V}_i$. The notation $\bigoplus v_i$ and $\{v_1, v_2, \cdots\}$ is used interchangeably to represent the same element in $\bigoplus \mathscr{V}_i$. Addition and scalar multiplication on $\bigoplus \mathscr{V}_i$ are defined respectively in the usual way: $\bigoplus u_i + \bigoplus v_i = \bigoplus (u_i + v_i)$; $\alpha(\bigoplus v_i) = \bigoplus (\alpha v_i)$, where $\bigoplus u_i, \bigoplus v_i \in \bigoplus \mathscr{V}_i$ and $\alpha \in \mathscr{K}$.

For $n > 0$ the linear space of all sequences $\bar{v}_n$ from I^n to $\mathscr{V}$ is denoted by $\tilde{\mathscr{S}}_n(\mathscr{V})$. If $n = 0$ then $\mathscr{S}_n(\mathscr{V}) \doteq \mathscr{V}$. Addition and scalar multiplication on $\tilde{\mathscr{S}}_n(\mathscr{V})$ are defined in the usual pointwise fashion: $(\tilde{v}_n + \tilde{u}_n)(i_1, \cdots, i_n) \doteq \tilde{v}_n(i_1, \cdots, i_n) + \tilde{u}_n(i_1, \cdots, i_n)$; $(\alpha \tilde{v}_n)(i_1, \cdots, i_n) \doteq \alpha \tilde{v}_n(i_1, \cdots, i_n)$ where $\tilde{v}_n, \tilde{u}_n \in \tilde{\mathscr{S}}_n(\mathscr{V})$, $\alpha \in \mathscr{K}$ and $(i_1, \cdots, i_n) \in I^n$.

For each $n > 0$ we define the linear space $\mathscr{S}_n(\mathscr{V})$ by the set of all formal series v_n such that

$$v_n(\lambda_1, \cdots, \lambda_n) = \sum_{i_1=0, \cdots, i_n=0}^{\infty, \cdots, \infty} \tilde{v}(i_1, \cdots, i_n) \lambda_1^{i_1} \cdots \lambda_n^{i_n} \tag{2.1}$$

where $\tilde{v}_n \in \tilde{\mathscr{S}}_n(\mathscr{V})$ and $\lambda_1, \cdots, \lambda_n$ are the indeterminates. If $n = 0$, then $\mathscr{S}_n(\mathscr{V}) \doteq \mathscr{V}$. Let Λ_n be the linear mapping defined by (2.1); i.e., $\Lambda_n \tilde{v}_n = v_n$. Then Λ_n is an isomorphism from $\tilde{\mathscr{S}}_n(\mathscr{V})$ onto $\mathscr{S}_n(\mathscr{V})$. We call v_n the Λ_n-*transform* of $\tilde{v}_n$, and always use $\tilde{v}_n$ to denote the unique element in $\tilde{\mathscr{S}}_n(\mathscr{V})$ given by $\Lambda_n \tilde{v}_n = v_n$.

Consider the following linear spaces:

$$\bigoplus_{1}^{\infty} \tilde{\mathscr{S}}_n(\mathscr{V}_n), \qquad \bigoplus_{1}^{\infty} \mathscr{S}_n(\mathscr{V}_n). \tag{2.2}$$

Then Λ is the isomorphism from $\bigoplus\tilde{\mathscr{S}}_n(\mathscr{V}_n)$ onto $\bigoplus\mathscr{S}_n(\mathscr{V}_n)$ defined by

$$\Lambda(\bigoplus\tilde{v}_n) \doteq \bigoplus\Lambda_n\tilde{v}_n, \tag{2.3}$$

where $\bigoplus\tilde{v}_n \in \bigoplus\tilde{\mathscr{S}}_n(\mathscr{V}_n)$. Let $\tilde{v} = \bigoplus\tilde{v}_n$, $v = \bigoplus v_n$ and $v = \Lambda\tilde{v}$; then v is called the Λ-*transform* of $\tilde{v}$. A ${}^{\sim}$ is always used over the unique element in $\bigoplus\tilde{\mathscr{S}}_n(\mathscr{V}_n)[\tilde{\mathscr{S}}_n(\mathscr{V}_n)]$ given by $\Lambda\tilde{v} = v[\Lambda_n\tilde{v}_n = v_n]$ respectively.

We say that $v_n \in \mathscr{S}_n(\mathscr{V}_n)$ is a *polynomial* if

$$v_n(\lambda_1, \cdots, \lambda_n) = \sum_{i_1=0,\cdots,i_n=0}^{\text{finite}} \tilde{v}_n(i_1, \cdots, i_n)\lambda_1^{i_1} \cdots \lambda_n^{i_n}, \tag{2.4}$$

where $\tilde{v}_n \in \tilde{\mathscr{S}}(\mathscr{V})$. A $v_n \in \mathscr{S}_n(\mathscr{V}_n)$ is called *rational* if there exists a scalar valued polynomial $d \in \mathscr{S}_n(\mathscr{K})$, $d(0, 0, \cdots, 0) \neq 0$, and a polynomial $q \in \mathscr{S}_n(\mathscr{V}_n)$, such that

$$d(\lambda_1, \cdots, \lambda_n)v_n(\lambda_1, \cdots, \lambda_n) = q(\lambda_1, \cdots, \lambda_n). \tag{2.5}$$

When (2.5) holds, then v_n is written as $v_n = q/d$. The condition $d(0) \neq 0$ guarantees that $d(\lambda_1, \cdots, \lambda_n)$ contains only strictly positive terms in its formal series. (The usual definition of rationality does not require that $d(0) \neq 0$. But when dealing with causal systems we always have $d(0) \neq 0$. So this additional assumption has been incorporated in our definition.)

We say $v \in \bigoplus\mathscr{S}_n(\mathscr{V}_n)$ is a *generalized polynomial* if $v = \bigoplus v_n$ and $v_n = 0$ for all large n. If $v = \bigoplus v_n$ is a generalized polynomial, this does not imply that v_n is a rational function for all n.

To complete this section we define several operators needed throughout this paper. The *backward shift operator* S_n is the linear operator from $\mathscr{S}_n(\mathscr{V})$ to $\mathscr{S}_n(\mathscr{V})$ defined by

$$S_nv_n(\lambda_1, \lambda_2, \cdots, \lambda_n) \doteq \frac{1}{\lambda_1}[v_n(\lambda_1, \lambda_2, \cdots, \lambda_n) - v_n(0, \lambda_2, \lambda_3, \cdots, \lambda_n)], \tag{2.6}$$

where $v_n \in \mathscr{S}_n(\mathscr{V})(n > 0)$. S_n is called the backward shift operator because S_n shifts the λ_1-coefficient in the formal power series expansion of v_n. For example, if $v_2 \in \mathscr{S}_2(\mathscr{V})$ is given by (2.1), then

$$S_2^k v_2 = \sum_{i=k,j=0}^{\infty,\infty} \tilde{v}_2(i, j)\lambda_1^{i-k}\lambda_2^{j}, \qquad k \geqq 0. \tag{2.7}$$

The *evaluation operator* E_n is the linear operator from $\mathscr{S}_n(\mathscr{V})$ to $\mathscr{S}_{n-1}(\mathscr{V})$ defined by

$$E_nv_n(\lambda_1, \lambda_2, \cdots, \lambda_n) \doteq v_n(0, \lambda_1, \lambda_2, \cdots, \lambda_{n-1}), \tag{2.8}$$

where $v_n \in \mathscr{S}_n(\mathscr{V})(n > 0)$. For example, if $v_3 \in \mathscr{S}_3(\mathscr{V})$ is given by (2.1), then

$$E_3v_3(\lambda_1, \lambda_2, \lambda_3) = \sum_{i=0,j=0}^{\infty,\infty} \tilde{v}_3(0, i, j)\lambda_1^{i}\lambda_2^{j}. \tag{2.9}$$

Note that the domain of S_n and E_n is $\mathscr{S}_n(\mathscr{V})$. The same symbol S_n, E_n is used to denote all operators defined by (2.6), (2.8) respectively, regardless of the particular space $\mathscr{V}$ we are working with.

The *generalized backward shift operator* S is the linear operator from $\oplus\mathcal{S}_n(\mathcal{V}_n)$ to $\oplus\mathcal{S}_n(\mathcal{V}_n)$ defined by

$$S\bigoplus_1^\infty v_n \doteq \bigoplus_1^\infty S_n v_n = \{S_1v_1, S_2v_2, S_3v_3, \cdots\}, \tag{2.10}$$

where $\oplus v_n \in \oplus\mathcal{S}_n(\mathcal{V}_n)$. The *generalized evaluation operator* E is the linear operator from $\oplus\mathcal{S}_n(\mathcal{V}_n)$ to $\mathcal{V}_1$ defined by

$$E\oplus v_n \doteq E_1v_1 = \tilde{v}_1(0) = v_1(0), \tag{2.11}$$

where $\oplus v_n \in \oplus\mathcal{S}_n(\mathcal{V}_n)$. The same symbol S, E is used to denote all operators defined by (2.10), (2.11) respectively, regardless of the particular space $\oplus\mathcal{S}_n(\mathcal{V}_n)$, $\mathcal{V}_n$ they are defined on.

3. The transfer function. Throughout the rest of this paper the input space $\mathcal{U}$ and output space $\mathcal{Y}$ are fixed linear spaces, unless stated otherwise. In this section the Λ-transform is used to define our transfer function for system (1.1) and $\Sigma = \{A, B, C, N, \mathcal{X}\}$.

To begin we define the following linear spaces:

$$\begin{aligned}\tilde{\mathcal{H}}(\mathcal{U};\mathcal{Y}) &\doteq \bigoplus_1^\infty \tilde{\mathcal{S}}_n(\mathcal{L}(\mathcal{U}^n;\mathcal{Y})),\\ \mathcal{H}(\mathcal{U};\mathcal{Y}) &\doteq \bigoplus_1^\infty \mathcal{S}_n(\mathcal{L}(\mathcal{U}^n;\mathcal{Y})).\end{aligned} \tag{3.1}$$

Since SLT input–output maps play an important role in our theory, the following spaces are also defined:

$$\begin{aligned}\tilde{\mathcal{H}}_\Delta(\mathcal{U};\mathcal{Y}) &\doteq \{\oplus\tilde{\theta}_m \in \tilde{\mathcal{H}}(\mathcal{U};\mathcal{Y}) | \{\tilde{\theta}_m\} \text{ is SLT, i.e., (1.2) holds}\},\\ \mathcal{H}_\Delta(\mathcal{U};\mathcal{Y}) &\doteq \Lambda\tilde{\mathcal{H}}_\Delta(\mathcal{U};\mathcal{Y}).\end{aligned} \tag{3.2}$$

For notational convenience we drop the emphasis on $\mathcal{U}$, $\mathcal{Y}$ in the above spaces, and simply write $\tilde{\mathcal{H}}$, $\mathcal{H}$, $\tilde{\mathcal{H}}_\Delta$, $\mathcal{H}_\Delta$ to represent the spaces defined in (3.1), (3.2) respectively. Clearly $\tilde{\mathcal{H}}_\Delta[\mathcal{H}_\Delta]$ is a linear subspace of $\tilde{\mathcal{H}}[\mathcal{H}]$.

Consider any input–output map of the form (1.1) with kernel sequence $\{\tilde{\theta}_m\}$. Then $\tilde{\theta} \doteq \oplus\tilde{\theta}_m$ is an element of $\tilde{\mathcal{H}}$, and $\theta \doteq \Lambda\tilde{\theta}$ is well defined. We call θ the *transfer function* for system (1.1). Transfer functions are elements in $\mathcal{H}$. If system (1.1) is SLT, then its transfer function θ is an element in $\mathcal{H}_\Delta$. Since Λ is an isomorphism, there is a one-to-one correspondence between elements in the space $\mathcal{H}[\mathcal{H}_\Delta]$ and input–output maps of the form (1.1) [SLT input–output maps of the form (1.1)].

Let $\Sigma = \{A, B, C, N, \mathcal{X}\}$. The transfer function for Σ is denoted by θ_Σ. Since bilinear systems generate SLT input–output maps (see (1.7)), $\theta_\Sigma \in \mathcal{H}_\Delta$.

LEMMA 3.1. *The transfer function θ_Σ corresponding to $\Sigma = \{A, B, C, N, \mathcal{X}\}$ is given by $\theta_\Sigma = \oplus\theta_{n\Sigma} \in \mathcal{H}_\Delta$, where the first two terms are $\theta_{1\Sigma} = CF_1B$, $\theta_{2\Sigma}[u_1, u_2] = CF_2N(u_2)F_1Bu_1$.*

The general term is

$$\theta_{n\Sigma}[u_1, u_2, \cdots, u_n] = CF_nN(u_n)F_{n-1}N(u_{n-1})\cdots N(u_2)F_1Bu_1, \qquad n > 0, \tag{3.3}$$

where $u_k \in \mathcal{U}$ and F_m is the formal series defined by

$$F_m = \sum_{i=0}^\infty A^i\lambda_m^{i+1}. \tag{3.4}$$

Remark 3.2. The bilinear operator N in (3.3) appears $(n-1)$ times. Therefore each $\theta_{n\Sigma}$ is a formal power series in $\lambda_1, \lambda_2, \cdots, \lambda_n$ with values in $\mathscr{L}(\mathscr{U}^n; \mathscr{Y})$. It is also important to note the "backward" order of the u_i's in (3.3). On occasion the u_i's are not inserted in (3.3) and the equation is written as

$$(3.5) \qquad \theta_{n\Sigma} = CF_nNF_{n-1}N \cdots NF_1B.$$

Proof. The kernel sequence $\{\tilde{\theta}_{n\Sigma}\}$ for Σ is given by (1.7). Thus $\theta_\Sigma = \Lambda(\bigoplus\tilde{\theta}_{n\Sigma}) = \bigoplus\theta_{n\Sigma}$. From the definition of Λ, (2.3) with (1.7) and (2.1), we see that (3.3) holds. □

Since Λ is an isomorphism, the bilinear realization question is equivalent to the following: Given a $\theta \in \mathscr{H}_\Delta$, then find a system $\Sigma = \{A, B, C, N, \mathscr{X}\}$ such that $\theta = \theta_\Sigma$.

4. The backward shift realization. In this section we obtain a solution to the bilinear realization problem. The approach is based on the following property of shift operators:

$$(4.1) \qquad E_1S_1^{i_n}E_2S_2^{i_{n-1}} \cdots E_nS_n^{i_1}v_n = \tilde{v}_n(i_1, i_2, \cdots, i_n), \qquad v_n \in \mathscr{S}_n(\mathscr{V}),$$

(recall that $v_n = \Lambda_n\tilde{v}_n$). Consider any $\theta = \bigoplus\theta_n \in \mathscr{H}_\Delta$. From (4.1),

$$(4.2) \qquad E_1S_1^{i_n+1}E_2S_2^{i_{n-1}+1} \cdots E_nS_n^{i_1+1}\theta_n = \tilde{\theta}_n(i_1+1, i_2+1, \cdots, i_n+1),$$

for all $(i_1, i_2, \cdots, i_n) \in I^n$ and $n > 0$. By (1.7) and (4.2), $\Sigma = \{A, B, C, N, \mathscr{X}\}$ is a realization of θ if

$$(4.3) \qquad \begin{aligned} &E_1S_1^{i_n+1}E_2S_2^{i_{n-1}+1} \cdots E_nS_n^{i_1+1}\theta_n[u_1, u_2, \cdots, u_n] \\ &\quad = CA^{i_n}N(u_n)A^{i_{n-1}}N(u_{n-1}) \cdots N(u_2)A^{i_1}Bu_1, \end{aligned}$$

for all $(i_1, i_2, \cdots, i_n) \in I^n$, all $u_k \in \mathscr{U}$, and all $n > 0$. So we obtain a solution to the bilinear realization problem by finding a system $\Sigma = \{A, B, C, N, \mathscr{X}\}$, consisting of shift operators and evaluation operators such that (4.3) holds.

We introduce several operators that will aid us in doing this. For each $x_n \in \mathscr{S}_n(\mathscr{L}(\mathscr{U}^n; \mathscr{Y}))$, $n > 0$, let $x_n: \mathscr{U} \to \mathscr{S}_n(\mathscr{L}(\mathscr{U}^{n-1}; \mathscr{Y}))$ be the operator defined by

$$(4.4) \qquad x_nu \doteq x_n(\lambda_1, \cdots, \lambda_n)[u, \cdot, \cdot, \cdots, \cdot], \qquad u \in \mathscr{U},$$

where $x_n(\lambda_1, \cdots, \lambda_n)[u, \cdot, \cdot, \cdots, \cdot]$ is a $(n-1)$-linear operator in $\mathscr{U}$, and $\lambda_1, \cdots, \lambda_n$ are indeterminates; i.e., $x_nu \in \mathscr{S}_n(\mathscr{L}(\mathscr{U}^{n-1}; \mathscr{Y}))$. The same symbol x_n is used to represent both the linear operator given by (4.4) and the element x_n in $\mathscr{S}_n(\mathscr{L}(\mathscr{U}^n; \mathscr{Y}))$.

The state space we choose for Σ is $\mathscr{X} = \mathscr{F}$, where

$$(4.5) \qquad \mathscr{F} = \mathscr{F}(\mathscr{U}; \mathscr{Y}) \doteq \bigoplus_{n=1}^{\infty} \mathscr{S}_n(\mathscr{L}(\mathscr{U}^{n-1}; \mathscr{Y})).$$

For each $x \in \mathscr{H}(\mathscr{U}; \mathscr{Y})$ we define a linear operator $x: \mathscr{U} \to \mathscr{F}(\mathscr{U}; \mathscr{Y})$ by

$$(4.6) \qquad xu \doteq \bigoplus x_nu = \bigoplus x_n(\lambda_1, \cdots, \lambda_n)[u, \cdot, \cdot, \cdots, \cdot],$$

where $u \in \mathscr{U}$, $x = \bigoplus x_n \in \mathscr{H}$. The same symbol x is used to denote both the operator given in (4.6) and the element x in $\mathscr{H}$. Finally, we introduce the bilinear operator $T: \mathscr{U} \times \mathscr{F} \to \mathscr{F}$:

$$(4.7) \qquad \begin{aligned} T(u)\bigoplus x_n &\doteq S\bigoplus E_{n+1}x_{n+1}u \\ &\doteq S\{x_2(0, \lambda_1)[u], x_3(0, \lambda_1, \lambda_2)[u, \cdot], x_4(0, \lambda_1, \lambda_2, \lambda_3)[u, \cdot, \cdot], \cdots\}. \end{aligned}$$

A system $\Sigma = \{A, B, C, N, \mathscr{X}\}$ that satisfies (4.3) is given by $\mathscr{X} = \mathscr{F}$, $A = S$, $C = E$, $N = T$, and $B = S\theta$, where

$$\begin{aligned} Bu &\doteq S\theta u = S \bigoplus \theta_n u \\ &= \bigoplus S_n \theta_n[u, \cdot, \cdot, \cdots, \cdot], \qquad u \in \mathscr{U}, \quad \theta = \bigoplus \theta_n. \end{aligned} \tag{4.8}$$

To show that Σ is a realization of θ consider the case $n = 2$ with $u_1, u_2 \in \mathscr{U}$:

$$\begin{aligned} CA^jN(u_2)A^iBu_1 &= ES^jT(u_2)S^{i+1}\theta u_1 \\ &= ES^jT(u_2)\{S_1^{i+1}\theta_1 u_1, S_2^{i+1}\theta_2 u_1, S_3^{i+1}\theta_3 u_1, \cdots\} \\ &= ES^jS\{E_2 S_2^{i+1}\theta_2[u_1, u_2], E_3 S_3^{i+1}\theta_3 \\ &\qquad \times [u_1, u_2, \cdot], E_4 S_4^{i+1}\theta_4[u_1, u_2, \cdot, \cdot], \cdots\} \\ &= E_1 S_1^{j+1} E_2 S_2^{i+1}\theta_2[u_1, u_2]; \end{aligned} \tag{4.9}$$

i.e., (4.3) holds. A calculation similar to (4.9) verifies that (4.3) holds for all $n > 0$. Thus

PROPOSITION 4.1. *If $\theta \in \mathscr{H}_\Delta$, then $\Gamma \doteq \{S, S\theta, E, T, \mathscr{F}\}$ is a realization of θ.*

Γ is called the *backward shift realization* (BSR) of θ. We need

DEFINITION 4.2

(i) If $\Sigma = \{A, B, C, N, \mathscr{X}\}$ is any bilinear system, then $\mathscr{X}_\Sigma$ is the linear space defined by:

$$\begin{aligned} \mathscr{X}_\Sigma \doteq \bigvee_{m \geqq 1} \bigvee\{A^{i_m}N(u_{i_m})A^{i_{m-1}}N(u_{i_{m-1}}) \cdots N(u_{i_2})A^{i_1}Bu_{i_1} \\ |(i_1, \cdots, i_m) \in I^m, u_i \in \mathscr{U} \text{ for } i \in I\}. \end{aligned} \tag{4.10}$$

(ii) If Γ is the BSR of θ, then $\mathscr{W}_\theta$ is the linear subspace of $\mathscr{F}$ defined by $\mathscr{W}_\theta \doteq (\mathscr{F}(\mathscr{U}; \mathscr{Y}))_\Gamma$.

The *restricted backward shift realization* (RBSR) of θ is obtained by restricting $S, S\theta, E, T$ to the space $\mathscr{W}_\theta$. More precisely, the RBSR of θ is defined by $\{S_\theta, S\theta, E_\theta, T_\theta, \mathscr{W}_\theta\}$, where the operators are

(i) $S_\theta: \mathscr{W}_\theta \to \mathscr{W}_\theta$, $S_\theta \doteq S|\mathscr{W}_\theta$,

(ii) $S\theta: \mathscr{U} \to \mathscr{W}_\theta$, $(S\theta)u = S\theta u$ if $u \in \mathscr{U}$,

(iii) $E_\theta: \mathscr{W}_\theta \to \mathscr{Y}$, $E_\theta = E|\mathscr{W}_\theta$,

(iv) T_θ: $\mathscr{U} \times \mathscr{W}_\theta \to \mathscr{W}_\theta$, $T_\theta(u)x = T(u)x$ if $x \in \mathscr{W}_\theta$ and $u \in \mathscr{U}$.

The same symbol $S\theta$ is used to denote both operators, $S\theta$ mapping $\mathscr{U}$ into $\mathscr{F}$ and $S\theta$ mapping $\mathscr{U}$ into $\mathscr{W}_\theta$.

Clearly the RBSR satisfies (4.3). Thus

PROPOSITION 4.3. *If $\theta \in \mathscr{H}_\Delta$, then the* RBSR $\{S_\theta, S\theta, E_\theta, T_\theta, \mathscr{W}_\theta\}$ *is a realization of θ.*

The RBSR is of lower dimension than the BSR. In the next section it is shown that the RBSR is a minimal realization of θ.

5. Minimal realizations. In this section we use the RBSR to develop a theory of minimality for bilinear systems. A linear operator H_Σ will play an important role in this theory. For each bilinear system $\Sigma = \{A, B, C, N, \mathscr{X}\}$ we define the operator H_Σ mapping $\mathscr{X}$ into $\mathscr{F}$ by

$$H_\Sigma x = S \bigoplus_1^\infty \Phi_{n\Sigma} x, \qquad x \in \mathscr{X}, \tag{5.1}$$

where F_M is given in (3.4) and $\Phi_{1\Sigma} x \doteq CF_1 x$, $\Phi_{2\Sigma}[u_2]x = CF_2 N(u_2)F_1 x, \cdots,$

$$\Phi_n[u_2, \cdots, u_n]x \doteq CF_n N(u_n)F_{n-1}N(u_{n-1}) \cdots N(u_2)F_1 x, \qquad n > 0, \quad u_k \in \mathscr{U}. \tag{5.2}$$

Clearly $\Phi_{n\Sigma} x \in \mathscr{S}_n(\mathscr{S}(\mathscr{U}^{n-1}; \mathscr{Y}))$ for all $n > 0$ and $x \in \mathscr{X}$.

The following is needed:

LEMMA 5.1. *If Σ is a bilinear realization of θ then H_Σ sends Σ into the* BSR *of θ.*

Proof. Let $\Sigma = \{A, B, C, N, \mathscr{X}\}$. First we make some observations concerning F_1 given by (3.4). Let $\hat{S}_1, \hat{E}_1$ be respectively the backward shift operator and evaluation operator on $\mathscr{S}_1(\mathscr{X})$; i.e., $\hat{S}_1 v(\lambda_1) = (v(\lambda_1) - v(0))/\lambda_1$, and $E_1 v(\lambda_1) = v(0)$ if $v \in \mathscr{S}_1(\mathscr{X})$. Reference to (3.4) shows that the following results are valid:

$$\hat{S}_1^2 F_1 = \hat{S}_1 F_1 A, \qquad \hat{E}_1 \hat{S}_1 F_1 = I_{\mathscr{X}}. \tag{5.3}$$

We now show that the required identities are valid. Because of (3.3) and (5.2), $\theta_{n\Sigma} = \Phi_{n\Sigma} B$ for all $n > 0$. Thus from (5.1) and $\theta_\Sigma = \theta$,

$$H_\Sigma B = S\theta_\Sigma = S\theta. \tag{5.4}$$

By the properties of E, S, and (5.3),

$$\begin{aligned} EH_\Sigma &= E \bigoplus S_n \Phi_{n\Sigma} = E_1 S_1 \Phi_{1\Sigma} \\ &= E_1 C \hat{S}_1 F_1 = C \hat{E}_1 \hat{S}_1 F_1 = C. \end{aligned} \tag{5.5}$$

From (5.3)

$$\begin{aligned} SH_\Sigma &= \bigoplus S_n^2 \Phi_{n\Sigma} = \bigoplus CF_n NF_{n-1} N \cdots N\hat{S}_1^2 F_1 \\ &= \bigoplus CF_n N \cdots N \hat{S}_1 F_1 A = \bigoplus S_n \Phi_{n\Sigma} A = H_\Sigma A. \end{aligned} \tag{5.6}$$

Finally, taking $u \in \mathscr{U}$ and using (4.7), (5.3)

$$\begin{aligned} T(u)H_\Sigma x &= T(u) \bigoplus S_n \Phi_{n\Sigma} \\ &= S \bigoplus (E_{n+1} S_{n+1} \Phi_{n+1\Sigma} u) \\ &= S \bigoplus (E_{n+1} CF_{n+1} N \cdots NF_2 N(u) \hat{S}_1 F_1) \\ &= S \bigoplus (CF_n N \cdots NF_1 N(u) \hat{E}_1 \hat{S}_1 F_1) \\ &= S \bigoplus \Phi_{n\Sigma} N(u) = H_\Sigma N(u). \end{aligned} \tag{5.7}$$

Equations (5.4)–(5.7) imply that H_Σ sends Σ into the BSR of θ. □

LEMMA 5.2. *If $\Sigma = \{A, B, C, N, \mathscr{X}\}$ is a realization of θ then $\mathscr{W}_\theta = H_\Sigma \mathscr{X}_\Sigma$. In particular, $\mathscr{W}_\theta \subseteq \mathscr{R}(H_\Sigma)$.*

Proof. Since $S\theta = H_\Sigma B$ we have

$$S^{i_n} T S^{i_{n-1}} T \cdots T S^{i_1} S\theta = S^{i_n} T S^{i_{n-1}} T \cdots T S^{i_1} H_\Sigma B. \tag{5.8}$$

From the definition of "into," H_Σ can be moved to the left-hand side of (5.8) to give

$$S^{i_n} T S^{i_{n-1}} T \cdots T S^{i_1} S\theta = H_\Sigma A^{i_n} N A^{i_{n-1}} N \cdots N A^{i_1} B, \tag{5.9}$$

for all $(i_1, \cdots, i_n) \in I^n$, $n > 0$. By Definition 4.2 this implies $\mathscr{W}_\theta = H_\Sigma \mathscr{X}_\Sigma$. □

Let $\Sigma = \{A, B, C, N, \mathscr{X}\}$ be a realization of θ. By Lemma 5.2, $\mathscr{W}_\theta \subseteq \mathscr{R}(H_\Sigma)$. From [30, Thm. (4.7.7)], $\dim(\mathscr{W}_\theta) \leqq \dim(\mathscr{R}(H_\Sigma)) \leqq \dim(\mathscr{X})$. Thus we have

PROPOSITION 5.3. *The* RBSR *of θ is a minimal realization of θ.*

COROLLARY 5.4. *$\theta \in \mathscr{H}_\Delta$ admits a finite dimensional bilinear realization if and only if* $\dim(\mathscr{W}_\theta) < \infty$.

Let $\Sigma = \{A, B, C, N, \mathscr{X}\}$ be a realization of θ. If $\mathscr{R}(H_\Sigma) = \mathscr{W}_\theta$, then we define the operator J_Σ mapping $\mathscr{X}$ onto $\mathscr{W}_\theta$ by: $J_\Sigma x = H_\Sigma x$ where $x \in \mathscr{X}$. In particular, J_Σ is defined when $\mathscr{X} = \mathscr{X}_\Sigma$ (see Lemma 5.2). Following the proof of Lemma 5.1, J_Σ sends Σ into the RBSR of θ. Combining these observations we have

LEMMA 5.5. *Let $\Sigma = \{A, B, C, N, \mathscr{X}\}$ be a realization of θ. If $\mathscr{X} = \mathscr{X}_\Sigma$ or $\mathscr{R}(H_\Sigma) = \mathscr{W}_\theta$ then J_Σ sends Σ into the* RBSR *of θ and J_Σ is onto $\mathscr{W}_\theta$.*

One of the main results of this section is

PROPOSITION 5.6. *Let $\theta \in \mathscr{H}_\Delta$ admit a finite dimensional bilinear realization. Then any minimal realization of θ is equivalent to the* RBSR of θ.

Proof. Let $\Sigma = \{A, B, C, N, \mathscr{X}\}$ be a minimal realization of θ. By Proposition 5.3, dim $(\mathscr{W}_\theta) = \dim(\mathscr{X}) < \infty$. From Lemma 5.2, $\mathscr{W}_\theta \subseteq \mathscr{R}(H_\Sigma)$. Thus $\mathscr{W}_\theta = \mathscr{R}(H_\Sigma)$. Applying Lemma (5.5) with dim $(\mathscr{W}_\theta) = \dim(\mathscr{X}) < \infty$ we see that J_Σ is an isomorphism that sends Σ into the RBSR. Hence Σ is equivalent to the RBSR. □

COROLLARY 5.7. *If θ admits a finite dimensional bilinear realization then any two minimal realizations of θ are equivalent.*

Proof. This follows from the fact that all minimal realizations of θ are equivalent to the RBSR and system equivalence is transitive. □

It is interesting to note that our theory of minimality for bilinear systems follows directly from the RBSR and does not make explicit use of the concepts of reachability and observability. In the following section we use the RBSR to develop a theory of reachability and observability without using the concept of minimality.

6. Span reachable and observable bilinear realizations. The purpose of this section is to show how the RBSR and H_Σ can be used, to offer simple proofs, to many of the standard results on reachability and observability for bilinear systems [7], [11], [26]. The results presented here hold for finite and infinite dimensional linear spaces. Thus our results are slightly more general than those existing in the literature.

To begin we establish some standard terminology.

DEFINITION 6.1. Let $\Sigma = \{A, B, C, N, \mathscr{X}\}$.

(i) Σ is *span reachable* if $\mathscr{H} = \mathscr{H}_\Sigma$.

(ii) Σ is *observable* if for all $x \in \mathscr{H}$ such that the $(n-1)$-linear operator $CA^{i_n}NA^{i_{n-1}}N \cdots NA^{i_1}x = 0$ for all $(i_1, \cdots, i_n) \in I^n$ and all $n > 0$, then $x = 0$.

Clearly, $\mathscr{X}_\Sigma$ is the linear span of the reachable set for Σ. If $N = 0$ then Σ is a linear system and these definitions reduce to the usual definitions of reachability and observability for linear systems [8]. The RBSR is span reachable. To show that the RBSR is observable, we need

LEMMA 6.2. *Let $\theta = \bigoplus \theta_n \in \mathscr{H}_\Delta$, and let $\mathscr{F}_1$ be the linear subspace of $\mathscr{F}$ defined by*

$$\mathscr{F}_1 \doteq \bigvee \left\{ \bigoplus_1^\infty (\lambda_2\lambda_3 \cdots \lambda_n\phi_n) | \phi_n \in \mathscr{S}_n(\mathscr{L}(\mathscr{U}^{n-1}; \mathscr{Y})), n > 0 \right\}. \tag{6.1}$$

Then $\mathscr{W}_\theta \subseteq \mathscr{F}_1$. In particular any $x \in \mathscr{W}_\theta$ can be put in the following form:

$$x = \bigoplus_1^\infty (\lambda_2\lambda_3 \cdots \lambda_n\phi_n) = \phi_1 \oplus \lambda_2\phi_2 \oplus \lambda_2\lambda_3\phi_3 \oplus \cdots, \tag{6.2}$$

where $\phi_n \in \mathscr{S}_n(\mathscr{L}(\mathscr{U}^{n-1}; \mathscr{Y}))$, $n > 0$.

Proof. First observe that $\mathscr{F}_1$ is an invariant subspace for S and $T(u)$, $u \in \mathscr{U}$. This is shown in the following calculation:

$$\begin{aligned} S \oplus (\lambda_2\lambda_3 \cdots \lambda_n\phi_n) &= \oplus(\lambda_2\lambda_3 \cdots \lambda_n S_n\phi_n) \in \mathscr{F}_1, \\ T(u) \oplus \lambda_2\lambda_3 \cdots \lambda_n\phi_n &= S \oplus \lambda_1\lambda_2 \cdots \lambda_n E_{n+1}\phi_{n+1}u \\ &= \oplus\lambda_2\lambda_3 \cdots \lambda_n E_{n+1}\phi_{n+1}u \in \mathscr{F}_1. \end{aligned} \tag{6.3}$$

The sequence $\{\tilde{\theta}_n\}$ is SLT. From (1.2) with the definition of Λ_n there exists a $\Phi_n \in \mathscr{S}_n(\mathscr{L}(\mathscr{U}^n; \mathscr{Y}))$ such that $\theta_n = (\lambda_1\lambda_2 \cdots \lambda_n)\Phi_n$ for all $n > 0$. Thus θ can be put in the form, $\theta = \bigoplus \lambda_1\lambda_2 \cdots \lambda_n \Phi_n$. Applying S with $u \in \mathscr{U}$, we obtain

$$S\theta u = \bigoplus_{1}^{\infty} (\lambda_2\lambda_3 \cdots \lambda_n \Phi_n u). \tag{6.4}$$

Thus $S\theta \subseteq \mathscr{F}_1$. Since $\mathscr{F}_1$ is an invariant subspace for S and $T(u), u \in \mathscr{U}$ we have $\mathscr{W}_\theta \subseteq \mathscr{F}_1$, and the proof is complete. □

PROPOSITION 6.3. *If $\theta \in \mathscr{H}_\Delta$ then the* RBSR *of θ is span reachable and observable.*

Proof. The RBSR is span reachable, so we show the RBSR is observable. Consider any $x \in \mathscr{W}_\theta$. From (6.2), $x = \bigoplus(\lambda_2\lambda_3 \cdots \lambda_n \phi_n)$. By applying an argument similar to that given in (4.9) with (4.1), we obtain

$$\begin{aligned} E_\theta S_\theta^j T_\theta S_\theta^i x &= ES^j T \bigoplus (\lambda_2\lambda_3 \cdots \lambda_n S_n^i \phi_n) \\ &= ES^j S \bigoplus (\lambda_1\lambda_2 \cdots \lambda_n E_{n+1} S_{n+1}^i \phi_{n+1}) \\ &= E \bigoplus (\lambda_2 \cdots \lambda_n S_n^j E_{n+1} S_{n+1}^i \phi_{n+1}) \\ &= E_1 S_1^j E_2 S_2^i \phi_2 = \tilde{\phi}_2(i, j). \end{aligned} \tag{6.5}$$

In the general case the calculations in (6.5) give

$$E_\theta S_\theta^{i_n} T_\theta S_\theta^{i_{n-1}} T_\theta \cdots T_\theta S_\theta^{i_1} x = E_1 S_1^{i_n} E_2 S_2^{i_{n-1}} \cdots E_n S_n^{i_1} \phi_n = \tilde{\phi}_n(i_1, \cdots, i_n). \tag{6.6}$$

Therefore (6.6) is zero for all $(i_1, \cdots, i_n) \in I^n$ and $n > 0$, if and only if $\tilde{\phi}_n = 0$ for all $n > 0$. From (6.2) the RBSR is observable. □

The operator J_Σ is the key to showing that any two span reachable and observable realizations of θ are equivalent. To this end we give

PROPOSITION 6.4. *If Σ is a span reachable and observable bilinear realization of θ, then Σ is equivalent to the* RBSR *of θ.*

Proof. Let $\Sigma = \{A, B, C, N, \mathscr{X}\}$ be a span reachable and observable realization of θ. From Lemma 5.5 with $\mathscr{X} = \mathscr{X}_\Sigma$, the operator J_Σ sends Σ into the RBSR of θ. Clearly J_Σ is onto. To complete the proof we show that J_Σ is one-to-one.

Consider any $x \in \mathscr{X}$ such that $J_\Sigma x = 0$. Then

$$E_\theta S_\theta^{i_n} T_\theta S_\theta^{i_{n-1}} T_\theta \cdots T_\theta S_\theta^{i_1} J_\Sigma x = 0, \tag{6.7}$$

for all $(i_1, \cdots, i_n) \in I^n$ and $n > 0$. From the definition of "sends into," the J_Σ moves to the left-hand side of (6.7); i.e.,

$$E_\theta J_\Sigma A^{i_n} N A^{i_{n-1}} N \cdots N A^{i_1} x = 0, \tag{6.8}$$

for all $(i_1, \cdots, i_n) \in I^n$ and $n > 0$. Since $E_\theta J_\Sigma = C$, (6.8) becomes: $CA^{i_n} NA^{i_{n-1}} N \cdots NA^{i_1} x = 0$ for all $(i_1, \cdots, i_n) \in I^n$ and $n > 0$. System Σ is observable, so $x = 0$ and J_Σ is one to one. □

From Proposition 6.4 and the transitivity of system equivalence the following is evident.

COROLLARY 6.5. *Any two span reachable and observable bilinear realizations of θ are equivalent.*

COROLLARY 6.6. *If Σ is a span reachable and observable bilinear realization of θ then Σ is a minimal bilinear realization of θ.*

The converse to Corollary 6.6 is not true. It is easy to construct infinite dimensional bilinear systems that are minimal and not span reachable and observable. In the finite dimensional setting this does not happen; i.e., Proposition 6.6. *Let $\theta \in \mathcal{H}_\Delta$ admit a finite dimensional bilinear realization. Σ is a minimal realization of θ if and only if Σ is a span reachable and observable realization of θ.*

Proof. Any minimal realization of θ is equivalent to the RBSR (Proposition 5.6). Any span reachable and observable realization of θ is equivalent to the RBSR (Proposition 6.4). Since system equivalence preserves minimality, span reachability, and observability, and the RBSR is minimal, span reachable, and observable, the proof is complete. □

7. Finite dimensional bilinear systems. In this section we use the RBSR to develop conditions for the existence of a finite dimensional bilinear realization. This section is broken up into two parts, A and B. In part A, all transfer functions for finite dimensional systems are shown to satisfy a special condition (called regularity). An example of how one uses this regularity condition to compute the RBSR is given. In part B the regularity condition on θ is used to develop a special rational form for θ (called factorable). This factorable form leads to necessary and sufficient conditions for θ to admit a finite dimensional bilinear realization.

A. The regularity conditions. To begin, we define the following linear subspaces of $\mathcal{W}_\theta(\theta \in \mathcal{H}_\Delta)$:

$$\mathcal{W}_\theta^1(k) \doteq \bigvee_{i=0}^{k} S^i S\theta\mathcal{U}, \qquad \mathcal{W}_\theta^1 \doteq \bigvee_{i=0}^{\infty} \mathcal{W}_\theta^1(i),$$

$$\mathcal{W}_\theta^q(k) \doteq \left[\bigvee_{i=0}^{k} \{S^i T(u)w \mid w \in \mathcal{W}_\theta^{q-1}, u \in \mathcal{U}\}\right] \vee \mathcal{W}_\theta^{q-1} \quad \text{if } q > 1, \tag{7.1}$$

$$\mathcal{W}_\theta^q \doteq \bigvee_{i=0}^{\infty} \mathcal{W}_\theta^q(i), \qquad \text{if } q \geqq 1.$$

From these definitions we obviously have:

$$\mathcal{W}_\theta^q(k) \subseteq \mathcal{W}_\theta^q(k+1) \subseteq \mathcal{W}_\theta^q, \qquad q \geqq 1, k \geqq 0, \tag{7.2}$$

$$\mathcal{W}_\theta^q \subseteq \mathcal{W}_\theta^{q+k} \subseteq \mathcal{W}_\theta, \qquad q \geqq 1, \tag{7.3}$$

$$\mathcal{W}_\theta = \bigvee_{q=1}^{\infty} \mathcal{W}_\theta^q, \tag{7.4}$$

$$\mathcal{W}_\theta^q = \bigvee_{m=1}^{q} \{S^{i_m} T(u_{i_m}) \cdots T(u_{i_2}) S^{i_1} S\theta u_{i_1} \mid (i_1, \cdots, i_m) \in I^m; u_i \in \mathcal{U} \text{ for } i \in I\}. \tag{7.5}$$

Let $\theta \in \mathcal{H}_\Delta$ admit a finite dimensional bilinear realization. By Corollary 5.4, dim $(\mathcal{W}_\theta) < \infty$. This implies that the dimensions of the spaces $\mathcal{W}_\theta^q(k)$, $\mathcal{W}_\theta^q$ are bounded for all integers k, q. This suggests the following.

Definition 7.1. Let $\theta \in \mathcal{H}_\Delta$.

(i) θ is *S-regular* if for all $q > 0$ there exists a $i_q \geqq 0$ such that $\mathcal{W}_\theta^q = \mathcal{W}_\theta^q(i_q)$.

(ii) θ is *T-regular* if there exists a finite integer $Q > 0$ such that $\mathcal{W}_\theta^Q = \mathcal{W}_\theta$.

(iii) θ is *regular* if θ is S-regular and T-regular.

Remark 7.2. If $\theta = \bigoplus \theta_n$ is a generalized polynomial, then θ is T-regular. In fact, if $\theta_n = 0$ for all $n \geqq Q$ then $\mathcal{W}_\theta = \mathcal{W}_\theta^Q$.

Remark 7.3. Let $\theta \in \mathcal{H}_\Delta$ be regular. Combining parts (i), (ii) of Definition 7.1 shows that there exists a finite integer $Q > 0$ and an upper bound m, such that $\mathcal{W}_\theta^k(m) = \mathcal{W}_\theta^k$ for $1 \leqq k \leqq Q$, and $\mathcal{W}_\theta^Q = \mathcal{W}_\theta^Q(m) = \mathcal{W}_\theta$. (Note that (7.2) implies $\mathcal{W}_\theta^k(i_k + j) = \mathcal{W}_\theta^k$ whenever

$\mathcal{W}^k_\theta(i_k)=\mathcal{W}^k_\theta$, so such an m can be chosen for finite Q.) From (7.5) with these bounds and m, Q, and θ regular, we have

$$\mathcal{W}_\theta=\bigvee_{n=1}^{Q}\{S^{i_n}T(u_{i_n})S^{i_{n-1}}T(u_{i_{n-1}})\cdots T(u_{i_2})S^{i_1}S\theta u_{i_1} \\ |0\leqq i_j\leqq m, j=1,\cdots,n \text{ and } u_i\in\mathcal{U} \text{ for } i\in I\}. \tag{7.6}$$

Finite dimensional systems have regular transfer functions. If not, then dim $\mathcal{W}^q_\theta(k)$) would approach infinity as k and q approach infinity, (7.1)–(7.4). However, not all regular θ's admit a finite dimensional bilinear realization. For example, $\theta=\{\lambda_1 I_{\mathcal{U}}, 0, 0,\cdots\}$ with $\dim(\mathcal{U})=\infty$ is regular and $\dim(\mathcal{W}_\theta)=\infty$. By Corollary 5.4, this θ does not admit a finite dimensional bilinear realization. The connection between regularity and finite dimensional bilinear systems is given in

PROPOSITION 7.4. *$\theta\in\mathcal{H}_\Delta$ admits a finite dimensional bilinear realization if and only if θ is regular and*

$$\dim(\mathcal{W}^q_\theta(i))\leqq M_{i,q}<\infty \quad \text{for all } i\geqq 0, q>0. \tag{7.7}$$

Proof. Let θ admit a finite dimensional bilinear realization. As shown above θ is regular. Condition (7.7) follows from Corollary 5.4 and $\mathcal{W}^q_\theta(i)\subseteq\mathcal{W}_\theta$ for all i, q. The converse follows from Corollary 5.4 and Remark 7.3; i.e., $\mathcal{W}_\theta=\mathcal{W}^Q_\theta(m)$ implies $\dim(\mathcal{W}_\theta)\leqq M_{m,Q}<\infty$. □

COROLLARY 7.5. *Let $\theta\in H_\Delta(\mathcal{U};\mathcal{Y})$ with* $\dim(\mathcal{U})<\infty$. *θ admits a finite dimensional bilinear realization if and only if θ is regular.*

Proof. If θ is regular, then Remark 7.3 with (7.6) and $\dim(\mathcal{U})<\infty$ implies $\dim(\mathcal{W}_\theta)<\infty$. The converse is obvious. □

COROLLARY 7.6. *Let $\theta\in\mathcal{H}(\mathcal{U};\mathcal{Y})$ be a generalized polynomial with* $\dim(\mathcal{U})<\infty$. *θ admits a finite dimensional bilinear realization if and only if θ is S-regular.*

Proof. Let θ be S-regular with $\dim(\mathcal{U})<\infty$. From Remark 7.2 θ is T-regular. Thus Corollary 7.5 implies that θ admits a finite dimensional bilinear realization. The converse is obvious. □

The hypothesis $\dim(\mathcal{U})<\infty$ in Corollary 7.5 cannot be replaced by $\dim(\mathcal{Y})<\infty$. For instance, consider the following: $\theta=\bigoplus\theta_n$,

$$\theta_1(\lambda_1)=(\lambda_1,\lambda_1^2,\lambda_1^3,\lambda_1^4,\cdots), \qquad \theta_n=0 \quad \text{if } n>1, \tag{7.8}$$

with $\mathcal{H}=\mathcal{R}$, the real numbers and $\mathcal{U}=\mathcal{R}^\infty$, the linear space of infinite tuples with compact support. By choosing $u_i\in\mathcal{U}$ to be the ith unit vector (the ith column is 1 and all other columns are zero), it is easy to show that $\dim(\mathcal{W}_\theta)=\infty$. Clearly $S^n\theta\mathcal{U}=S\theta\mathcal{U}$ for all $n>0$. Thus θ is regular, $\dim(\mathcal{Y})=1$ and $\dim(\mathcal{W}_\theta)=\infty$.

The following can be used to determine whether or not θ is regular.

LEMMA 7.7. *Let $\theta\in\mathcal{H}_\Delta$.*

(i) *$\mathcal{W}^q_\theta=\mathcal{W}^q_\theta(i_q)$ if and only if*

$$\mathcal{W}^q_\theta(i_q+1)=\mathcal{W}^q_\theta(i_q). \tag{7.9}$$

(ii) *$\mathcal{W}_\theta=\mathcal{W}^Q_\theta$ if and only if*

$$\{T(u)w|w\in\mathcal{W}^Q_\theta, \text{ and } u\in\mathcal{U}\}\subseteq\mathcal{W}^Q_\theta. \tag{7.10}$$

Proof.

(i) If (7.9) holds, then $\mathcal{W}^q_\theta(i_q)$ is an invariant subspace for S. By (7.1) $\mathcal{W}^q_\theta=\mathcal{W}^q_\theta(i_q)$. The converse is obvious.

(ii) If (7.10) holds, then $\mathscr{W}_\theta^Q$ is an invariant subspace for all $T(u)$, $u \in \mathscr{U}$ and S. By (7.1) to (7.4) $\mathscr{W}_\theta^Q = \mathscr{W}_\theta$. The converse is obvious. □

Remark 7.8. Let $\theta \in \mathscr{H}_\Delta$. From Definition 7.1 and Lemma 7.7 the following is evident.

(i) θ is S-regular if and only if for all $q > 0$ there exists an $i_q \geqq 0$ such that (7.9) holds.

(ii) θ is T-regular if and only if there exists a (finite integer) $Q > 0$ such that (7.10) holds.

(iii) Finally, θ is regular if and only if there exist $Q > 0$, $i_q \geqq 0$ such that $\mathscr{W}_\theta^q(i_q) = \mathscr{W}_\theta^q(i_q + 1)$ for $q = 1, \cdots, Q$ and (7.10) holds. □

If θ is regular, then (7.9) and (7.10) give us a recursive procedure for calculating the space $\mathscr{W}_\theta$. First, set $q = 1$ and iterate on i_q until (7.9) becomes valid. Once (7.9) holds, set $\mathscr{W}_\theta^q = \mathscr{W}_\theta^q(i_q)$, $q = q + 1$, and repeat the above procedure until (7.10) holds. Once (7.10) becomes valid, $\mathscr{W}_\theta = \mathscr{W}_\theta^q$ (with $q = Q$). If (7.9) or (7.10) never become valid, then θ is not regular and a finite dimensional bilinear realization of θ does not exist. Upon obtaining $\mathscr{W}_\theta$ (i.e., a valid (7.9) and (7.10)), it is a simple matter to find a basis for $\mathscr{W}_\theta$ and a matrix representation for the RBSR. In fact we obtain a basis for $\mathscr{W}_\theta$ as we recursively compute each $\mathscr{W}_\theta^q(i_q)$. To demonstrate this procedure we give

Example 7.9. Let $\theta \in \mathscr{H}_\Delta(\mathscr{R}, \mathscr{R})$ be given by

$$\theta = \left\{\frac{\lambda_1}{(1-\lambda_1)}, \frac{\lambda_1\lambda_2^2}{(1-2\lambda_1)(1-3\lambda_2)}, 0, 0, \cdots\right\}, \tag{7.11}$$

where $\mathscr{R}$ is the field of real numbers. (Note that $(1-\alpha\lambda)^{-1}$ is the formal series given by

$$(1-\alpha\lambda)^{-1} \doteq \sum_{i=0}^{\infty} (\alpha\lambda)^i, \tag{7.12}$$

where $\alpha \in \mathscr{R}$ and λ is an indeterminate.) For this θ we find a basis for $\mathscr{W}_\theta$ and a matrix representation for the RBSR of θ.

First, some notation is established. $\mathscr{K}^n$ denotes the linear space of n-column vectors, and e_i is the ith unit vector; i.e., the ith column of e_i is 1 and all other columns of e_i are zero. Since $\mathscr{U} = \mathscr{R} = \mathscr{K}$, we identify $S\theta$: $\mathscr{U} \to \mathscr{F}$ with the element $S\theta \in \mathscr{F}$ and the bilinear operator T: $\mathscr{U} \times \mathscr{F} \to \mathscr{F}$ with the appropriate linear operator T: $\mathscr{F} \to \mathscr{F}$ (because $u = u \cdot 1$ we have $T(u)x = uT(1)x \doteq uTx$ when $u \in \mathscr{U} = \mathscr{K}$ and $x \in \mathscr{F}$.)

Using $S_n(1-\alpha\lambda_1)^{-1} = \alpha(1-\alpha\lambda_1)^{-1}$ where $\alpha \in \mathscr{R}$, $n > 0$ (see (2.6)) we have:

$$S\theta = \left\{\frac{1}{(1-\lambda_1)}, \frac{\lambda_2^2}{(1-2\lambda_1)(1-3\lambda_2)}, 0, 0, \cdots\right\} \doteq w_1, \tag{7.13}$$

$$SS\theta = Sw_1 = \left\{\frac{1}{(1-\lambda_1)}, \frac{2\lambda_2^2}{(1-2\lambda_1)(1-3\lambda_2)}, 0, 0, \cdots\right\} \doteq w_2, \tag{7.14}$$

$$S^2S\theta = Sw_2 = \left\{\frac{1}{1-\lambda_1}, \frac{4\lambda_2^2}{(1-2\lambda_1)(1-3\lambda_2)}, 0, 0, \cdots\right\} = -2w_1 + 3w_2. \tag{7.15}$$

Thus (7.9) holds and w_1, w_2 span $\mathscr{W}_\theta^1$. Using the definition of S and T we have:

$$Tw_1 = \left\{\frac{\lambda_1}{(1-3\lambda_1)}, 0, 0, \cdots\right\} \doteq w_3, \tag{7.16}$$

$$Tw_2 = \left\{\frac{2\lambda_1}{(1-3\lambda_1)}, 0, 0, \cdots\right\} = 2w_3, \tag{7.17}$$

$$STw_1 = Sw_3 = \left\{\frac{1}{(1-3\lambda_1)}, 0, 0, \cdots\right\} \doteq w_4, \tag{7.18}$$

$$STw_2 = 2w_4, \qquad S^2Tw_2 = 6w_4, \tag{7.19}$$

$$S^2Tw_1 = Sw_4 = 3w_4. \tag{7.20}$$

Thus (7.9) holds and $\{w_i\}_{i=1}^4$ span $\mathscr{W}_\theta^2$. Clearly $Tw_1 = w_3$, $Tw_2 = 2w_3$, $Tw_3 = Tw_4 = 0$. So (7.10) is valid $(Q = 2)$, and $\{w_i\}_{i=1}^4$ spans $\mathscr{W}_\theta = \mathscr{W}_\theta^2$.

Let $\{A, B, C, N, \mathscr{R}^4\}$ be a matrix representation of the RBSR $\{S_\theta, S\theta, E_\theta, T_\theta, \mathscr{W}_\theta\}$ where w_i is represented in $\mathscr{R}^4$ by e_i for $i = 1, \cdots, 4$. From this representation with (7.14), (7.15), (7.18), (7.20), we have: $Ae_1 = e_2$, $Ae_2 = -2e_1 + 3e_2$, $Ae_3 = e_4$, $Ae_4 = 3e_4$, or

$$A = \begin{bmatrix} 0 & -2 & 0 & 0 \\ 1 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 3 \end{bmatrix}. \tag{7.21}$$

From (7.13) with B representing $S\theta$ we have $B = e_1$; i.e.,

$$B = [1 \quad 0 \quad 0 \quad 0]^t, \tag{7.22}$$

where t denotes transpose. Since C represents E, $C = [Ew_1, Ew_2, Ew_3, Ew_4]$. From (7.13), (7.14), (7.16), (7.18) C becomes

$$C = [1 \quad 1 \quad 0 \quad 1]. \tag{7.23}$$

Finally, using (7.16), (7.17), $Tw_3 = Tw_4 = 0$ we have: $Ne_1 = e_3$, $Ne_2 = 2e_3$, $Ne_3 = 0$, $Ne_4 = 0$ or

$$N = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}. \tag{7.24}$$

The system $\{A, B, C, N, \mathscr{R}^4\}$ is a matrix representation for the RBSR. By Proposition 5.3, $\{A, B, C, N, \mathscr{R}^4\}$ is a minimal realization of θ.

The method outlined in Example 7.9 leads to a procedure for calculating the minimal bilinear realization of θ. Use Lemma 7.7 and S, E, T, to find a matrix representation for the RBSR of θ.

B. The factorable form. In linear system theory it is shown that the transfer function admits a finite dimensional linear realization if and only if it is rational. The natural generalization of this result is not true even if the transfer function is a generalized polynomial and $\mathscr{U} = \mathscr{K} = \mathscr{Y}$.

Example 7.10. Let $\theta \in \mathscr{H}_\Delta(\mathscr{K}, \mathscr{K})$ be given by $\theta = \{0, \lambda_1\lambda_2(1-\lambda_1\lambda_2)^{-1}, 0, 0, \cdots\}$, where $(1-\lambda_1\lambda_2)^{-1}$ is defined by (7.12). θ does not admit a finite dimensional bilinear realization. This follows from Corollary 7.6 and the fact that θ is not S-regular, as the following argument shows:

$$S^iS\theta u = \left\{0, \frac{\lambda_2^{i+1}u}{(1-\lambda_1\lambda_2)}, 0, 0, \cdots\right\}, \tag{7.25}$$

where $u \in \mathscr{U}$.

Let $\Sigma=\{A, B, C, N, \mathscr{X}\}$ be a finite dimensional bilinear system. From the Cayley–Hamilton Theorem:

$$F_m=\frac{1}{d(\lambda_m)}\sum_{i=0}^{M} R_i\lambda_m^i, \tag{7.26}$$

where $d(\lambda_m)$ is a scalar valued polynomial in λ_m, $d(0)\neq 0$, F_m is defined in (3.4), and $R_i\in\mathscr{L}(\mathscr{X};\mathscr{X})$ for $i=0, 1, \cdots, M$. Substituting (7.26) into (3.3) or (3.5) gives:

$$\theta_{n\Sigma}=\left[\sum_{i_n=0,\cdots,i_1=0}^{M,\cdots,M} CR_{i_n}NR_{i_{n-1}}N\cdots NR_{i_1}B\lambda_1^{i_1}\cdots\lambda_n^{i_n}\right]\left[\prod_{i=1}^{n} d(\lambda_i)\right]^{-1}, \tag{7.27}$$

and $\theta_\Sigma=\bigoplus\theta_{n\Sigma}$. Equation (7.27) is the form of interest.

Definition 7.11. Let $\theta=\bigoplus\theta_n\in\mathscr{H}(\mathscr{U};\mathscr{Y})$. θ is said to be *factorable* if and only if there exists a universal bound $M<\infty$ and a scalar valued polynomial $d(\lambda)$,

$$d(\lambda)=\sum_{i=0}^{M} d_i\lambda^i, \qquad d(0)\neq 0, \tag{7.28}$$

such that

$$\theta_n=\left[\sum_{i_n=0,\cdots,i_1=0}^{M,\cdots,M} K_n(i_1,\cdots,i_n)\lambda_1^{i_1}\cdots\lambda_n^{i_n}\right]\left[\prod_{i=1}^{n} d(\lambda_i)\right]^{-1}, \tag{7.29}$$

for all $n>0$ where $K_n(i_1,\cdots,i_n)\in\mathscr{L}(\mathscr{U}^n;\mathscr{Y})$ for $0\leqq i_j\leqq M$, $j=1, 2, \cdots, n$.

Remark 7.12. Definition 7.11 says that $\theta=\bigoplus\theta_n$ is factorable if and only if θ can be put in the form displayed by (7.29). For example the θ given in (7.11) is not in the form expressed by (7.29). By multiplying the numerator and denominator of θ_n by the appropriate terms, this $\theta=\bigoplus\theta_n$ can be put in the form (7.29); i.e., the θ given in (7.11) is factorable. Throughout this paper we use the same M in (7.28) and (7.29). A bound M is usually obtained by setting the appropriate $K(i_1,\cdots,i_n)$ or d_i equal to zero.

Remark 7.13. Definition 7.11 is stated for transfer functions in $\mathscr{H}$. This will be useful in studying state-affine systems.

By (7.27) all finite dimensional bilinear systems have factorable transfer functions. From this, one might suspect that all factorable θ's admit a finite dimensional bilinear realization. This is not true even if $\mathscr{U}=\mathscr{X}=\mathscr{Y}$. It turns out that all factorable θ's are S-regular and not necessarily T-regular.

Example 7.14. Let $\theta\in\mathscr{H}_\Delta(\mathscr{X},\mathscr{X})$ be given by

$$\theta=\bigoplus_{n=1}^{\infty}\left(\frac{1}{n!}\prod_{i=1}^{n}\frac{\lambda_i}{(1-\lambda_i)}\right). \tag{7.30}$$

Clearly θ is factorable. Applying T recursively with $S_1(1-\lambda_1)^{-1}=(1-\lambda_1)^{-1}$ and input $u=1$ gives:

$$T^kS\theta=\bigoplus_{n=1}^{\infty}\left[\frac{1}{(n+k)!}\left(\frac{1}{1-\lambda_1}\right)\prod_{i=2}^{n}\left(\frac{\lambda_i}{1-\lambda_i}\right)\right]. \tag{7.31}$$

Using the fact that $e^\lambda-1=\sum_{n=1}^{\infty}\lambda^n/n!$ is not rational, it is easy to show that $\{T^kS\theta\}_{k=0}^{\infty}$ is a linearly independent set. Thus θ is not T-regular.

To show that all factorable θ's are S-regular we need

Lemma 7.15. *Let $v\in\mathscr{S}_1(\mathscr{V})$ be rational; i.e.,*

$$v=\left(\sum_{i=0}^{M}\Psi_i\lambda_1^i\right)\frac{1}{d(\lambda_1)}, \tag{7.32}$$

where $d(\lambda)$ is the scalar valued polynomial given by (7.28) *and $\Psi_i \in \mathcal{V}$ for $i = 0, 1, \cdots, M$. Then*

$$(7.33)\qquad \bigvee_{i=0}^{\infty} \hat{S}_1^i v \subseteq \operatorname{span}\left\{\frac{\Psi_i \lambda_1^j}{d(\lambda_1)}\right\}_{i=0,j=0}^{M,M}.$$

In particular,

$$\dim\left(\bigvee_{i \geqq 0} \hat{S}_1^i \hat{S}_1 v\right) \leqq (M+1)^2 \quad \textit{and} \quad \bigvee_{i=0}^{\infty} \hat{S}_1^i v = \bigvee_{i=0}^{(M+1)^2} \hat{S}_1^i v.$$

Remark 7.16. $\hat{S}_1$ is the backward shift operator on $\mathcal{S}_1(\mathcal{V})$, i.e., $\hat{S}_1 u(\lambda_1) = (u(\lambda_1) - u(0))/\lambda_1$ if $u \in \mathcal{S}_1(\mathcal{V})$.

Proof. Without loss of generality, assume $d(0) = 1$. By applying $\hat{S}_1$ to (7.32):

$$(7.34)\qquad \hat{S}_1 v = \left[\sum_{i=1}^{M} (\Psi_i - \Psi_0 d_i)\lambda_1^{i-1}\right] \frac{1}{d(\lambda_1)}.$$

Repeated application of $\hat{S}_1$ to (7.34) leads to (7.33). □

PROPOSITION 7.17. *If $\theta \in \mathcal{H}$ is factorable, then θ is S-regular. Further, if M is an upper bound for θ given in Definition* 7.11 *then*

$$(7.35)\qquad \mathcal{W}_\theta^q = \mathcal{W}_\theta^q((M+1)^2), \quad \textit{for all } q > 0.$$

Proof. Let $\theta = \bigoplus \theta_n$ be given by (7.29). θ_n can be written as

$$(7.36)\qquad \theta_n = \frac{1}{d(\lambda_1)}\left[\sum_{i=0}^{M} P_{n,i}(\lambda_2, \cdots, \lambda_n)\lambda_1^i\right],$$

for all $n > 0$, where $P_{n,i} \in \mathcal{S}_n(\mathcal{L}(\mathcal{U}^n; \mathcal{Y}))$ and $P_{n,i}$ contains no λ_1^k terms ($k > 0$). For $u \in \mathcal{U}$, this implies

$$(7.37)\qquad \theta u = \bigoplus_{n=1}^{\infty} \theta_n u = \frac{1}{d(\lambda_1)}\left[\sum_{i=0}^{M} \lambda_1^i \left(\bigoplus_{n=1}^{\infty} P_{n,i} u\right)\right].$$

By Lemma 7.15 with $S = \hat{S}_1$, $\mathcal{V} = \mathcal{F}$, and (7.37):

$$(7.38)\qquad \bigvee_{i=0}^{\infty} S^i S\theta u = \bigvee_{i=0}^{(M+1)^2} S^i S\theta u.$$

Since M in (7.29) is independent of u, (7.35) holds if $q = 1$. For $q \geqq 1$ we note that the operators $T(u)$, $u \in \mathcal{U}$, and S do not destroy the form of θ displayed in (7.29) (this is shown by applying S, T to θ and using (7.34)). Thus (7.35) holds for all $q > 0$. □

From this proposition with Proposition 7.4, Corollary 7.5, and Corollary 7.6, we immediately obtain the following results.

COROLLARY 7.18. *$\theta \in \mathcal{H}_\Delta$ admits a finite dimensional bilinear realization if and only if θ is factorable, T-regular and condition* 7.7 *holds.*

COROLLARY 7.19. *Let $\theta \in \mathcal{H}_\Delta(\mathcal{U}; \mathcal{Y})$ with* $\dim(\mathcal{U}) < \infty$. *$\theta$ admits a finite dimensional bilinear realization if and only if θ is factorable and T-regular.*

COROLLARY 7.20. *Let $\theta \in \mathcal{H}_\Delta(\mathcal{U}; \mathcal{Y})$ be a generalized polynomial with* $\dim(\mathcal{U}) < \infty$. *$\theta$ admits a finite dimensional bilinear realization if and only if θ is factorable.*

Consider any $\theta = \bigoplus \theta_n \in \mathcal{H}_\Delta$ where each θ_n is rational. This condition is necessary for θ_n to admit a finite dimensional bilinear realization. To see if θ admits a finite dimensional bilinear realization, one can check for factorability, T-regularity and condition (7.7). Checking for factorability can become rather involved, i.e., one may have to factor polynomials of several variables. An alternate way to check for the existence of a finite dimensional bilinear realization is to apply Lemma 7.7 with Conditions (7.9), (7.10). This was done in Example 7.9 and Example 7.10. Note that the operators S_n, E_n can be applied directly to the rational functions θ_n, see (2.6) and (2.8). Therefore, Lemma 7.7 with Examples 7.9 and 7.10 gives a recursive procedure to find $\mathscr{W}_\theta$ and determine whether or not θ admits a finite dimensional bilinear realization.

Another result of this section is that it gives us a procedure to determine a minimal bilinear matrix realization of θ. Assuming θ admits a finite dimensional bilinear realization, an algorithm is summarized [17]: Let $\theta = \bigoplus \theta_n$ be given in rational form, i.e., each θ_n is a rational function. (i) Using Lemma (7.7), find a basis for $\mathscr{W}_\theta$. (ii) Using this basis for $\mathscr{W}_\theta$, find a matrix representation $\Sigma = \{A, B, C, N, \mathscr{K}^n\}$ for the RBSR of $\theta, \{S_\theta, S\theta, E_\theta, T_\theta, \mathscr{W}_\theta\}$. Thus Σ is a minimal realization of θ. Note that the above algorithm lives in the transform domain, see Example (7.9). It is easy to implement because the operators S_n and E_n can be directly applied to rational functions, see Example 7.9, Example 7.10, and (2.6) and (2.8). The simplicity of the algorithm is due to the shift operators S, E and our transform for a Volterra series input–output map. (It is this transform theory that lead to rational kernels θ_n which gives an easy implementation of the algorithm.)

8. Conclusion. In this paper we have used shift operators to present a theory of bilinear systems. Shift operators can be used to solve other nonlinear problems [2]. Our shift operator approach to nonlinear system theory is summarized in the following steps.

(i) Find the input–output map for Σ (Σ is a state-space representation).

(ii) Use shift operators to find a backward shift realization (BSR) for an input–output map.

(iii) Restrict the state-space in the BSR to obtain the restricted backward shift realization (RBSR) of the input–output map.

(iv) Find the linear operator H_Σ that sends Σ into the BSR.

(v) Use H_Σ and the RBSR to develop a theory of minimality, reachability and observability.

(vi) Use the RBSR to develop a theory of regularity and finite dimensional systems.

The importance of a transform representation for a Volterra series is well demonstrated in [4]. In this paper, we have introduced a new transform for a Volterra series. Our transform is defined as the n-dimensional z-transform of the kernels in the lower triangular Volterra series. For n-homogeneous systems, our transfer function [factorable transfer function] collapses into the regular transfer function [recognizable transfer function] given in [9], [10], and [29] respectively. The transform theory allowed us to use the concepts of rationality, factorability and regularity to further develop the theory of finite dimensional bilinear systems. The transform representation also provided an ideal setting for a realization algorithm; see Example 7.9 and [17]. This algorithm lives in the frequency domain. It is extremely easy to implement, because the operators S_n and E_n are naturally suited to act on rational functions; see (2.6) and (2.8). Obviously our realization theory can be converted to the time domain. (Recall that Λ is an isomorphism.) However, in this setting the concepts of rationality and factorability becomes obscure. Even more distressing is that a time domain realization algorithm is much harder to implement. One becomes rapidly convinced of this by working out several examples in both the time and frequency domain. Note a time domain version of our algorithm involves checking for linear independence of vectors $\bigoplus \tilde{v}_n(i_1, \cdots, i_n)$ formed by an infinite direct sum of infinite multivariable sequences. Therefore even constructing a matrix realization for a relatively "simple" kernel sequence $\{\theta_n\}$, will demonstrate the advantage of our transform theory.

Conclusion: Our transform representation provides a compact notation that is a real asset for finding minimal matrix realizations for a Volterra series! Finally, it is noted that our transform representation can be applied to other problems in nonlinear systems; it is used to determine the stability of certain homogeneous systems [10].

Other transforms for a Volterra series have also been used, [1], [4]. A particularly interesting approach involves the noncommutative formal series [13], [14], [15]. Here a finite number of noncommutative indeterminates $x_0, x_1, \cdots, x_n$ are used to obtain a transfer function for a bilinear system. The number n of noncommutative variables depends on the dimension of the input space $\mathcal{U}$. The noncommutativity of $x_0, x_1, \cdots, x_n$ is the price one pays for eliminating our (possibly) infinite number of commutating variables $\lambda_1, \lambda_2, \cdots$. In other words, there is a trade-off: one can choose between finitely many noncommutating variables and infinitely many commutating variables. Note that if the Volterra series is a polynomial, i.e., if θ is a generalized polynomial, then both theories use a finite number of indeterminates. However, ours commute.

Other approaches to realization theory are given in [25], [26], [31] and elsewhere. Basically these papers arrange the input–output data into one huge "Hankel matrix" and then use this matrix to solve the realization problem. In this paper, Hankel matrices were not used; the input–output data are given by the transfer function θ where θ sits in the space $\mathcal{H}$. We have also developed a theory of minimality, reachability and observability directly from the RBSR and the operator H_Σ. This approach is believed to be new. For other methods on proving some of the results in §§ 5 and 6, see [26], [31].

Acknowledgment. Conversations with Professor E. G. Gilbert, The University of Michigan, and Professor W. J. Rugh, The Johns Hopkins University, during the course of the research, have been most helpful.

REFERENCES

[1] P. Alper, *A consideration of the discrete Volterra series*, IEEE Trans. Automat. Control, AC-10 (1965), pp. 322–327.

[2] A. V. Balakrishnan, *On the state space theory of nonlinear systems*, in Functional Analysis and Optimization, E. R. Caianiello, ed., Academic Press, New York, 1966, pp. 15–36.

[3] J. S. Baras and R. W. Brockett, *H^2-functions and infinite-dimensional realization theory*, SIAM J. Control, 13 (1975), pp. 221–241.

[4] E. Bedrosain and S. Rice, *The output properties of Volterra systems (nonlinear systems with memory) driven by harmonic and Gaussian inputs*, Proc. Inst. Elec. Engr., 59 (1971), pp. 1688–1707.

[5] R. W. Brockett, *On the algebraic structure of bilinear systems*, in Theory and Applications of Variable Structure Systems, R. Mohler and A. Ruberti, eds., Academic Press, New York, 1972, pp. 153–168.

[6] ———, *Finite and infinite dimensional bilinear realizations*, J. Franklin Inst., 301 (1976), pp. 509–520.

[7] C. Bruni, G. DiPillo and G. Koch, *Bilinear systems: an appealing class of nearly linear systems in theory and applications*, IEEE Trans. Automat. Control, AC-19 (1974), pp. 334–348.

[8] C. T. Chen, *Introduction to Linear System Theory*, Holt, Rinehart, and Winston, New York, 1970.

[9] S. J. Clancy and W. J. Rugh, *On the realization problem for stationary, homogeneous discrete-time systems*, Automatica, 14 (1978), pp. 357–366.

[10] S. J. Clancy, G. E. Mitzel and W. J. Rugh, *On transfer function representations for homogeneous nonlinear systems*, IEEE Trans. Automat. Control, 24 (1979), pp. 242–249.

[11] P. D'Allessandro, A. Isidori and A. Ruberti, *Realization and structure theory of bilinear dynamical systems*, SIAM J. Control, 12 (1974), pp. 517–535.

[12] J. L. Doob, *Stochastic Processes*, John Wiley, New York,

[13] M. Fliess, *Un codage non commutatif pour certains systèmes echantillonnes non lineaires*, Inform. and Control, 38 (1978), pp. 264–287.

[14] ———, *Un outil algebrique: les series formelles non commutatures*, in Lecture Notes in Economics and Mathematical Systems, G. Marchesini and S. Mitter, eds., Springer-Verlag, Berlin, 1975, pp. 122–148.

[15] ———, *Matrices de Hankel*, J. Math. Pures Appl., 53 (1974), pp. 197–224.

[16] E. FORNASINI AND G. MARCHESINI, *Algebraic realization theory of bilinear discrete-time input–output maps*, J. Franklin Inst., 301 (1976), pp. 143–159.

[17] A. E. FRAZHO, *Shift operators and bilinear system theory*, Proceedings of the 1979 IEEE Conference on Decision and Control, San Diego, California, pp. 551–556.

[18] ———, *State-affine realization theory*, Proceedings of the 1979 Conference on Information Sciences and Systems, The Johns Hopkins University, Baltimore, MD, pp. 214–219.

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

[20] E. G. GILBERT, *Functional expansions for the response of nonlinear differential systems*, IEEE Trans. Automat. Control, AC-22 (1977), pp. 909–921.

[21] ———, *Bilinear and 2-power input–output maps: finite dimensional realizations and the role of the functional series*, IEEE Trans. Automatic Control, to appear.

[22] W. GREUB, *Linear Algebra*, Springer-Verlag, New York, 1975.

[23] H. HELSON, *Lectures on Invariant Subspaces*, Academic Press, New York, 1964.

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

[25] A. ISIDORI, *Direct construction of minimal bilinear realizations from nonlinear input-output maps*, IEEE Trans. Automat. Control, AC-18 (1973), pp. 626–631.

[26] A. ISIDORI AND A. RUBERTI, *Realization theory of bilinear systems*, in Geometric Methods in System Theory, D. Q. Mayne and R. W. Brockett, eds., D. Reidel, Dordrecht, 1973.

[27] G. KOCH, *A realization theorem for infinite dimensional bilinear systems*, Ricerche di Automatica, 3 (1973).

[28] Y. H. KU AND A. A. WOLF, *Volterra-Wiener functionals for the analysis of nonlinear systems*, J. Franklin Inst., 271 (1966), pp. 9–26.

[29] G. E. MITZEL AND W. J. RUGH, *Reaslization of stationary homogeneous systems: the degree* 2 *case*, Proceedings of the 1977 IEEE Conference on Decision and Control, New Orleans, LA, pp. 783–788.

[30] A. W. NAYLOR AND G. R. SELL, *Linear Operator Theory in Engineering and Science*, Holt, Rinehart, and Winston, New York, 1971.

[31] E. D. SONTAG, *Realization theory of discrete-time nonlinear systems, I. The bounded case*, IEEE Trans. Circuits and Systems, CAS-26 (1979), pp. 342–356.

[32] B. SZ.-NAGY AND C. FOIAS, *Harmonic Analysis of Operators on Hilbert Space*, North-Holland, Amsterdam, 1970.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 6, November 1980

0363-0129/80/1806-0007 $01.00/0

NEWTON'S METHOD AND THE GOLDSTEIN STEP-LENGTH RULE FOR CONSTRAINED MINIMIZATION PROBLEMS*

J. C. DUNN†

Abstract. A relaxed form of Newton's method is analyzed for the problem, $\min_\Omega F$, with Ω a convex subset of a real Banach space X, and $F:X\to\mathbb{R}^1$ twice differentiable in the sense of Fréchet. In this iterative scheme, feasible directions are gotten by minimizing local quadratic approximations Q to F, and the relaxation parameters, or step lengths, are obtained from Goldstein's rule. The local and global convergence theorems established here yield two significant extensions of an earlier theorem of Goldstein for the special case $\Omega=X=$ a Hilbert space. In one extension, growth rate conditions on the local approximation Q subsume the classical uniform positivity restriction on F''; connections are made here with a recently formulated classification scheme for singular and nonsingular extremals. In the second extension, uniform growth rate conditions are replaced by assumptions of the compactness and boundedness type. This development establishes global convergence of the Newton-Goldstein algorithm for a large class of problems with singular critical points.

1. Introduction. Newton's method has a natural extension for the constrained minimization problem,

$$\min_{y\in\Omega} F(y), \tag{1.1}$$

with Ω a nonempty subset of a real Banach space X and $F:X\to\mathbb{R}^1$ twice differentiable in the sense of Fréchet. Thus, for each $x\in\Omega$, let $Q(x,\cdot)$ denote the corresponding local quadratic approximation to $F(\cdot)-F(x)$, i.e.,

$$Q(x,y)=\langle F'(x),y-x\rangle+\tfrac{1}{2}\langle F''(x)(y-x),y-x\rangle \tag{1.2}$$

where $\langle u,v\rangle$ signifies the value of a linear functional $u\in X^*$ ($=$ dual of X) at $v\in X$, and $F'(x)$ and $F''(x)$ are, respectively, the first and second Fréchet derivatives of F at x. With x fixed, consider the problem

$$\min_{y\in\Omega} Q(x,y), \tag{1.3}$$

in place of (1.1), and let $T_Q(x)$ denote the corresponding solution set,

$$T_Q(x)=\{\hat{x}\in\Omega\,|\,Q(x,\hat{x})=\inf_{y\in\Omega}Q(x,y)\}. \tag{1.4}$$

As x ranges over Ω, (1.4) defines a set-valued map $T_Q:\Omega\to 2^\Omega$. If Ω and F are convex, the fixed points ξ of this map are precisely the minimizers of F in Ω; i.e., $F(\xi)=\inf_{y\in\Omega}F(y)$ if and only if $\xi\in T_Q(\xi)$. More generally, if Ω is convex but F is not convex, the fixed points of T_Q and the minimizers of F are always extremals, i.e., fixed points of the operator T_L defined by,

$$T_L(x)=\{\bar{x}\in\Omega\,|\,L(x,\bar{x})=\inf_{y\in\Omega}L(x,y)\}, \tag{1.5}$$

where $L(x,\cdot)$ is the local *linear* approximation to $F(\cdot)-F(x)$,

$$L(x,y)=\langle F'(x),y-x\rangle. \tag{1.6}$$

* Received by the editors November 21, 1978, and in final revised form March 10, 1980. This investigation was supported by the National Science Foundation under Research Grant #ENG 78-03385.

† Department of Mathematics, North Carolina State University, Raleigh, North Carolina, 27650.

If neither Ω nor F is convex, no simple relationship exists between the fixed points of T_Q or T_L and the solutions of (1.1); however, even at this level of generality certain types of regular fixed points of T_Q do provide strong *local minimizers* of F in Ω; if F'' is continuous, these special fixed points are also strong *local attractors* for the Picard iteration scheme,

$$x_{n+1} \in T_Q(x_n), \qquad x_0 \in \Omega. \tag{1.7}$$

When $\Omega = X$ and F is convex, (1.7) reduces to the classical Newton recursion

$$F'(x_n) + F''(x_n)(x_{n+1} - x_n) = 0. \tag{1.8}$$

Kantorovich [1], Goldstein [2], and others have investigated (1.8), and the more general process (1.7) has been analyzed by Levitin and Polyak [3] under conditions somewhat more restrictive than those invoked in the present article. More recently, Pshenichnyi [4] and Robinson [5] have studied a different extension of the Newton process for solving mixed systems of equations and inequalities, and Robinson [6] has investigated convergence rates for a class of nonlinear programming algorithms containing (1.7) as a special case.

In [7], [8] it is shown that the asymptotic behavior of conditional gradient sequences is sensitive to the rate at which the local linear approximation $L(\xi, y)$ grows as y moves away from an extremal ξ into Ω. It turns out that the Newton iterates (1.7) are governed by the growth rate of the corresponding local quadratic approximation $Q(\xi, \cdot)$ near a fixed point ξ of T_Q. For the special case $\Omega = X$, a modification of Goldstein's development in [2] shows that local Lipschitz conditions of the sort,

$$x \in \Omega, \quad \hat{x} \in T_Q(x), \quad \|x - \xi\| \leqq \sigma \Rightarrow \|\hat{x} - \xi\| \leqq K(\sigma)\|x - \xi\|, \tag{1.9a}$$

are satisfied by the Newton operator T_Q, with

$$\lim_{\sigma \to 0^+} K(\sigma) = 0, \tag{1.9b}$$

provided F'' is continuous and the growth condition,

$$\exists \gamma > 0, \quad \forall y \in \Omega, \quad Q(\xi, y) \geqq \gamma \|y - \xi\|^2 \tag{1.10}$$

is satisfied, with $Q(\xi, y) = \frac{1}{2}\langle F''(\xi)(y - \xi), y - \xi\rangle$. Growth conditions of this type also arise naturally in analyses of gradient methods and quasi-Newton methods [9]. In § 3 it is shown that (1.10) is also the key to (1.9) when Ω is an arbitrary subset of X, except that now $F'(\xi)$ need not vanish, and therefore $Q(\xi, y) = \langle F'(\xi), y - \xi\rangle + \frac{1}{2}\langle F''(\xi)(y - \xi), y - \xi\rangle$. The regularity condition (1.10) implies that ξ is a strong local minimizer of F in Ω. The Lipschitz condition (1.9) implies that any sequence $\{x_n\} \subset \Omega$ generated by (1.7) converges to ξ superlinearly (i.e., faster than any geometric progression), provided x_0 is sufficiently close to ξ, e.g., close enough to make $K(\|x_0 - \xi\|) < 1$. The *strong regularity* condition,

$$\exists \gamma_s > 0, \quad \forall y \in \Omega, \quad Q(\xi, y) \geqq \gamma_s \|y - \xi\|, \tag{1.11}$$

also implies that ξ is a strong local minimizer of F, and for continuous F'', guarantees that $T_Q(x) = \xi$ for x near ξ; evidently, this last condition insures that Newton iterates which converge to ξ must actually *terminate* at ξ beyond some finite value of n.

If Ω and F are convex and ξ is a minimizer of F, then $\langle F'(\xi), y - \xi\rangle \geqq 0$ and $\langle F''(\xi)(y - \xi), y - \xi\rangle \geqq 0$ for all $y \in \Omega$. Under these circumstances (1.10) will hold if *either* of the following conditions is satisfied:

$$\exists \alpha > 0, \quad \forall y \in \Omega, \quad \langle F'(\xi), y - \xi\rangle \geqq \alpha \|y - \xi\|^2, \tag{1.12}$$

or

$$\exists \mu > 0, \quad \forall y \in \Omega, \quad \langle F''(\xi)(y-\xi), y-\xi\rangle \geqq \mu\|y-\xi\|^2. \tag{1.13}$$

The second condition is related to the curvature of the level surface $\{x \in X | F(x) = F(\xi)\}$ at ξ, and is automatically satisfied if F'' is positive definite, as in [2] and [3]. On the other hand, when $F'(\xi) \neq 0$ the first condition is a manifestation of curvature in the boundary of Ω at ξ, and plays an important role in the analysis of conditional gradient processes; extremals ξ satisfying (1.12) are said to be *regular* in [7]. If $F'(\xi) \neq 0$, and if Ω obeys the uniform convexity condition,

$$\exists \nu > 0, \quad \forall x, y \in \Omega, \quad \|z\| \leqq \nu\|x-y\|^2 \Rightarrow \frac{x+y}{2} + z \in \Omega, \tag{1.14}$$

then (1.12) holds with $\alpha = 2\|F'(\xi)\|\nu$. However, (1.12) does not require uniform convexity; in particular, an L^1 version of (1.12) is closely related to classical notions of nonsingularity for optimal control problems on hypercubes. These points are considered at length in [7].

The strong regularity condition,

$$\exists \alpha_s > 0, \quad \forall y \in \Omega, \quad \langle F'(\xi), y-\xi\rangle \geqq \alpha_s\|y-\xi\|, \tag{1.15}$$

is also introduced in [7]. At every extremal ξ, the vector $-F'(\xi)$ must lie in $K_\Omega(\xi)$ = the cone of normals to Ω at ξ; however at a strongly regular extremal, $-F'(\xi)$ falls in the *interior* of $K_\Omega(\xi)$. (This can happen, for instance, at the vertices of polyhedral convex sets Ω). For convex F, condition (1.15) implies (1.11).

While convergent Newton iterates tend to converge rapidly, it can easily happen that $\{x_n\}$ does not converge to *any* limit even though F is convex and x_0 is quite close to an extremal ξ. For example, on $\Omega = X = \mathbb{R}^1$, consider the strictly convex function F defined by the conditions,

$$\frac{d^2F}{dx^2}(x) = a + \frac{1}{1+4(x/\Delta)^2} \geqq a,$$

$$\frac{dF}{dx}(0) = F(0) = 0,$$

$$a = \tfrac{1}{2}\tan^{-1} 2 - \tfrac{2}{5} > 0,$$

where Δ is an arbitrarily small fixed positive number. This function has a unique minimizer at $\xi = 0$. Moreover, the associated Newton operator T_Q is single-valued and,

$$x \geqq \Delta \Rightarrow T_Q(x) \leqq -\Delta,$$

$$x \leqq -\Delta \Rightarrow T_Q(x) \geqq \Delta.$$

Consequently, the corresponding Newton iterates diverge if $|x_0 - \xi| = |x_0| \geqq \Delta$.

On the other hand, for *convex* F and Ω it is sometimes possible to induce convergence from all remote starting points $x_0 \in \Omega$ with certain *relaxed* Newton schemes of the form

$$x_{n+1} = x_n + \omega_n(\hat{x}_n - x_n), \qquad x_0 \in \Omega, \tag{1.16a}$$

$$\hat{x}_n \in T_Q(x_n), \tag{1.16b}$$

where the relaxation parameters, or "step lengths," $\omega_n \in [0, 1]$ are usually chosen to secure a decrease in the functional F from its value at x_n. In the classical line minimization scheme, ω_n is determined implicitly by

$$F(x_n+\omega_n(\hat{x}_n-x_n))=\inf_{0\leqq\omega\leqq 1} F(x_n+\omega(\hat{x}_n-x_n)).$$

In the implicit scheme of Goldstein [2], [10], ω_n is determined as follows. Fix δ in $(0, \frac{1}{2})$ and put

$$g(x, \hat{x}; \omega)=\frac{F(x)-F(x+\omega(\hat{x}-x))}{\omega\langle F'(x), x-\hat{x}\rangle},$$

when $\langle F'(x), x-\hat{x}\rangle\neq 0$. For continuous F', g is continuous in ω on $[0, 1]$ with x and $\hat{x}$ fixed. Moreover, $\lim_{\omega\to 0^+} g(x, \hat{x}; \omega)=1$; consequently if $g(x, \hat{x}; 1)<\delta$, the set

$$W_\delta(x, \hat{x})=\{\omega\in[0, 1]|\delta\leqq g(x, \hat{x}; \omega)\leqq 1-\delta\}$$

is not empty. For convex F, one can show that $\langle F'(x), x-\hat{x}\rangle$ is always nonnegative for $\hat{x}\in T_Q(x)$ (see inequality (4.3)), hence $\hat{x}-x$ specifies a "descent direction" for F, and the corresponding set,

$$\bar{W}_\delta(x, \hat{x})=\begin{cases}\{0\} & \text{if } \langle F'(x), x-\hat{x}\rangle=0,\\ \{1\} & \text{if } \langle F'(x), x-\hat{x}\rangle>0 \text{ and } g(x, \hat{x}; 1)\geqq\delta,\\ W_\delta(x, \hat{x}) & \text{if } \langle F'(x), x-\hat{x}\rangle>0 \text{ and } g(x, \hat{x}; 1)<\delta,\end{cases}$$

is well defined and nonempty at each $(x, \hat{x})$ with $\hat{x}\in T_Q(x)$. The associated step-length rule now requires that

$$\omega_n\in\bar{W}_\delta(x_n, \hat{x}_n) \tag{1.17}$$

in (1.16), or equivalently,

$$x_{n+1}\in T_{Q,\delta}(x_n), \qquad x_0\in\Omega, \tag{1.18}$$

where $T_{Q,\delta}$ is the set-valued operator defined by

$$T_{Q,\delta}(x)=\{y\in\Omega|\exists\hat{x}\in T_Q(x), \exists\omega\in\bar{W}_\delta(x, \hat{x}), y=x+\omega(\hat{x}-x)\} \tag{1.19}$$

as x ranges over Ω. In a very rough sense, (1.17) approximates the line minimization condition along the feasible direction specified by $\hat{x}_n-x_n$ (see [8]). A related but somewhat simpler step-size scheme proposed by Armijo in [11] for gradient methods has also been adapted for classical Newton processes (Polak [12]).

When $\Omega=X=$ a Hilbert space, Goldstein has shown that the iterates generated by (1.18) always produce minimizing sequences with extremal limit points, provided that on the level set

$$\Omega_0=\{x\in\Omega|F(x)\leqq F(x_0)\}, \tag{1.20}$$

F'' is continuous and uniformly bounded, and satisfies the uniform convexity condition

$$\langle F''(x)v, v\rangle\geqq\mu\|v\|^2 \tag{1.21}$$

for some $\mu>0$ and all $v\in X$ and all $x\in\Omega_0$. Moreover, under these conditions, one finds that for sufficiently large n, either $F_n=\inf F$ or $\omega_n=1$ (i.e., (1.18) reduces to the basic Newton scheme (1.7)). Extensions of these results are established in § 4 below for convex Ω in a Banach space. In one of these extensions the uniform convexity assumption (1.21) is replaced by a weaker uniform growth condition on Q, viz.,

$$Q(x, y)-Q(x, \hat{x})\geqq\gamma\|y-\hat{x}\|^2, \tag{1.22}$$

for some $\gamma > 0$, all $y \in \Omega$, all $\hat{x} \in T_Q(x)$, uniformly for $x \in \Omega_0$. In another extension, compactness of Ω_0 is the dominant assumption.

The present analysis reveals that the convergence behavior of Newton methods depends on the structure of the feasible set Ω only indirectly, through the growth rate of the quadratic functions Q. However, when it comes to implementing (1.7) or (1.18), the nature of Ω assumes a new and more immediate significance. For polyhedral Ω, the subproblem (1.3) can be attacked with extensions of the simplex algorithm such as Wolf's method (cf. [13]), or still more specialized procedures suited to the particular type of linear constraints which specify Ω. One can also devise effective specialized algorithms for (1.3) when Ω is a sphere, or the intersection of a sphere and a hyperplane, and so on. Beyond this, there is little that can be said here about the difficulty of (1.3) vis-a-vis (1.1) when Ω is an arbitrary set, or even an arbitrary convex set. The related question of how imperfections in the solution of (1.3) propagate through a computation with (1.7) or (1.18) has practical significance but receives no further consideration here.

2. General lemmas. The following results are used later on in the analysis of the Newton algorithms (1.7) and (1.18).

LEMMA 2.1. *Let X and Y be real Banach spaces and let G be a real functional on the Cartesian product $X \times Y$. Let Ω be a nonempty subset of Y and define the associated functional $\bar{\psi}: X \to (-\infty, \infty]$ by the rule*

$$\bar{\psi}(x) = \sup_{y \in \Omega} G(x, y) \leqq \infty.$$

Suppose that for each fixed $y \in \Omega$, the functional $G(\cdot, y): X \to \mathbb{R}^1$ is continuous at $x \in \Omega$. Then $\bar{\psi}$ is lower semicontinuous at x.

Proof. If $\lim_{n \to \infty} \|x_n - x\| = 0$, then for each fixed $y \in \Omega$,

$$\begin{aligned} \varliminf_{n \to \infty} \bar{\psi}(x_n) &\geqq \varliminf_{n \to \infty} G(x_n, y) \\ &= \lim_{n \to \infty} G(x_n, y) \\ &= G(x, y). \end{aligned}$$

Consequently,

$$\begin{aligned} \varliminf_{n \to \infty} \bar{\psi}(x_n) &\geqq \sup_{y \in \Omega} G(x, y) \\ &= \bar{\psi}(x). \end{aligned}$$

Q.E.D.

Note 2.1. See [14] for an extensive treatment of the continuity properties of maximum and minimum sets.

LEMMA 2.2. *Let X be a real Banach space, let $F: X \to \mathbb{R}^1$ have a second Fréchet derivative F'' near $x \in \Omega$, and suppose that F'' is continuous at x. Then for all $u, v \in X$,*

$$\langle F''(x)u, v \rangle = \langle F''(x)v, u \rangle. \tag{2.1}$$

Proof. The following proof is a straightforward generalization from the special case $X = \mathbb{R}^2$ treated in [15]; still more general symmetry theorems are proved in [16], [17] for Nth order derivatives.

For fixed $h \geqq 0$, put

$$D(u, v) = F(x + hu + hv) - F(x + hu) - F(x + hv) + F(x),$$

and

$$\phi(t) = F(x + tu + hv) - F(x + tu),$$

with $0 \leqq t \leqq h$. Then $D = \phi(h) - \phi(0)$, and consequently for some $\tau_1 \in [0, h]$

$$\begin{aligned} D &= \frac{d}{dt}\phi(t)|_{t=\tau_1} \cdot h \\ &= \{\langle F'(x + \tau_1 u + hv), u\rangle - \langle F'(x + \tau_1 u), u\rangle\}h. \end{aligned}$$

A second application of the mean value theorem now yields

$$D = \langle F''(x + \tau_1 u + \sigma_1 v)v, u\rangle h^2, \tag{2.2}$$

for some $\sigma_1 \in [0, h]$. On the other hand, if one puts

$$\phi(t) = F(x + hu + tv) - F(x + tv),$$

then $D = \phi(h) - \phi(0)$, and a similar argument yields

$$D = \langle F''(x + \tau_2 u + \sigma_2 v)u, v\rangle h^2, \tag{2.3}$$

for some $\tau_2, \sigma_2 \in [0, h]$. The symmetry condition (2.1) follows from (2.2) and (2.3) in the limit as $h \to 0$ Q.E.D.

LEMMA 2.3. *Let X be a real Banach space, let Ω be a nonempty convex subset of X, and let $F: X \to \mathbb{R}^1$ have a second Fréchet derivative F''. Furthermore, let $L(x, \cdot)$ and $Q(x, \cdot)$ signify the local linear approximation and the local quadratic approximation to $F(\cdot) - F(x)$ at x, defined by* (1.6) *and* (1.2) *respectively. Finally, let T_L and T_Q denote the set-valued operators in* (1.5) *and* (1.4), *and let Ω_F be the set of minimizers of F in Ω. Then*

$$\xi \in \Omega_F \Rightarrow \xi \in T_L(\xi), \tag{2.4}$$

and

$$\xi \in T_Q(\xi) \Rightarrow \xi \in T_L(\xi). \tag{2.5}$$

If F is convex, the converses of (2.4) *and* (2.5) *also hold.*

Proof. Proofs for (2.4) and its converse for convex F may be found in standard references, e.g., [18], [19].

If $\xi \notin T_L(\xi)$, there is a z in Ω for which

$$\langle F'(\xi), z - \xi\rangle \leqq -\varepsilon < 0.$$

For $\alpha \in [0, 1]$, put $z_\alpha = \xi + \alpha(z - \xi)$. Then $z_\alpha \in \Omega$, and

$$\begin{aligned} Q(\xi, z_\alpha) &= -\alpha\varepsilon + \tfrac{1}{2}\alpha^2\langle F''(\xi)(z - \xi), z - \xi\rangle \\ &\leqq -\alpha\varepsilon + \tfrac{1}{2}\alpha^2\|F''(\xi)\|\|z - \xi\|^2. \end{aligned}$$

Since the right side of this inequality is negative for sufficiently small $\alpha > 0$, it follows that $\xi \notin T_Q(\xi)$. This establishes (2.5). If F is convex and $\xi \in T_L(\xi)$, then

$$\begin{aligned} Q(\xi, y) &= \langle F'(\xi), y - \xi\rangle + \tfrac{1}{2}\langle F''(\xi)(y - \xi), y - \xi\rangle\rangle \\ &\geqq 0 \\ &= Q(\xi, \xi), \end{aligned}$$

for all $y \in \Omega$, i.e., $\xi \in T_Q(\xi)$. Q.E.D.

Note 2.2. If F is not convex, the condition $\xi \in T_Q(\xi)$ is neither necessary nor sufficient for $\xi \in \Omega_F$. For example, consider $F(x) = x^3$ on $\Omega = [-1, 1] \subset \mathbb{R}^1$. F achieves its global minimum over Ω at $\xi = -1$ and nowhere else; however, $Q(-1, y) = \frac{dF}{dx}(-1)(y + 1) + \frac{1}{2}\frac{d^2F}{dx^2}(-1)(y + 1)^2 = 3(y + 1) - 3(y + 1)^2$, and this quadratic function attains its minimum over Ω at $y=1$. Thus, $-1 \notin T_Q(-1)=\{1\}$. Moreover, at $\xi=0$, one has $\frac{dF}{dx}(0)=\frac{d^2F}{dx^2}(0)=0$; consequently $Q(0,y)\equiv 0$ and therefore $0 \in T_Q(0)=[-1,1]$, even though the inflection point 0 is not even a *local* minimizer of F over Ω.

DEFINITION 2.1.

i) x is a *regular point* of the operator T_Q in (1.4) if and only if $T_Q(x)$ consists of a single element $\{\hat{x}\}$, and

$$Q(x,y)-Q(x,\hat{x}) \geqq \gamma\|y-\hat{x}\|^2, \tag{2.6}$$

for some $\gamma>0$ and all $y \in \Omega$. In particular, ξ is a *regular fixed point* of T_Q if and only if $T_Q(\xi)=\{\xi\}$ and $Q(\xi,y)$ satisfies (1.10); i.e., for all y in Ω.

$$Q(\xi,y) \geqq \gamma\|y-\xi\|^2.$$

ii) x is a *strongly regular point* of T_Q if and only if $T_Q(x)=\{\hat{x}\}$ and

$$Q(x,y)-Q(x,\hat{x}) \geqq \gamma_s\|y-\hat{x}\|, \tag{2.7}$$

for some $\gamma_s>0$ and all $y \in \Omega$. In particular, ξ is a *strongly regular fixed point* of T_Q if and only if $T_Q(\xi)=\{\xi\}$ and $Q(\xi,y)$ satisfies (1.11).

Note 2.3. According to Lemma 2.3, $\xi \in T_Q(\xi) \Rightarrow \xi \in T_L(\xi)$, therefore when $\Omega = X$ one has $F'(\xi)=0$ and consequently $Q(\xi,y)=\frac{1}{2}\langle F''(\xi)(y-\xi), y-\xi\rangle$ at all fixed points of T_Q. In this special case, regularity of ξ relative to T_Q means that $F''(\xi)$ is positive definite. More generally, when Ω is a convex subset of X, the linear term $\langle F'(\xi), y-\xi\rangle$ is always nonnegative for $y \in \Omega$ at an extremal ξ; consequently if $F''(\xi)$ is positive definite, then ξ is a regular fixed point of T_Q even though ξ may be a *singular* extremal in the sense of [7] (i.e., $\langle F'(\xi), y-\xi\rangle = 0$ for some $y \in \Omega$, $y \notin \xi$). On the other hand, ξ can be a regular fixed point of T_Q even though $F''(\xi)$ is indefinite, provided the linear term in $Q(\xi,y)$ grows rapidly enough because of "curvature" in the boundary of Ω at ξ; in particular, if $F''(\xi)$ is positive semidefinite and ξ is a regular extremal (i.e., (1.12) holds at ξ), then ξ is a regular fixed point of T_Q. Thus, if F is convex and ξ is an extremal, then (1.12) ⇒ (1.10) and (1.13) ⇒ (1.10); furthermore, (1.15) ⇒ (1.11). Finally, if Ω is bounded, then (2.7 ⇒ (2.6) and (1.11) ⇒ (1.10).

LEMMA 2.4. *Let X be a real Banach space, let Ω be a nonempty subset of X, and suppose that $F: X \to \mathbb{R}^1$ has a second Fréchet derivative F''. If F'' is continuous at ξ and if ξ is a regular or strongly regular fixed point of the operator T_Q in* (1.4), *then ξ is a proper strong local minimizer of F over Ω.*

Proof. If (1.10) holds at ξ, then for $y \in \Omega$, there is a ζ on the line segment joining ξ to y for which

$$\begin{aligned} F(y)-F(\xi) &= \langle F'(\xi), y-\xi\rangle + \tfrac{1}{2}\langle F''(\zeta)(y-\xi), y-\xi\rangle \\ &= Q(\xi,y)+\tfrac{1}{2}\langle (F''(\zeta)-F''(\xi))(y-\xi), y-\xi\rangle \\ &\geqq (\gamma-\tfrac{1}{2}\|F''(\zeta)-F''(\xi)\|)\|y-\xi\|^2. \end{aligned}$$

Choose $\sigma>0$ such that $\|z-\xi\|<\sigma \Rightarrow \|F''(z)-F''(\xi)\|<2\gamma$. Then $0<\|y-\xi\|<\sigma$ and $y \in \Omega \Rightarrow \|\zeta-\xi\|<\sigma$ and $F(y)-F(\xi)>0$. A similar argument leads to the same conclusion when (1.11) holds at ξ. Q.E.D.

LEMMA 2.5. *Let Ω be a nonempty subset of a real Banach space X and let $F: X \to \mathbb{R}^1$ have a second Fréchet derivative F''. Suppose that ξ is a regular fixed point of the operator T_Q in* (1.4) *and that F'' is continuous at ξ. Then for all $x \in \Omega$ sufficiently close to ξ, T_Q satisfies the local Lipschitz conditions* (1.9). *Moreover, if ξ also satisfies the strong regularity condition* (1.11), *then $T_Q(x)=\{\xi\}$ for all x sufficiently close to ξ.*

Proof. With some manipulation, it follows from (1.2) and Lemma 2.2 that for all $\hat{x} \in T_Q(x)$,

$$\begin{aligned} 0 &\leqq Q(x,\xi) - Q(x,\hat{x}) \\ &= -\langle F'(x), \hat{x}-\xi\rangle - \tfrac{1}{2}\langle F''(x)(\hat{x}-\xi), \hat{x}-\xi\rangle + \langle F''(x)(x-\xi), \hat{x}-\xi\rangle. \end{aligned}$$

Consequently, the mean value theorem yields

$$\begin{aligned} \langle F'(\xi), \hat{x}-\xi\rangle &\leqq \langle F'(\xi), \hat{x}-\xi\rangle + Q(x,\xi) - Q(x,\hat{x}) \\ &= \langle (F''(x) - F''(\zeta))(x-\xi), \hat{x}-\xi\rangle - \tfrac{1}{2}\langle F''(x)(\hat{x}-\xi), \hat{x}-\xi\rangle, \end{aligned}$$

for some ζ on the line segment joining x to ξ; equivalently,

$$\langle (F''(x) - F''(\zeta))(x-\xi), \hat{x}-\xi\rangle \geqq Q(\xi, \hat{x}) + \tfrac{1}{2}\langle (F''(x) - F''(\xi))(\hat{x}-\xi), \hat{x}-\xi\rangle.$$

According to (1.10), one then has

$$\|F''(x)-F''(\zeta)\| \|x-\xi\| \|\hat{x}-\xi\| \geqq (\gamma - \tfrac{1}{2}\|F''(x)-F''(\xi)\|)\|\hat{x}-\xi\|^2. \tag{2.8}$$

Since F'' is continuous at ξ, it follows that for sufficiently small $\sigma > 0$,

$$\varepsilon(\sigma) \triangleq \sup_{\|y-\xi\| \leqq \sigma} \|F''(y) - F''(\xi)\| < \infty, \tag{2.9}$$

with

$$\lim_{\sigma \to 0^+} \varepsilon(\sigma) = 0.$$

Consequently, for $\|x-\xi\|$ sufficiently small, conditions (1.9) hold with

$$K(\sigma) = \frac{2\varepsilon(\sigma)}{\gamma - \frac{1}{2}\varepsilon(\sigma)}. \tag{2.10}$$

Finally, if (1.11) also holds, one obtains

$$\|F''(x)-F''(\zeta)\| \|x-\xi\| \|\hat{x}-\xi\| \geqq (\gamma_s - \tfrac{1}{2}\|F''(x)-F''(\xi)\| \|\hat{x}-\xi\|)\|\hat{x}-\xi\| \tag{2.11}$$

along with (2.8), where ζ is once again somewhere on the line segment joining x to ξ. But in view of (1.9) and the continuity of F'' at ξ, it follows that for all $\hat{x} \in T_Q(x)$,

$$\|F''(x)-F''(\zeta)\| \|x-\xi\| < \gamma_s - \tfrac{1}{2}\|F''(x)-F''(\xi)\| \|\hat{x}-\xi\|,$$

provided x is sufficiently close to ξ. Therefore (2.11) implies that $\hat{x} = \xi$ for all $\hat{x} \in T_Q(x)$ with $x \in \Omega$ in some sufficiently small neighborhood of ξ. Q.E.D.

LEMMA 2.6. *Let Ω be a nonempty convex subset of a real Banach space X, and let $F: X \to \mathbb{R}^1$ have a second Fréchet derivative F''. Suppose that ξ is a regular fixed point of the operator T_Q in* (1.4) *and that F'' is continuous at ξ. Then for all $\varepsilon > 0$ there is a corresponding $d(\varepsilon) > 0$ such that*

$$x \in \Omega, \quad \|x-\xi\| < d(\varepsilon), \quad \text{and} \quad \hat{x} \in T_Q(x) \Rightarrow Q(x,x) - Q(x,\hat{x}) \geqq (\gamma - \varepsilon)\|x-\hat{x}\|^2, \tag{2.12}$$

where $\gamma > 0$ is the constant in the regularity condition (1.10).

Proof. Since $Q(x,x) = 0$, one has

$$\begin{aligned} Q(x,x) - Q(x,\hat{x}) &= -Q(x,\hat{x}) \\ &= \langle F'(x), x-\hat{x}\rangle - \tfrac{1}{2}\langle F''(x)(\hat{x}-x), \hat{x}-x\rangle. \end{aligned} \tag{2.13}$$

Furthermore, since Ω is convex and $\hat{x}$ minimizes $Q(x, \cdot)$ over Ω, it follows that

$$\begin{aligned}0 &\leqq \langle Q'_y(x, \hat{x}), y-\hat{x}\rangle \\ &= \langle F'(x)+F''(x)(\hat{x}-x), y-\hat{x}\rangle,\end{aligned} \tag{2.14}$$

for all $y \in \Omega$. In particular, for $y=\xi$, this gives

$$\langle F'(x), \xi-\hat{x}\rangle \geqq -\langle F''(\xi)(\hat{x}-x), \xi-\hat{x}\rangle,$$

and consequently,

$$\begin{aligned}Q(x,x)-Q(x,\hat{x}) \geqq \langle F'(x), x-\xi\rangle - \langle F''(x)(\hat{x}-x), \xi-\hat{x}\rangle \\ -\tfrac{1}{2}\langle F''(x)(\hat{x}-x), \hat{x}-x\rangle.\end{aligned} \tag{2.15}$$

By writing

$$\begin{aligned}\langle F''(x)(x-\xi), x-\xi\rangle &= \langle F''(x)(x-\hat{x}+\hat{x}-\xi), x-\hat{x}+\hat{x}-\xi\rangle \\ &= \langle F''(x)(x-\hat{x}), x-\hat{x}\rangle + \langle F''(x)(x-\hat{x}), \hat{x}-\xi\rangle \\ &\quad + \langle F''(x)(\hat{x}-\xi), x-\hat{x}\rangle + \langle F''(x)(\hat{x}-\xi), \hat{x}-\xi\rangle,\end{aligned}$$

and applying Lemma 2.2, one can carry (2.15) further to

$$\begin{aligned}Q(x,x)-Q(x,\hat{x}) &\geqq \langle F'(x), x-\xi\rangle - \tfrac{1}{2}\langle F''(x)(x-\xi), x-\xi\rangle + \tfrac{1}{2}\langle F''(x)(\hat{x}-\xi), \hat{x}-\xi\rangle \\ &= Q(\xi, x) + \langle F'(x)-F'(\xi), x-\xi\rangle - \langle F''(x)(x-\xi), x-\xi\rangle \\ &\quad + \tfrac{1}{2}\langle F''(x)(\hat{x}-\xi), \hat{x}-\xi\rangle.\end{aligned}$$

For all x sufficiently close to ξ, condition (1.10), Lemma 2.5, and the mean value theorem now yield

$$\begin{aligned}Q(x,x)-Q(x,\hat{x}) &\geqq Q(\xi, x) + \langle (F''(\zeta)-F''(x))(x-\xi), x-\xi\rangle \\ &\quad + \tfrac{1}{2}\langle F''(x)(\hat{x}-\xi), \hat{x}-\xi\rangle \\ &\geqq [\gamma - \|F''(\zeta)-F''(x)\| - \tfrac{1}{2}\|F''(x)\|(K(\|x-\xi\|))^2]\|x-\xi\|^2,\end{aligned} \tag{2.16}$$

where ζ is somewhere on the line segment joining x to ξ, and K is the Lipschitz constant in (2.10). A second application of Lemma 2.5 gives

$$\begin{aligned}\|x-\xi\| &= \|x-\hat{x}+\hat{x}-\xi\| \\ &\geqq \|x-\hat{x}\| - \|\hat{x}-\xi\| \\ &\geqq \|x-\hat{x}\| - K(\|x-\xi\|)\|x-\xi\|,\end{aligned}$$

and therefore

$$\|x-\xi\|^2 \geqq \frac{1}{(1+K(\|x-\xi\|))^2} \cdot \|x-\hat{x}\|^2, \tag{2.17}$$

for x near ξ and $\hat{x} \in T_Q(x)$. It follows from (2.16) and (2.17) that for all x sufficiently close to ξ, and all $\hat{x} \in T_Q(x)$,

$$Q(x,x)-Q(x,\hat{x}) \geqq \gamma'\|x-\hat{x}\|^2,$$

with

$$\gamma' = [\gamma - \|F''(\zeta)-F''(x)\| - \tfrac{1}{2}\|F''(x)\|(K(\|x-\xi\|))^2] \cdot [1+K(\|x-\xi\|)]^{-2},$$

and ζ somewhere on the line segment joining x to ξ. Condition (2.12) is now an immediate consequence of (1.9b) and the continuity of F'' at ξ. QED

LEMMA 2.7. *Let Ω, X, F, and ξ satisfy the conditions of Lemma* 2.6, *and in addition, suppose that F is convex. Fix δ in* $(0, \frac{1}{2})$. *Then for all x sufficiently close to ξ,*

$$T_{Q,\delta}(x) = T_Q(x);$$

i.e., the Newton-Goldstein operator in (1.19) *reduces to the Newton operator in* (1.4) *for x near ξ.*

Proof. For $\hat{\xi} \in T_Q(\xi)$, the Lipschitz conditions (1.9) established in Lemma 2.5 yield $\|\hat{\xi} - \xi\| = 0$, and therefore $T_{Q,\delta}(\xi) = T_Q(\xi) = \{\xi\}$. According to Lemma 2.4, ξ is a proper local minimizer of F in Ω, and for convex Ω and F, this means that ξ is the unique *global* minimizer of F over Ω. By Lemma 2.3, it then follows that there are no fixed points of T_Q other than ξ; hence $x \notin T_Q(x)$ for x in Ω and $x \neq \xi$. Fix γ^* in $(0, \gamma)$. Since F is convex, Lemma 2.6 gives

$$\begin{aligned}\langle F'(x), x - \hat{x}\rangle &= Q(x, x) - Q(x, \hat{x}) + \tfrac{1}{2}\langle F''(x)(\hat{x} - x), \hat{x} - x\rangle \\ &\geqq \gamma^*\|\hat{x} - x\|^2 \\ &> 0,\end{aligned}$$

for $\hat{x} \in T_Q(x)$ and all $x \in \Omega$ sufficiently near but not equal to ξ. With Taylor's formula and (2.14) one then gets

$$\begin{aligned}g(x, \hat{x}; 1) &= \frac{F(x) - F(\hat{x})}{\langle F'(x), x - \hat{x}\rangle} \\ &= 1 - \frac{\langle F''(x)(\hat{x} - x), \hat{x} - x\rangle}{2\langle F'(x), x - \hat{x}\rangle} - \frac{\langle (F''(\zeta) - F''(x), (\hat{x} - x), \hat{x} - x\rangle}{2\langle F'(x), x - \hat{x}\rangle} \\ &\geqq \frac{1}{2} - \frac{\|F''(\zeta) - F''(x)\|}{\gamma^*},\end{aligned}$$

with ζ somewhere on the line joining x to $\hat{x}$. It now follows from Lemma 2.5 and the continuity of F'' at ξ that $g(x, \hat{x}; 1) \geqq \delta$, and therefore $T_{Q,\delta}(x) = T_Q(x)$, for all $x \neq \xi$ sufficiently close to ξ in Ω. Q.E.D.

3. Convergence near regular fixed points of T_Q.

The following theorems are straightforward corollaries of Lemmas 2.5 and 2.7. Notice that convexity of F or Ω is not invoked in the first theorem. The principal assumption is that the local quadratic approximation $Q(\xi, y)$ grows rapidly enough as y moves away from the fixed point ξ inside Ω. This regularity assumption is also a basic sufficient condition for strong local minimality (Lemma 2.4).

THEOREM 3.1. *Let Ω be a nonempty subset of a real Banach space X, and let $F: X \to \mathbb{R}^1$ have a second Fréchet derivative F''. Suppose that ξ is a regular fixed point of the operator T_Q in* (1.4), *and that F'' is continuous at ξ. Then for all $x_0 \in \Omega$ sufficiently close to ξ, any sequence of Newton iterates $\{x_n\}$ generated by* (1.7) *converges superlinearly to ξ. Moreover, if ξ is also strongly regular, then $x_n = \xi$ for n sufficiently large.*

Proof. According to Lemma 2.5, there is a $\sigma_0 > 0$ such that $x \in \Omega$, $\|x - \xi\| \leqq \sigma < \sigma_0 \Rightarrow K(\sigma) \leqq K(\sigma_0) < 1$, and $\|\hat{x} - \xi\| \leqq K(\sigma)\|x - \xi\|$ for all $\hat{x} \in T_Q(x)$, where $K(\sigma)$ is defined in (2.9)–(2.10). If $\{x_n\}$ satisfies (1.7) with $\|x_0 - \xi\| < \sigma_0$, then $\|x_n - \xi\| \leqq [K(\sigma_0)]^n \cdot \|x_0 - \xi\|$ for all $n \geqq 0$, by induction, and consequently $\lim \|x_n - \xi\| = 0$. Furthermore, the local Lipschitz conditions in Lemma 2.5 also insure that $T_Q(\xi) = \{\xi\}$; therefore $x_N = \xi \Rightarrow x_n = \xi$ for all $n \geqq N$. On the other hand, if $\|x_n - \xi\| > 0$ for all $n \geqq 0$, one finds that

$$\lim_{n\to 0}\frac{\|x_{n+1}-\xi\|}{\|x_n-\xi\|}=\lim_{n\to 0}K(\|x_n-\xi\|)=0.$$

Finally, if (1.11) holds, then $T_Q(x)=\{\xi\}$ for all $x\in\Omega$ sufficiently close to ξ and therefore $x_n\to\xi\Rightarrow x_n=\xi$ for n sufficiently large. Q.E.D.

Note 3.1. If F'' is Lipschitz continuous, it follows from (2.9)–(2.10) that $K(\|x-\xi\|)\leqq\text{cons.}\,\|x-\xi\|$; the Newton iterates in (1.7) then converge *quadratically* to ξ for x_0 sufficiently close to ξ.

THEOREM 3.2. *Let Ω, X, F, and ξ satisfy the conditions of Theorem* 3.1 *and in addition, let Ω and F be convex. If $\{x_n\}\subset\Omega$ is generated by the Newton-Goldstein algorithm* (1.18) *and if $\{x_n\}$ converges to ξ, then $\{x_n\}$ eventually satisfies the Newton recursion* (1.7) *for sufficiently large n. In this case, $\{x_n\}$ converges superlinearly to ξ; moreover if ξ is a strongly regular fixed point of T_Q, then $x_n=\xi$ for n sufficiently large. Finally, if x_0 is sufficiently close to ξ, $\{x_n\}$ must converge to ξ.*

Proof. Immediate from Lemma 2.7 and Theorem 3.1. Q.E.D.

4. Global convergence theorems for convex functionals. In their analysis of the basic Newton scheme (1.7), Levitin and Polyak [3] assume that Ω is a convex subset of a Hilbert space X and that F is uniformly convex, with F'' Lipschitz continuous on Ω and

$$\lambda\|v\|^2\geqq\langle F''(x)v,v\rangle\geqq\mu\|v\|^2, \tag{4.1}$$

for some λ, $\mu>0$, all $v\in X$, and all $x\in\Omega$. Although these global constraints on F are much stronger than the local regularity condition (1.10) imposed in Theorem 3.1, the corresponding convergence theorem in [3] is still a local theorem; indeed, the example in § 1 shows that no global convergence theorem is possible for (1.7) on the class of functionals treated by Levitin and Polyak (the analysis in [3] does produce a test which can sometimes decide at the outset whether x_0 falls in the domain of attraction of a minimizer of F; however this test is often a rather conservative sufficient condition for convergence, and to apply it one must have values for μ in (4.1) and a Lipschitz constant for F''). On the other hand, for convex F, Goldstein has shown that the relaxed process (1.18) can converge from all remote starting points x_0 under circumstances where (1.7) converges only locally. The following results significantly extend the principal theorem in [2].

THEOREM 4.1. *Let Ω be a nonempty convex subset of a real Banach space X, and let $F:X\to\mathbb{R}^1$ be convex, twice Fréchet differentiable, and bounded below on Ω. For a given $x_0\in\Omega$ suppose that the second derivative F'' is continuous on the corresponding level set Ω_0 in* (1.20) *and also uniformly bounded there; i.e.,*

$$\|F''(x)\|\leqq\lambda, \tag{4.2}$$

for some $\lambda>0$ and all $x\in\Omega_0$. Finally, with δ fixed in $(0,\frac{1}{2})$ let $\{x_n\}\subset\Omega$, $\{\hat{x}_n\}\subset\Omega$, and $\{\omega_n\}\subset[0,1]$ satisfy the Newton-Goldstein conditions (1.16)–(1.17) *and suppose that the defect sequence $\{\hat{x}_n-x_n\}$ is bounded. Then $\{F_n\}$ converges monotonically downward to $l\geqq\inf_\Omega F>-\infty$, and x_n belongs to Ω_0 for all $n\geqq 0$. Furthermore, if $\{x_n\}$ has a limit point ξ in Ω_0 then ξ lies in the set Ω_F of minimizers of F over Ω, and $l=\inf_\Omega F$; in particular, if Ω_0 is compact then $l=\inf_\Omega F$, Ω_F is not empty, and $\{x_n\}$ converges to the set Ω_F; i.e.,*

$$\lim_{n\to\infty}\inf_{x\in\Omega}\|x_n-x\|=0.$$

Proof. Since $Q(x_n, x_n) = 0$, $\hat{x}_n \in T_Q(x_n)$, and F is convex, one has

$$\begin{aligned}\langle F'_n, x_n - \hat{x}_n\rangle &= [Q(x_n, x_n) - Q(x_n, \hat{x}_n)] + \tfrac{1}{2}\langle F''_n(x_n - \hat{x}_n), x_n - \hat{x}_n\rangle \\ &\geqq 0.\end{aligned} \tag{4.3}$$

Consequently, (1.16)–(1.17) yields

$$F_n - F_{n+1} \geqq \omega_n \delta \langle F'_n, x_n - \hat{x}_n\rangle \geqq 0, \tag{4.4}$$

for all $n \geqq 0$. Thus $\{F_n\}$ is monotone nonincreasing and bounded below, and therefore must converge to some limit $l \geqq \inf_\Omega F > -\infty$; it follows that $F_n \leqq F_0$, and hence $x_n \in \Omega_0$, for all $n \geqq 0$.

According to (1.16)–(1.17), (4.3), and Lemma 2.3, $\langle F'_n, x_n - \hat{x}_n\rangle = 0 \Rightarrow x_n \in T_Q(x_n)$, and $\omega_n = 0 \Rightarrow x_{n+1} = x_n \in \Omega_F \Rightarrow \langle F'_{n+1}, x_{n+1} - \hat{x}_{n+1}\rangle = 0$; consequently $\langle F'_N, x_N - \hat{x}_N\rangle = 0 \Rightarrow x_n = x_N \in \Omega_F$ for all $n \geqq N$, by induction. If $\langle F'_n, x_n - \hat{x}_n\rangle \neq 0$ for all $n \geqq 0$, one either has $\omega_n = 1$ and

$$F_n - F_{n+1} \geqq \delta\langle F'_n, x_n - \hat{x}_n\rangle > 0, \tag{4.5}$$

or else

$$1 - \delta \geqq \frac{F_n - F_{n+1}}{\omega_n\langle F'_n, x_n - \hat{x}_n\rangle} \geqq \delta,$$

in view of (1.16)–(1.17). In the latter case, Taylor's formula yields,

$$\begin{aligned}1 - \delta &\geqq \frac{F_n - F_{n+1}}{\omega_n\langle F'_n, x_n - \hat{x}_n\rangle} \\ &= 1 - \frac{\omega_n}{2}\,\frac{\langle F''(\zeta)(x_n - \hat{x}_n), x_n - \hat{x}_n\rangle}{\langle F'_n, x_n - \hat{x}_n\rangle},\end{aligned}$$

where ζ is somewhere on the line segment joining x_n to x_{n+1} in the convex set Ω_0. This inequality and (4.2) now produce,

$$\omega_n \geqq \frac{2\delta\langle F'_n, x_n - \hat{x}_n\rangle}{\lambda\|x_n - \hat{x}_n\|^2}, \tag{4.6}$$

and consequently,

$$\begin{aligned}F_n - F_{n+1} &\geqq \frac{2\delta^2\langle F'_n, x_n - \hat{x}_n\rangle^2}{\lambda\|x_n - \hat{x}_n\|^2} \\ &\geqq \frac{2\delta^2}{\lambda d^2}\cdot\langle F'_n, x_n - \hat{x}_n\rangle^2,\end{aligned} \tag{4.7}$$

where d is an upper bound on the defect norms $\|x_n - \hat{x}_n\|$. Together, (4.5) and (4.7) give

$$F_n - F_{n+1} \geqq \min\{\delta\langle F'_n, x_n - \hat{x}_n\rangle, \frac{2\delta^2}{\lambda d^2}\langle F'_n, x_n - \hat{x}_n\rangle^2\}. \tag{4.8}$$

Since $F_n \to l \Rightarrow F_n - F_{n+1} \to 0$, it now follows that

$$\lim_{n\to\infty}\langle F'_n, x_n - \hat{x}_n\rangle = 0. \tag{4.9}$$

In view of (4.3) this yields

$$\lim_{n\to\infty}\psi(x_n) = 0 \tag{4.10a}$$

where

$$\psi(x) = \sup_{y\in\Omega}(-Q(x, y)) \tag{4.10b}$$
$$= Q(x, x) - \inf_{y\in\Omega} Q(x, y) \geqq 0.$$

For each fixed $y \in \Omega$, $Q(\cdot, y)$ is continuous on Ω_0 because $F''(\cdot)$ is continuous on Ω_0. Therefore by Lemma 2.1, the nonnegative function ψ is lower semicontinuous, and it follows from (4.10) and Lemma 2.3 that $x_{n_k} \to \xi \in \Omega_0 \Rightarrow \psi(\xi) = 0 \Rightarrow \xi \in \Omega_F$. Furthermore, by the continuity of F, $x_{n_k} \to \xi \in \Omega_F \Rightarrow F_{n_k} \to F(\xi) = \inf_\Omega F$. If Ω_0 is compact, then $\{x_n\}$ necessarily has limit points ξ in $\Omega_F \neq \emptyset$, and so $l = \inf_\Omega F$. Finally, if $\{x_n\}$ does not converge to Ω_F one obtains the contradiction, $\inf_{x\in\Omega_F} \|x - \xi\| \geqq \varepsilon$ for some $\varepsilon > 0$ and some limit point ξ of $\{x_n\}$. Q.E.D.

THEOREM 4.2. *Let Ω, Ω_0 and F satisfy the conditions of Theorem* 4.1, *and suppose that the local quadratic approximations Q obey the uniform growth condition* (1.22) *on Ω_0. For fixed δ in $(0, \frac{1}{2})$, let $\{x_n\} \subset \Omega$, $\{\hat{x}_n\} \subset \Omega$ and $\{\omega_n\} \subset [0, 1]$ satisfy the Newton-Goldstein conditions* (1.16)–(1.17). *Then F has at most one minimizer ξ in Ω, $\{F_n\}$ converges monotonically downward to $l \geqq \inf_\Omega F > -\infty$, $x_n \in \Omega_0$ for all $n \geqq 0$, and* $\lim_{n\to\infty} \|x_n - \hat{x}_n\| = 0$. *Furthermore, if $\{x_n\}$ is bounded, then $l = \inf_\Omega F$ and every weak limit point of $\{x_n\}$ coincides with ξ.*

Proof. If $\xi \in \Omega_F$, then $\xi \in T_Q(\xi)$, by Lemma 2.3. Furthermore (1.22) $\Rightarrow$ (1.10) for all $y \in \Omega$; consequently ξ is a regular fixed point of T_Q. According to Lemma 2.4, ξ is then a proper local minimizer of F in Ω, and for convex F and Ω this means that ξ is the unique global minimizer of F over Ω.

Inequalities (4.2), and (4.3) through (4.6), remain valid under the present hypotheses; therefore $\{F_n\}$ converges monotonically downward to some limit $l \geqq \inf_\Omega F > -\infty$, and so $x_n \in \Omega_0$ for all $n \geqq 0$. Furthermore, as in the proof of Theorem 4.1, one finds that either $x_n = x_N = \xi$ for all n beyond some N, or else $\langle F'_n, x_n - \hat{x}_n\rangle > 0$ for all $n \geqq 0$. In the first case, it follows from (1.10) that $\hat{x}_n = \xi$ and therefore $\hat{x}_n - x_n = 0$ for $n \geqq N$; in the second case,

$$F_n - F_{n+1} \geqq \min\left\{\delta, \frac{2\delta^2}{\lambda}\frac{\langle F'_n, x_n - \hat{x}_n\rangle}{\|x_n - \hat{x}_n\|^2}\right\} \cdot \langle F'_n, x_n - \hat{x}_n\rangle. \tag{4.11}$$

For convex F, the second term on the right in (4.3) is nonnegative, and so the growth condition (1.22) gives

$$\langle F'_n, x_n - \hat{x}_n\rangle \geqq \gamma\|x_n - \hat{x}_n\|^2. \tag{4.12}$$

With (4.11) and (4.12) one now gets

$$F_n - F_{n+1} \geqq \min\left\{\delta, \frac{2\delta^2\gamma}{\lambda}\right\} \cdot \langle F'_n, x_n - \hat{x}_n\rangle \tag{4.13}$$
$$\geqq \min\left\{\delta\gamma, \frac{2\delta^2\gamma^2}{\lambda}\right\} \cdot \|\hat{x}_n - x_n\|^2,$$

for all $n \geqq 0$, and consequently,

$$\lim_{n\to\infty} \langle F'_n, x_n - \hat{x}_n\rangle = \lim_{n\to\infty} \|x_n - \hat{x}_n\|^2 = 0, \tag{4.14}$$

since $F_n \to l \Rightarrow F_n - F_{n+1} \to 0$.

For convex F, one also has

$$\begin{aligned} F(y) &\geqq F_n + \langle F'_n, y - x_n\rangle \\ &= F_n + \langle F'_n, y - \hat{x}_n\rangle + \langle F'_n, \hat{x}_n - x_n\rangle, \end{aligned}$$

for all $n \geqq 0$ and $y \in X$; therefore it follows from (2.14) and (4.2) that

$$\begin{aligned} F(y) &\geqq F_n - \langle F''_n(\hat{x}_n - x_n), y - \hat{x}_n\rangle - \langle F'_n, x_n - \hat{x}_n\rangle \\ &\geqq F_n - \lambda\|\hat{x}_n - x_n\|\,\|y - \hat{x}\| - \langle F'_n, x_n - \hat{x}_n\rangle, \end{aligned} \tag{4.15}$$

for all $y \in \Omega$. But in view of (4.14), $\{\hat{x}_n\}$ is bounded because $\{x_n\}$ is bounded; hence (4.14) and (4.15) give $F(y) \geqq \lim_{n\to\infty} F_n = l \geqq \inf_\Omega$ for all $y \in \Omega$, in which case $l = \inf_\Omega F$.

Finally, a continuous convex functional F is also weakly lower semicontinuous; therefore $F_n \to \inf_\Omega F \Rightarrow$ all weak limit points of $\{x_n\}$ fall in the minimizer set $\Omega_F = \{\xi\}$. Q.E.D.

Note 4.1. For compact Ω_0, (4.2) automatically follows from the continuity of F''. For continuous F, the level set

$$S_0 = \{x \in X \mid F(x) \leqq F(x_0)\} \tag{4.16}$$

is closed; consequently if Ω is compact and F'' is continuous, then $\Omega_0 (= \Omega \cap S_0)$ is compact, (4.2) holds, and $\{\hat{x}_n - x_n\}$ is bounded. If S_0 is compact and Ω is closed but not bounded, then Ω_0 is again compact and (4.2) holds, but $\{\hat{x}_n - x_n\}$ is not necessarily bounded. Nevertheless, the latter condition can be verified for many interesting problems on unbounded closed Ω's. As a simple illustration, consider $F(x) = x^4$ on $\Omega = X = \mathbb{R}^1$. The unique minimizer for F, viz. $\xi = 0$, is "singular to second order," in the sense that both the local linear approximation *and* the local quadratic approximation to F at ξ have multiple minimizers (in fact, both functions *vanish identically*); consequently Goldstein's Theorem [2] and Theorem 4.2 of the present article, are inapplicable. On the other hand, Theorem 4.1 can be applied here. The level sets S_0 for F are compact, and at each $x \neq 0$, $T_Q(x)$ is single-valued with

$$T_Q(x) = \{\hat{x}\} = \left\{x - \frac{F'(x)}{F''(x)}\right\} = \{\tfrac{2}{3}x\}.$$

At the minimizer $\xi = 0$, Q vanishes identically hence $T_Q(0) = X = \mathbb{R}^1$, and therefore the single-valued branches of T_Q are defined by

$$\hat{x} = \begin{cases} \tfrac{2}{3}x, & x \neq 0, \\ y, & x = 0, \end{cases} \tag{4.17}$$

with y an arbitrary real number. Evidently, each of these branches is bounded on bounded sets. Moreover, in the first part of the proof of Theorem 4.1, it is shown that sequences $\{x_n\}$ generated by (1.16)–(1.17) remain in Ω_0 irrespective of whether $\{\hat{x}_n - x_n\}$ is bounded. But in the present case, $\{x_n\} \subset \Omega_0 \Rightarrow \{x_n\}$ is bounded $\Rightarrow \{\hat{x}_n\}$ is bounded; consequently $\{x_n\}$ must converge to ξ from arbitrary x_0, according to Theorem 4.1. (Notice that sequences $\{x_n\}$ generated by the basic Newton scheme (1.7) also converge to 0 from arbitrary x_0; however, because of the singularity in $F''(x)^{-1}$ at x_0, the convergence rate is now merely *linear* with ratio $\frac{2}{3}$). A similar development is possible for a large class of functions $F: \mathbb{R}^N \to \mathbb{R}^1$ with singular critical points; the basic requirement here is that the minimum norm solution $V^\dagger(x)$ of $F'(x) + F''(x)v = 0$ should remain bounded as x ranges over any bounded set. In another recent investigation, Reddien [20] characterizes the local behavior of classical Newton iterates near a certain type of singular zero.

Note 4.2. For all $x, y \in X$ and $\hat{x} \in T_Q(x)$ one has

$$Q(x, y) - Q(x, \hat{x}) = \langle Q'_y(x, \hat{x}), y - \hat{x} \rangle + \tfrac{1}{2} \langle F''(x)(y - \hat{x}), y - \hat{x} \rangle. \tag{4.18}$$

If $\Omega = X$, then $Q'_y(x, \hat{x}) = 0$ and the growth condition (1.22) is equivalent to the right side of (4.1). More generally, if Ω is an arbitrary convex subset of X, it is still true that

$$\langle Q'_y(x, \hat{x}), y - \hat{x} \rangle \geqq 0, \tag{4.19}$$

for all $y \in \Omega$; consequently (4.1) $\Rightarrow$ (1.22). On the other hand, (1.22) can hold even when F is not convex. Thus, if Ω satisfies the uniform convexity condition (1.14) and if the derivative norms $\|Q'_y(x, \hat{x})\|$ are uniformly bounded away from zero on Ω, i.e., if for some $\varepsilon > 0$,

$$\|Q'_y(x, \hat{x})\| \geqq \varepsilon, \tag{4.20}$$

for all $x \in \Omega$, $\hat{x} \in T_Q(x)$, then it can be shown that

$$\langle Q'_y(x, \hat{x}), y - \hat{x} \rangle \geqq 2\nu\varepsilon \|y - \hat{x}\|^2,$$

for all $x, y \in \Omega$, $\hat{x} \in T_Q(x)$ (see the proof of Theorem 3.4 in [7]). Therefore, for any convex F (1.14) and (4.20) $\Rightarrow$ (1.22), since the second term on the right in (4.18) is nonnegative; in fact (1.22) will still follow even for a nonconvex F, provided

$$2\nu\varepsilon \geqq \inf_{\substack{y \in \Omega \\ y \neq \hat{x}}} \frac{\langle F''(x)(y - \hat{x}), y - \hat{x} \rangle}{\|y - \hat{x}\|^2},$$

for all $x \in \Omega_0$, $\hat{x} \in T_Q(x)$.

Note 4.3. Let F be convex and let ξ be a regular extremal of F in Ω, i.e., (1.12) holds at ξ. By Lemma 5.2 in [7], it follows that ξ is the unique minimizer of F in Ω, and that every minimizing sequence converges to ξ, i.e., $F(x_n) \to \inf_\Omega F \Rightarrow x_n \to \xi$. Moreover, if Ω satisfies the uniform convexity condition (1.14) and if $\|Q'_y(\xi, \xi)\| = \|F'(\xi)\| \neq 0$, then ξ is automatically regular (Theorem 3.4, [7]).

Note 4.4. Let X be reflexive and let Ω be closed and convex. Furthermore, let F satisfy the right side of (4.1). Then it can be shown that the level set Ω_0 is bounded, closed and convex, and therefore weakly compact. It now follows from Theorems 4.1 and 4.2, and [18, Chapt. 1], that F has a unique minimizer ξ and every minimizing sequence converges to ξ.

Note 4.5. For convex F, Lemma 2.3 asserts that $\xi \in T_Q(\xi)$; in this case, (1.22) $\Rightarrow$ (1.10).

In light of these observations, one can now see that Theorems 3.1, 3.2, 4.1, and 4.2 contain and substantially extend the principal theorem of Goldstein in [2] for $\Omega = X =$ a Hilbert space.

REFERENCES

[1] L. V. Kantorovich, *Functional analysis and applied mathematics*, Uspehi Mat. Nauk., 3 (1948), pp. 89–185.

[2] A. A. Goldstein, *On Newton's method*, Numer. Math., 7 (1965), pp. 391–393.

[3] E. S. Levitin and B. T. Polyak, *Constrained minimization methods*, U.S.S.R. Computational Math. and Math. Phys., 6 (1966), pp. 1–50.

[4] B. N. Pshenichnyi, *Newton's method for the solution of systems of equalities and inequalities*, English translation in Math. Notes, 8 (1970), pp. 827–830.

[5] S. M. Robinson, *Extensions of Newton's method to nonlinear functions with values in a cone*, Numer. Math., 19 (1972), pp. 341–347.

[6] ———, *Perturbed Kuhn-Tucker points and rates of convergence for a class of nonlinear programming algorithms*, Math. Programming, 7 (1974), pp. 1–16.

[7] J. C. DUNN, *Rates of convergence for conditional gradient algorithms near singular and nonsingular extremals*, this Journal, 17 (1979), pp. 187–211.

[8] ——, *Convergence rates for conditional gradient sequences generated by implicit step length rules*, this Journal, 18 (1980), pp. 473–487.

[9] D. G. LUENBERGER, *Optimization by Vector Space Methods*, Academic Press, New York, 1962.

[10] A. A. GOLDSTEIN, *On steepest descent*, SIAM J. Control Ser. A, 3 (1965), pp. 147–151.

[11] L. ARMIJO, *Minimization of functions having Lipschitz continuous first partial derivatives*, Pacific J. Math., 16 (1966), pp. 1–3.

[12] E. POLAK, *Computational Methods in Optimization: A Unified Approach*, Academic Press, New York, 1971.

[13] G. B. DANTZIG, *Linear Programming and Extensions*, Princeton University Press, Princeton, NJ, 1963.

[14] G. B. DANTZIG, J. FOLKMAN AND N. SHAPIRO, *On the continuity of the minimum set of a continuous function*, J. Math. Anal. Appl., 17 (1967), pp. 519–548.

[15] A. E. TAYLOR AND W. R. MANN, *Advanced Calculus*, 2nd edition, Xerox College Publishing, Lexington, MA, 1972.

[16] L. M. GRAVES, *Riemann integration and Taylor's theorem in general analysis*, Trans. Amer. Math. Soc., January 1927, pp. 163–177.

[17] L. M. GRAVES AND T. H. HILDERBRANDT, *Implicit functions and their differentials in general analysis*, Trans. Amer. Math. Soc., January 1927, pp. 127–153.

[18] V. F. DEMYANOV AND A. M. RUBINOV, *Approximate Methods in Optimization Problems*, American Elsevier, New York, 1970.

[19] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[20] G. W. REDDIEN, *On Newton's method for singular problems*, SIAM J. Numer. Anal., 15 (1978), pp. 993–996.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 6, November 1980

0363-0129/80/1806-0008/$01.00/0

FACTORIZATIONS OF TRANSFER FUNCTIONS*

H. BART†, I. GOHBERG‡, M. A. KAASHOEK† AND P. VAN DOOREN§

Abstract. This paper is concerned with minimal factorizations of rational matrix functions. The treatment is based on a new geometrical principle. In fact, it is shown that there is a one-to-one correspondence between minimal factorizations on the one hand and certain projections on the other. Considerable attention is given to the problem of stability of a minimal factorization. Also the numerical aspects are discussed. Along the way, a stability theorem for solutions of the matrix Riccati equation is obtained.

Introduction. The problem of factorizing a rational matrix-valued function $W(\lambda)$ into "simpler" rational factors has network theory as one of its origins. In this theory $W(\lambda)$ appears as a transfer function of a network. Its minimal factorizations (see Chapter II) are of particular interest because it allows one to obtain the network by a cascade connection of elementary sections which have the simplest synthesis [6], [22].

In the present paper the treatment of the factorization problem is based on a new geometrical principle. This principle has been observed independently by the first three authors and by the fourth (and has been communicated at a miniconference on Operators and System Theory held at Amsterdam and Delft, February, 1978). For the fourth author network theory [22], [23] has been the main motivation, while the first three authors were inspired by [3], [7], [20].

The new geometrical principle referred to allows for a unifying approach to seemingly disjoint topics such as the network problems mentioned above, the matrix Riccati equation [19], the factorization theory of characteristic functions for linear operators [7], the theory of Wiener-Hopf (or spectral) factorization [10], [11] and the divisibility theory of operator polynomials [3], [12], [13]. Here we treat only the first two topics; the other connections will be investigated in detail in a forthcoming publication [5].

The problem of computing numerically the minimal factors of a transfer function led us to investigate the stability of divisors under small perturbations. We pay considerable attention to the measure of stability.

The matrix functions studied here are viewed as transfer functions of systems. A system consists of three matrices A, B, and C, of appropriate sizes, and the corresponding transfer functions are of the form

$$W(\lambda)=I+C(\lambda I-A)^{-1}B,$$

where λ is the complex variable and I the identity matrix. In the first chapter multiplication and division of transfer functions are described in terms of systems. Applications to matrix Riccati equations are also considered here. The special type of minimal factorization and its properties are studied in Chapter II. In geometrical terms an explicit description of all minimal factors is given. Stability and numerical aspects are studied in the last two chapters. Throughout the paper we confine ourselves to the finite dimensional case, but with minor modifications the results of Chapters I and III are also valid in the infinite dimensional situation (see [5]).

* Received by the editors June 22, 1979, and in final revised form March 10, 1980.

† Wiskundig Seminarium, Vrije Universiteit, Amsterdam, The Netherlands.

‡ Department of Mathematical Sciences, Tel-Aviv University, Ramat Aviv and Department of Pure Mathematics, Weizmann Institute of Science, Rehovot, Israel.

§ Division of Applied Mathematics and Computer Science, Katholieke Universiteit, Leuven, Belgium.

As far as notation and terminology is concerned we stipulate the following. The term *linear space* stands for a complex vector space. All linear spaces appearing below are assumed to be finite dimensional. In Chapters III and IV it is also assumed that they are endowed with a norm, which is always denoted by $\|\cdot\|$. By an *operator* we mean a linear transformation between two linear spaces. The null space and range of an operator T are denoted by Ker T and Im T, respectively. The identity operator on a linear space X is always denoted by I. The symbol I_n is used for the $n \times n$ identity matrix. Whenever this is convenient, an $m \times n$ matrix A will be identified with the operator from $\mathbb{C}^n$ into $\mathbb{C}^m$ given by the canonical action of A with respect to the standard bases in $\mathbb{C}^n$ and $\mathbb{C}^m$. In particular a rational $n \times n$ matrix function may be viewed as a rational function whose values are operators acting on $\mathbb{C}^n$.

I. Divisibility of transfer functions and the Riccati equation. In this chapter multiplication and division of transfer functions are described in terms of systems. The main result on factorization is presented in § 1.1. A slightly more sophisticated factorization theorem, involving the notion of an angular operator, is given in § 1.2. In § 1.3 we discuss the operator Riccati equation.

1.1. Multiplication and divisibility of systems. A *system* is a quintet $\theta = (A, B, C; X, Y)$ of two linear spaces X, Y and three operators $A: X \rightarrow X$, $B: Y \rightarrow X$ and $C: X \rightarrow Y$. The space X is called the *state space*; the space Y is called the *input/output space*. The operator A is referred to as the *state space* or *main operator.* A common way to give systems is to specify three matrices of appropriate sizes. To be more specific, if A is a $\delta \times \delta$ matrix, B is a $\delta \times n$ matrix and C is an $n \times \delta$ matrix, then (identifying A, B, and C in the usual way with operators) the quintet $(A, B, C; \mathbb{C}^\delta, \mathbb{C}^n)$ is a system.

Two systems $\theta_1 = (A_1, B_1, C_1; X_1, Y)$ and $\theta_2 = (A_2, B_2, C_2; X_2, Y)$ are said to be *similar*, written $\theta_1 \simeq \theta_2$, if there exists an invertible operator $S: X_1 \rightarrow X_2$, called a *system similarity*, between θ_1 and θ_2 such that

$$A_1 = S^{-1}A_2S, \quad B_1 = S^{-1}B_2, \quad C_1 = C_2S.$$

The relation $\simeq$ is reflexive, symmetric and transitive.

Let $\theta = (A, B, C; X, Y)$ be a system, and put

$$W(\lambda) = I + C(\lambda I - A)^{-1}B. \tag{1.1}$$

Then $W(\lambda)$ is a rational operator function and $W(\infty) = I$. This function is called the *transfer function* of θ, and is denoted by W_θ. Obviously, similar systems have the same transfer function.

If $W(\lambda)$ is any rational function whose values are operators acting on Y and $W(\infty) = I$, then it is known from system theory (cf. [2]) that $W(\lambda)$ can be represented in the form (1.1). Such a representation is called a *realization* for $W(\lambda)$; we also use this term for the system $(A, B, C; X, Y)$.

Our terminology is taken from system theory, where the transfer function (1.1) is used to describe the input/output behavior of the linear dynamical system

$$\dot{x}(t) = Ax(t) + Bu(t), \qquad y(t) = Cx(t) + u(t).$$

In the theory of characteristic operator functions, certain systems with special properties are called *nodes* (see, for instance, [7]). The connections with this theory are further developed in [5]. In the next paragraph we shall define the product of two systems. The definition is motivated by the notion of a series connection of two linear dynamical systems. For details, the reader is referred to [18] (cf. also [7]).

Let $\theta_1 = (A_1, B_1, C_1; X_1, Y)$ and $\theta_2 = (A_2, B_2, C_2; X_2, Y)$ be systems. Put $X = X_1 \oplus X_2$, and

$$A = \begin{bmatrix} A_1 & B_1C_2 \\ 0 & A_2 \end{bmatrix}, \quad B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}, \quad C = [C_1 \quad C_2].$$

Then $(A, B, C; X, Y)$ is a system. It is called the *product* of θ_1 and θ_2 and denoted by $\theta_1\theta_2$. A straightforward calculation shows that

$$W_{\theta_1\theta_2}(\lambda) = W_{\theta_1}(\lambda) W_{\theta_2}(\lambda). \tag{1.2}$$

So if θ_1 and θ_2 are realizations for $W_1(\lambda)$ and $W_2(\lambda)$, respectively, then $\theta_1\theta_2$ is a realization for $W_1(\lambda)W_2(\lambda)$.

If $\theta = (A, B, C; X, Y)$ is a realization for the rational operator function $W(\lambda)$, then

$$\theta^x = (A - BC, B, -C; X, Y)$$

is a realization for $W(\lambda)^{-1}$. We call θ^x the *associate system* of θ. The operator $A - BC$ is called the *associate* (*main*) *operator* of θ. By abuse of notation, we write $A^x = A - BC$. Note that A^x depends not only on A, but also on the other operators appearing in the system θ. One checks without difficulty that $(\theta^x)^x = \theta$ (so in particular $(A^x)^x = A$) and $(\theta_1\theta_2)^x \simeq \theta_2^x\theta_1^x$, the natural identification of $X_1 \oplus X_2$ and $X_2 \oplus X_1$ being a system similarity.

Consider the system $\theta = (A, B, C; X, Y)$ and let Π be a projection of X. So Π is an idempotent operator on X. With respect to the decomposition $X = \operatorname{Ker} \Pi \oplus \operatorname{Im} \Pi$, we write

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}, \quad C = [C_1 \quad C_2]. \tag{1.3}$$

The system $\operatorname{pr}_\Pi(\theta) = (A_{22}, B_2, C_2; \operatorname{Im} \Pi, Y)$ is called the *projection* of θ associated with Π (cf. [7]). Observe that $\operatorname{pr}_{I-\Pi}(\theta) = (A_{11}, B_1, C_1; \operatorname{Ker} \Pi, Y)$. One easily verifies that $\operatorname{pr}_\Pi(\theta)^x = \operatorname{pr}_\Pi(\theta^x)$. The projection Π is said to be a *supporting projection* for θ if

$$A[\operatorname{Ker} \Pi] \subset \operatorname{Ker} \Pi, \qquad A^x[\operatorname{Im} \Pi] \subset \operatorname{Im} \Pi. \tag{1.4}$$

If Π is a supporting projection for θ, then $I - \Pi$ is one for θ^x. The second part of (1.4) is equivalent to the rank condition.

$$\operatorname{rank} \begin{bmatrix} A_{12} & B_1 \\ C_2 & I \end{bmatrix} = \dim Y. \tag{1.5}$$

This is immediate from the fact that the left-hand side of (1.5) is equal to rank $(A_{12} - B_1C_2) + \dim Y$.

The following theorem admits a very simple proof. Nevertheless it is one of the cornerstones for the rest of the present paper. A somewhat more sophisticated factorization theorem will be presented in § 1.2.

THEOREM 1.1. *Let Π be a supporting projection for the system $\theta = (A, B, C; X, Y)$. Then*

$$\theta = \operatorname{pr}_{I-\Pi}(\theta) \cdot \operatorname{pr}_\Pi(\theta). \tag{1.6}$$

If $W(\lambda)$, $W_1(\lambda)$ and $W_2(\lambda)$ are the transfer functions of θ, $\operatorname{pr}_{I-\Pi}(\theta)$ and $\operatorname{pr}_\Pi(\theta)$, respectively, then $W(\lambda) = W_1(\lambda)W_2(\lambda)$. In other words,

$$I + C(\lambda I - A)^{-1}B = [I + C(\lambda I - A)^{-1}(I - \Pi)B][I + C\Pi(\lambda I - A)^{-1}B].$$

Proof. With respect to the decomposition $X = \operatorname{Ker}\Pi \oplus \operatorname{Im}\Pi$, we write the operators A, B, and C as in (1.3). Then $A^\times$ may be written as

$$A^\times = \begin{bmatrix} A_{11} - B_1C_1 & A_{12} - B_1C_2 \\ A_{21} - B_2C_1 & A_{22} - B_2C_2 \end{bmatrix}.$$

Hence (1.4) is equivalent to $A_{21} = 0$ and $A_{12} - B_1C_2 = 0$. It follows that

$$A = \begin{bmatrix} A_{11} & B_1C_2 \\ 0 & A_{22} \end{bmatrix}.$$

But then (1.6) is clear from the definition of the product of two systems. The second part of the theorem is now an immediate consequence of formula (1.2).

In a certain sense Theorem 1.1 gives a complete description of all possible factorizations of the system θ. Indeed, if $\theta \simeq \theta_1\theta_2$ for some systems θ_1 and θ_2, then there exists a supporting projection Π for θ such that $\theta_1 \simeq \operatorname{pr}_{I-\Pi}(\theta)$ and $\theta_2 \simeq \operatorname{pr}_{\Pi}(\theta)$.

1.2. Angular operators and the division theorem. Throughout this section, X is a linear space and Π is a projection of X onto X_2 along X_1. (Block) matrix representations of operators acting on X will always be taken with respect to the decomposition $X = X_1 \oplus X_2$.

A subspace N of X is called *angular* with respect to Π if $X = \operatorname{Ker}\Pi \oplus N$. If R is an operator from X_2 into X_1, then the space

$$N_R = \{Rx + x \mid x \in X_2\}$$

is angular with respect to Π. The next proposition shows that every angular subspace is of this form.

PROPOSITION 1.2. *Let N be a subspace of X. Then N is angular with respect to* Π *if and only if $N = N_R$ for some operator R from X_2 into X_1.*

Proof. We have already observed that if $N = N_R$, then N is angular with respect to Π. To prove the converse, assume that N is angular with respect to Π, and let Q be the projection of X onto N along X_1. Put $Rx = (Q - \Pi)x$ for $x \in X_2$. Then $N = N_R$.

Given an angular subspace N, the operator R for which $N = N_R$ is uniquely determined. It is called the *angular operator* for N with respect to Π. This notion was introduced by M. G. Krein in [17]. We are now in a position to bring the division theorem for systems into a slightly more general form.

THEOREM 1.3. *Let $\theta = (A, B, C; X, Y)$ be a system, let* Π *be a projection of X onto X_2 along X_1, and let N be an angular subspace of X with respect to* Π. *Assume that*

$$A[X_1] \subset X_1, \qquad A^\times[N] \subset N. \tag{1.7}$$

Further, let

$$A = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}, \quad B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}, \quad C = [C_1 \quad C_2],$$

be the matrix representations of A, B and C with respect to the decomposition $X = X_1 \oplus X_2$, let R be the angular operator for N with respect to Π, *and put*

$$\theta_1 = (A_{11}, B_1 - RB_2, C_1; X_1, Y), \tag{1.8}$$

$$\theta_2 = (A_{22}, B_2, C_1R + C_2; X_2, Y). \tag{1.9}$$

Then $\theta \simeq \theta_1\theta_2$. More precisely,

$$\theta_1\theta_2 = (E^{-1}AE, E^{-1}B, CE; X, Y),$$

where E is the invertible operator

$$E=\begin{bmatrix} I & R \\ 0 & I \end{bmatrix}.$$

Proof. For convenience, put $\hat{A}=E^{-1}AE$, $\hat{B}=E^{-1}B$, $\hat{C}=CE$ and $\hat{\theta}=(\hat{A},\hat{B},\hat{C};X,Y)$. Observe that $\hat{A}^x=E^{-1}A^xE$. Now E maps X_1 onto X_1 and X_2 onto N. Thus (1.7) implies that

$$\hat{A}[X_1]\subset X_1, \qquad \hat{A}^x[X_2]\subset X_2.$$

Apply now Theorem 1.1 to show that

$$\hat{\theta}=\mathrm{pr}_{I-\Pi}(\hat{\theta})\cdot \mathrm{pr}_{\Pi}(\hat{\theta}).$$

But $\mathrm{pr}_{I-\Pi}(\hat{\theta})=\theta_1$ and $\mathrm{pr}_{\Pi}(\hat{\theta})=\theta_2$, and the proof is complete.

Suppose that the angular subspace N in Theorem 1.3 is the image of X_2 under some invertible operator

$$S=\begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}: X_1\oplus X_2\to X_1\oplus X_2.$$

Then it is not difficult to prove that S_{22} is invertible. Moreover the angular operator R for N is given by $R=S_{12}S_{22}^{-1}$. By substituting this in (1.8) and (1.9), we get

$$\theta_1=(A_{11}, B_1-S_{12}S_{22}^{-1}B_2, C_1; X_1, Y),$$

$$\theta_2=(A_{22}, B_1, C_1S_{12}S_{22}^{-1}+C_2; X_2, Y).$$

This together with formula (1.2), can be used to give a quick proof of Theorem 4 in Sahnovič's paper [20].

1.3. The Riccati equation. As in the previous section, X is a linear space and Π is a projection of X onto X_2 along X_1. In view of Theorem 1.3 the following question is of interest. Given an angular subspace N of X and an operator T on X, when is N invariant under T? The next proposition shows that the answer involves an operator Riccati equation.

PROPOSITION 1.4. *Let N be an angular subspace of X with respect to* Π, *and let*

$$T=\begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix}: X_1\oplus X_2\to X_1\oplus X_2$$

be an operator on X. Then N is invariant under T if and only if the angular operator R for N satisfies the Riccati equation

$$RT_{21}R+RT_{22}-T_{11}R-T_{12}=0. \tag{1.10}$$

Proof. The operator

$$E=\begin{bmatrix} I & R \\ 0 & I \end{bmatrix}: X_1\oplus X_2\to X_1\oplus X_2$$

is invertible and maps X_2 onto N. So

$$E^{-1}TE=\begin{bmatrix} T_{11}-RT_{21} & -RT_{21}R-RT_{22}+T_{11}R+T_{12} \\ T_{21} & T_{22}+T_{21}R \end{bmatrix}$$

leaves invariant X_2 if and only if T leaves invariant N. But $E^{-1}TE$ leaves invariant X_2 if and only if (1.10) is satisfied, and the proof is complete.

In view of formula (1.2) and Theorem 1.3, the problem of finding factorizations for transfer functions of systems is related to that of solving a certain Riccati operator equation. As a matter of fact, the condition $A^{\times}[N]\subset N$ is equivalent to the requirement

$$RB_2C_1R+R(B_2C_2-A_{22})+(A_{11}-B_1C_1)R+A_{12}-B_1C_2=0.$$

Here we use the notation of Theorem 1.3.

Now let us introduce some more notation and terminology. Let T be an operator on X and let μ be an eigenvalue of T. The subspace $\operatorname{Ker}(\mu I-T)^m$, where m is the dimension of X, is called the *generalized eigenspace* of T corresponding to μ. If $\lambda_1,\cdots,\lambda_r$ are eigenvalues of T, the space

$$\operatorname{Ker}(\lambda_1 I-T)^m\oplus\cdots\oplus\operatorname{Ker}(\lambda_r I-T)^m \tag{1.11}$$

is called the *spectral subspace* for T corresponding to the eigenvalues $\lambda_1,\cdots,\lambda_r$. This spectral subspace can also be described as follows. Let Γ be a contour in $\mathbb{C}$ such that $\lambda_1,\cdots,\lambda_r$ are inside and the remaining eigenvalues of T are outside Γ. Put

$$P(T;\Gamma)=\frac{1}{2\pi i}\int_\Gamma(\lambda I-T)^{-1}\,d\lambda.$$

Then the spectral subspace (1.11) coincides with the image of $P(T;\Gamma)$. In view of this, (1.11) is also called the spectral subspace for T corresponding to Γ. The operator $P(T;\Gamma)$ is a projection of X, called the *Riesz projection* corresponding to T and Γ (or $\lambda_1,\cdots,\lambda_r$).

PROPOSITION 1.5. *Let N be an angular subspace of X with respect to* Π, *and let*

$$T=\begin{bmatrix}T_{11} & T_{12}\\ T_{21} & T_{22}\end{bmatrix}:X_1\oplus X_2\to X_1\oplus X_2,$$

be an operator on X. Then N is a spectral subspace for T if and only if the angular operator R for N satisfies the Riccati equation (1.10) *and the operators $T_{11}-RT_{21}$ and $T_{22}+T_{21}R$ have no common eigenvalues.*

It will appear from the proof that if N is the spectral subspace for T corresponding to the contour Γ, then the eigenvalues of $T_{22}+T_{21}R$ are inside Γ and the eigenvalues of $T_{11}-RT_{21}$ are outside Γ.

Proof. Define E as in the proof of Proposition 1.4. It is clear that N is a spectral subspace for T (corresponding to a contour Γ) if and only if X_2 is a spectral subspace (corresponding to the same contour Γ) for $S=E^{-1}TE$. With respect to the decomposition $X=X_1\oplus X_2$, we write

$$S=\begin{bmatrix}S_{11} & S_{12}\\ S_{21} & S_{22}\end{bmatrix}.$$

Recall that $S_{11}=T_{11}-RT_{21}$, $S_{12}=-RT_{21}R-RT_{22}+T_{11}R+T_{12}$, $S_{21}=T_{21}$, and $S_{22}=T_{22}+T_{21}R$.

Now suppose that $S_{12}=0$ and that S_{11} and S_{22} have no common eigenvalues. Let Γ be a Cauchy contour such that the eigenvalues of S_{11} are outside and the eigenvalues of S_{22} are inside Γ. Then $P(S;\Gamma)$ has the form

$$P(S;\Gamma)=\begin{bmatrix}0 & 0\\ * & I\end{bmatrix},$$

and it follows that $X_2=\operatorname{Im}P(S;\Gamma)$.

Next assume that X_2 is the spectral subspace for S corresponding to the contour Γ.

Then in particular X_2 is S-invariant and so $S_{12}=0$. Write $P=P(S;\Gamma)$. The operator S_{22} is the restriction of S to Im P. Thus the eigenvalues of S_{22} are precisely the eigenvalues of S lying inside Γ. Let S_0 be the restriction of S to Ker P. Then the eigenvalues of S_0 are precisely the eigenvalues of S lying outside Γ. In particular S_{22} and S_0 have no common eigenvalues. It remains to prove that S_0 and S_{11} have the same eigenvalues.

Since $\operatorname{Im} P = X_2 = \operatorname{Im} \Pi$, we have $I-P=(I-P)(I-\Pi)$ and the map

$$F=(I-P)|X_1 : X_1 \to \operatorname{Ker} P$$

is an invertible operator. One easily verifies that $S_0F=FS_{11}$. So S_0 and S_{11} are similar, and the proof is complete.

II. Minimality and minimal factorizations. In this chapter we discuss minimal systems and minimal factorizations of rational matrix functions. The main result is Theorem 2.2, which shows that there is a one-to-one correspondence between minimal factorizations and supporting projections of minimal systems.

2.1. Minimal nodes. Let X and Y be linear spaces. A pair of operators (A, B),

$$A: X \to X, \qquad B: Y \to X,$$

is called *controllable* if, for k sufficiently large,

$$\operatorname{Im} B + \operatorname{Im} AB + \cdots + \operatorname{Im} A^{k-1}B = X. \tag{2.1}$$

Similarly, a pair (A, C)

$$A: X \to X, \qquad C: X \to Y,$$

is said to be *observable* if, for k sufficiently large,

$$\operatorname{Ker} C \cap \operatorname{Ker} CA \cap \cdots \cap \operatorname{Ker} CA^{k-1} = (0). \tag{2.2}$$

Observe that the left-hand sides of (2.1) and (2.2) are independent of k, provided k is larger than or equal to the degree of the minimal polynomial of A.

A system $\theta=(A,B,C;X,Y)$ is called *minimal* if (A,B) is controllable and (A,C) is observable. Such systems play an important role in the sequel. Below we collect together a number of facts concerning minimal systems that are either wellknown or easy to prove (cf. [2], [16] and the references given there).

If θ is minimal, then so is θ^x. Similarity of systems implies that their transfer functions coincide. The converse of this is not true in general. However, if θ and Δ are minimal systems for which $W_\theta = W_\Delta$, then θ and Δ are similar. This result is known as the state space isomorphism theorem. If S is a system similarity between two minimal systems, then S is uniquely determined. In other words, the only system similarity between a minimal system and itself is the identity operator. Given a system θ, there exists a minimal system (unique up to similarity) whose transfer function coincides with that of θ. The product of two minimal systems need not be minimal. However, if the product of two systems is minimal, then so are the factors. In particular, if Π is a supporting projection for the minimal system θ, then $\mathrm{pr}_\Pi(\theta)$ and $\mathrm{pr}_{I-\Pi}(\theta)$ are both minimal.

2.2. Minimality and McMillan degree. Let $W(\lambda)$ be a rational $n\times n$ matrix function, and let λ_0 be a complex number. Then λ_0 is at worst a pole of $W(\lambda)$. So, taking p sufficiently large, we may write

$$W(\lambda)=\sum_{j=-p}^{\infty}(\lambda-\lambda_0)^j W_j,$$

the expansion being valid in some deleted neighborhood of λ_0. The rank of the block Hankel matrix

$$\begin{bmatrix} W_{-1} & W_{-2} & \cdots & W_{-p} \\ W_{-2} & \cdots & W_{-p} & 0 \\ \vdots & & & \vdots \\ W_{-p} & 0 & \cdots & 0 \end{bmatrix},$$

is called the *degree* of W at λ_0. It is denoted by $\delta(W; \lambda_0)$. Observe that $\delta(W; \lambda_0)$ does not depend on the choice of p. For equivalent definitions and generalizations, see [15]. We also define the degree $\delta(W; \infty)$ of W at ∞ to be the degree of $W(\lambda^{-1})$ at 0.

It is clear that $\delta(W; \mu) = 0$ if and only if $W(\lambda)$ is analytic at μ. Therefore it makes sense to put

$$\delta(W) = \sum_{\mu \in \mathbb{C}_\infty} \delta(W; \mu).$$

Here $\mathbb{C}_\infty = \mathbb{C} \cup \{\infty\}$. The number $\delta(W)$ is called the *McMillan degree* of W.

Assume that $W(\infty) = I_n$. Then $W(\lambda)$ admits a realization of the form

$$W(\lambda) = I_n + C(\lambda I_\delta - A)^{-1}B. \tag{2.3}$$

The system $\theta = (A, B, C; \mathbb{C}^\delta, \mathbb{C}^n)$ has $W(\lambda)$ as its transfer function. We call the realization (2.3) *minimal* if θ is a minimal system. In fact (2.3) is minimal if and only if δ is equal to the McMillan degree $\delta(W)$ of W. If (2.3) is not minimal, then $\delta > \delta(W)$. From (2.3) it is clear that each pole of $W(\lambda)$ is an eigenvalue of A. In general the converse is not true, but if the realization is minimal, each eigenvalue of A is a pole of $W(\lambda)$. So in that case the set of poles of $W(\lambda)$ coincides with the set of eigenvalues of A. Similarly, if (2.3) is minimal, the set of poles of $W(\lambda)^{-1}$ coincides with the set of eigenvalues of $A^\times = A - BC$. Poles of $W(\lambda)^{-1}$ are usually called *zeros* of $W(\lambda)$.

2.3. Minimal factorizations. Let $W(\lambda)$, $W_1(\lambda)$ and $W_2(\lambda)$ be rational $n \times n$ matrix functions, and assume that

$$W(\lambda) = W_1(\lambda) W_2(\lambda). \tag{2.4}$$

Then it is known (cf., e.g., [26]) that $\delta(W) \leqq \delta(W_1) + \delta(W_2)$. In fact this inequality holds pointwise in the following sense:

$$\delta(W; \mu) \leqq \delta(W_1; \mu) + \delta(W_2; \mu), \qquad \mu \in \mathbb{C}_\infty. \tag{2.5}$$

The factorization (2.4) is called *minimal* if $\delta(W) = \delta(W_1) + \delta(W_2)$. An equivalent requirement is that in (2.5) we have equality for all $\mu \in \mathbb{C}_\infty$.

In dealing with minimal factorizations, we shall always suppose that det $W(\lambda) \not\equiv 0$. This implies the existence of $a \in \mathbb{C}$ such that $W(a)$ is invertible. Put $\tilde{W}(\lambda) = W(a)^{-1} W(\lambda^{-1} + a)$. Then clearly $\tilde{W}(\infty) = I_n$. There is a one-to-one correspondence between the (minimal) factorizations of $W(\lambda)$ and those of $\tilde{W}(\lambda)$. So (from a theoretical point of view) there is no loss of generality in assuming that $W(\infty) = I_n$.

Suppose $W(\infty) = I_n$. We are interested in the minimal factorizations of $W(\lambda)$. We claim that it suffices to consider only those factorizations (2.4) of $W(\lambda)$ for which $W_1(\infty) = W_2(\infty) = I_n$. To make this claim more precise, assume that (2.4) is a minimal factorization of $W(\lambda)$. Then $\delta(W_1; \infty) + \delta(W_2; \infty) = \delta(W; \infty) = 0$, because W is analytic at ∞. Hence $\delta(W_1; \infty) = \delta(W_2; \infty) = 0$, or, in other words, W_1 and W_2 are analytic at ∞. Moreover $I_n = W(\infty) = W_1(\infty) W_2(\infty)$, and so $W_1(\infty)$ and $W_2(\infty)$ are each other's inverse. Put $U = W_1(\infty)^{-1}$. By multiplying $W_1(\lambda)$ from the right with U and

$W_2(\lambda)$ from the left with U^{-1}, we obtain a minimal factorization of $W(\lambda)$ whose factors have the value I_n at ∞.

These considerations justify the fact that, *from now on, without further mentioning, we only deal with rational matrix functions that are analytic at ∞ with value the identity matrix.* In other words, the rational matrix functions considered below will be viewed as transfer functions of systems.

PROPOSITION 2.1. *Let $W(\lambda)=W_1(\lambda)W_2(\lambda)$ be a factorization of the rational matrix function $W(\lambda)$, let θ_1 be a minimal realization for $W_1(\lambda)$ and let θ_2 be a minimal realization for $W_2(\lambda)$. Then the factorization is minimal if and only if the product $\theta_1\theta_2$ is a minimal system.*

Proof. Let $\theta=(A, B, C; \mathbb{C}^\delta, \mathbb{C}^n)$ be a realization for the rational matrix function $W(\lambda)$, i.e., formula (2.3) is satisfied. Then θ is a minimal system if and only if $\delta=\delta(W)$. From this and the definition of the product of two systems, the desired result is clear.

We now come to the main result of this chapter.

THEOREM 2.2. *Let θ be a minimal realization of the rational matrix function $W(\lambda)$.*

(i) *If Π is a supporting projection for θ, $W_1(\lambda)$ is the transfer function of* $\mathrm{pr}_{I-\Pi}(\theta)$ *and $W_2(\lambda)$ is the transfer function of* $\mathrm{pr}_\Pi(\theta)$, *then $W(\lambda)=W_1(\lambda)W_2(\lambda)$ is a minimal factorization of $W(\lambda)$.*

(ii) *If $W(\lambda)=W_1(\lambda)W_2(\lambda)$ is a minimal factorization, then there exists a unique supporting projection Π for the system θ such that $W_1(\lambda)$ and $W_2(\lambda)$ are the transfer functions of* $\mathrm{pr}_{I-\Pi}(\theta)$ *and* $\mathrm{pr}_\Pi(\theta)$, *respectively.*

Proof. Statement (i) is an immediate consequence of Proposition 2.1. Therefore we concentrate on (ii). Assume that $W(\lambda)=W_1(\lambda)W_2(\lambda)$ is a minimal factorization. For $i=1,2$, let θ_i be a minimal realization of $W_i(\lambda)$ with state space $\mathbb{C}^{\delta_i}$. Here δ_i is the McMillan degree $\delta(W_i)$ of W_i. By Proposition 2.1, the product $\theta_1\theta_2$ is a minimal realization for $W(\lambda)$. Hence $\theta_1\theta_2$ and θ are similar, say with system similarity S: $\mathbb{C}^{\delta_1}\oplus\mathbb{C}^{\delta_2}\rightarrow\mathbb{C}^\delta$, where $\delta=\delta_1+\delta_2=\delta(W)$. Let Π be the projection of $\mathbb{C}^\delta$ along $S[\mathbb{C}^{\delta_1}]$ onto $S[\mathbb{C}^{\delta_2}]$. Then Π is a supporting projection for θ. Moreover $\mathrm{pr}_{I-\Pi}(\theta)$ is similar to θ_1 and $\mathrm{pr}_\Pi(\theta)$ is similar to θ_2. It remains to prove the unicity of Π.

Suppose P is another supporting projection of θ such that $\mathrm{pr}_P(\theta)$ and $\mathrm{pr}_{I-P}(\theta)$ are realizations of $W_2(\lambda)$ and $W_1(\lambda)$ respectively. Then $\mathrm{pr}_P(\theta)$ and $\mathrm{pr}_{I-P}(\theta)$ are minimal again. Hence $\mathrm{pr}_{I-\Pi}(\theta)$ and $\mathrm{pr}_{I-P}(\theta)$ are similar, say with system similarity U: Ker $\Pi\rightarrow$ Ker P, and $\mathrm{pr}_\Pi(\theta)$ and $\mathrm{pr}_P(\theta)$ are similar, say with system similarity V: Im $\Pi\rightarrow$ Im P. Define T on $\mathbb{C}^\delta$ by

$$T=\begin{bmatrix} U & 0\\ 0 & V\end{bmatrix}: \mathrm{Ker}\,\Pi\oplus\mathrm{Im}\,\Pi\rightarrow\mathrm{Ker}\,P\oplus\mathrm{Im}\,P.$$

Then T is a system similarity between θ and itself. Since θ is minimal, it follows that T is the identity operator on $\mathbb{C}^\delta$. But then $\Pi=P$, and the proof is complete.

III. Stability of spectral divisors. The problem of computing numerically the minimal factorizations of a given transfer function leads in a natural way to questions concerning the stability of divisors of a system. These and related questions form the main topic of this chapter.

3.1. Examples and first results. The property of having nontrivial minimal factorizations may be ill-conditioned. To see this, consider the following example. Let

$$W_\varepsilon(\lambda)=\begin{bmatrix} 1+\dfrac{1}{\lambda} & \dfrac{\varepsilon}{\lambda^2}\\ 0 & 1+\dfrac{1}{\lambda}\end{bmatrix}. \tag{3.1}$$

For each ε this is the transfer function of the minimal system $\theta_\varepsilon = (A, I, I; \mathbb{C}^2, \mathbb{C}^2)$, where $I = I_2$ and

$$A_\varepsilon = \begin{bmatrix} 0 & \varepsilon \\ 0 & 0 \end{bmatrix}.$$

To find a nontrivial minimal factorization of the function (3.1), we have to find nontrivial supporting projections of the system θ_ε (cf. Theorem 2.2); i.e., we must look for nontrivial subspaces M and M^x of $\mathbb{C}^2$, invariant under A_ε and $A_\varepsilon^x = A_\varepsilon - I$, respectively, such that $M \oplus M^x = \mathbb{C}^2$. The operators A_ε and $A_\varepsilon - I$ have the same invariant subspaces, and for $\varepsilon \neq 0$ there is only one such subspace of dimension one, namely the first coordinate space. It follows that for $\varepsilon \neq 0$, the rational matrix function (3.1) has no nontrivial minimal factorizations. For $\varepsilon = 0$, we have

$$W_0(\lambda) = \begin{bmatrix} 1+\dfrac{1}{\lambda} & 0 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 \\ 0 & 1+\dfrac{1}{\lambda} \end{bmatrix},$$

and this factorization is minimal, because the McMillan degree of $W_0(\lambda)$ is equal to 2 and the McMillan degree of each factor is 1.

The next theorem shows that under certain conditions the existence of a minimal factorization is a stable property.

THEOREM 3.1. *Consider the rational matrix function*

$$W_0(\lambda) = I_n + C_0(\lambda I_\delta - A_0)^{-1} B_0. \tag{3.2}$$

Assume that the realization (3.2) *is minimal and that* $W_0(\lambda)$ *admits a minimal factorization* $W_0(\lambda) = W_{01}(\lambda) W_{02}(\lambda)$,

$$W_{0i}(\lambda) = I_n + C_{0i}(\lambda I_{\delta_i} - A_{0i})^{-1} B_{0i}, \qquad i = 1, 2, \tag{3.3}$$

where $\delta = \delta_1 + \delta_2$ *and the factors* $W_{01}(\lambda)$ *and* $W_{02}(\lambda)$ *have no common zeros and no common poles. Then there exist positive constants* ω *and* ε *such that the following holds. If* A, B, *and* C *are matrices* (*of the appropriate sizes*) *such that*

$$\|A - A_0\| + \|B - B_0\| + \|C - C_0\| < \omega, \tag{3.4}$$

then the realization $W(\lambda) = I_n + C(\lambda I_\delta - A)^{-1} B$ *is minimal and* $W(\lambda)$ *admits a minimal factorization* $W(\lambda) = W_1(\lambda) W_2(\lambda)$,

$$W_i(\lambda) = I_n + C_i(\lambda I_{\delta_i} - A_i)^{-1} B_i, \qquad i = 1, 2, \tag{3.5}$$

such that the factors $W_1(\lambda)$ *and* $W_2(\lambda)$ *have no common zeros and no common poles, and*

$$\|A_i - A_{0i}\| + \|B_i - B_{0i}\| + \|C_i - C_{0i}\| \leqq \varepsilon (\|A - A_0\| + \|B - B_0\| + \|C - C_0\|),$$

for $i = 1, 2$.

The above theorem deals with "spectral" minimal factorizations. The stability of nonspectral minimal factorizations is investigated in [5]. The theorem is proved in § 3.3. The proof provides explicit estimates for ω and ε.

3.2. Opening between subspaces and angular operators. Let M and N be subspaces of the linear space X, and let $\|\cdot\|$ be a norm on X. The number

$$\eta(M, N) = \inf \{\|x + y\| \mid x \in M, y \in N, \max\{\|x\|, \|y\|\} = 1\}$$

is called the *minimal opening* between M and N. Note that always $0 \leqq \eta(M, N) \leqq 1$, except when both M and N are trivial, in which case $\eta(M, N)=\infty$. It is well known (see [14, Lemma 1]) that $\eta(M, N)>0$ if and only if $M\cap N=(0)$. If Π is a projection of the space X, then

$$\max\{\|\Pi\|, \|I-\Pi\|\}\leqq\frac{1}{\eta(\operatorname{Ker}\Pi, \operatorname{Im}\Pi)}. \tag{3.6}$$

Sometimes it will be convenient to describe $\eta(M, N)$ in terms of the *minimal angle* $\phi_{\min}$ between M and N. By definition this quantity is given by the following formulas:

$$0\leqq\phi_{\min}\leqq\frac{\Pi}{2},\qquad \sin\phi_{\min}=\eta(M, N).$$

(cf. [14]). Put

$$\rho(M, N)=\sup_{0\neq x\in M}\inf_{y\in N}\frac{\|x-y\|}{\|x\|}.$$

The number

$$\operatorname{gap}(M, N)=\max\{\rho(M, N), \rho(N, M)\}$$

is called the *gap* or *maximal opening* between M and N. There is an extensive literature on this concept[1].

Now let us assume that $X=\mathbb{C}^n$, endowed with the usual Euclidean norm. Let P and Q be the orthogonal projections of X onto M and N, respectively. It can be shown that $\operatorname{gap}(M, N)=\|P-Q\|$ (cf. [1]). Also

$$1-\eta(M, N)^2=\sup_{0\neq x\in M}\frac{\|Qx\|^2}{\|x\|^2}=\sup_{0\neq y\in N}\frac{\|Py\|^2}{\|y\|^2}, \tag{3.7}$$

provided both M and N are nontrivial.

LEMMA 3.2. *Let* Π_0, Π *and* Π_1 *be projections of the linear space* X, *and assume that* $\operatorname{Ker}\Pi_0=\operatorname{Ker}\Pi=\operatorname{Ker}\Pi_1$. *Let* R *be the angular operator of* $\operatorname{Im}\Pi$ *with respect to* Π_0 *and let* R_1 *be that of* $\operatorname{Im}\Pi_1$. *The following assertions hold*:

(i) $\eta(\operatorname{Ker}\Pi_0, \operatorname{Im}\Pi_0)\cdot\rho(\operatorname{Im}\Pi_1, \operatorname{Im}\Pi)\leqq\|R_1-R\|$.

(ii) *If* $\rho(\operatorname{Im}\Pi_1, \operatorname{Im}\Pi)<\eta(\operatorname{Ker}\Pi, \operatorname{Im}\Pi)$, *then*

$$\|R_1-R\|\leqq\frac{\rho(\operatorname{Im}\Pi_1, \operatorname{Im}\Pi)\cdot(1+\|R\|)}{\eta(\operatorname{Ker}\Pi, \operatorname{Im}\Pi)-\rho(\operatorname{Im}\Pi_1, \operatorname{Im}\Pi)}.$$

In particular, if $\rho(\operatorname{Im}\Pi_1, \operatorname{Im}\Pi_0)<\eta(\operatorname{Ker}\Pi_0, \operatorname{Im}\Pi_0)$, *then*

$$\|R_1\|\leqq\frac{\rho(\operatorname{Im}\Pi_1, \operatorname{Im}\Pi_0)}{\eta(\operatorname{Ker}\Pi_0, \operatorname{Im}\Pi_0)-\rho(\operatorname{Im}\Pi_1, \operatorname{Im}\Pi_0)}.$$

Results of this type seem to be well known. Therefore we omit the proof. We proceed with a lemma which will be most useful in the next section.

LEMMA 3.3. *Let* P, P^x, Q *and* Q^x *be projections of the linear space* X, *and put* $\alpha_0=\frac{1}{6}\eta(\operatorname{Im}P, \operatorname{Im}P^x)\cdot(1+\|P^x\|)^{-1}$. *Assume* $X=\operatorname{Im}P\oplus\operatorname{Im}P^x$ *and*

$$\|P-Q\|+\|P^x-Q^x\|<\alpha_0. \tag{3.8}$$

[1] For details, see T. Kato: *Perturbation Theory For Linear Operators*, Springer, Berlin–Heidelberg–New York, 1966, and the references given there.

Then $X=\operatorname{Im} Q\oplus\operatorname{Im} Q^x$ *and there exists an invertible operator* $S: X\to X$ *such that*

(i) $S[\operatorname{Im} Q]=\operatorname{Im} P,\quad S[\operatorname{Im} Q^x]=\operatorname{Im} P^x,$

(ii) $\max\{\|S-I\|,\|S^{-1}-I\|\}\leqq\beta(\|P-Q\|+\|P^x-Q^x\|),$

where $\beta=2[\alpha_0\eta(\operatorname{Im} P,\operatorname{Im} P^x)]^{-1}$.

Proof. For simplicity we put $d=\|P-Q\|+\|P^x-Q^x\|$ and $\eta=\eta(\operatorname{Im} P,\operatorname{Im} P^x)$. Since $\operatorname{gap}(\operatorname{Im} P,\operatorname{Im} Q)\leqq\|P-Q\|$ and $\operatorname{gap}(\operatorname{Im} P^x,\operatorname{Im} Q^x)\leqq\|P^x-Q^x\|$, condition (3.8) implies that

$$2\operatorname{gap}(\operatorname{Im} P,\operatorname{Im} Q)+2\operatorname{gap}(\operatorname{Im} P^x,\operatorname{Im} Q^x)<\eta.$$

But then we may apply Theorem 2 from [14] to show that $X=\operatorname{Im} Q\oplus\operatorname{Im} Q^x$.

Note that (3.8) also implies that $\|P-Q\|<\frac{1}{4}$. Hence $S_1=I+P-Q$ is invertible and we can write $S_1^{-1}=I+V$ with $\|V\|\leqq\frac{4}{3}\|P-Q\|<\frac{1}{3}$. As $I-P+Q$ is invertible too, we have $S_1[\operatorname{Im} Q]=\operatorname{Im} P$. By direct calculation, it can be shown that

$$\|S_1Q^xS_1^{-1}-P^x\|\leqq 3\|P^x-Q^x\|+3\cdot\|P^x\|\cdot\|P-Q\|,$$

and hence

$$\rho(\operatorname{Im} S_1Q^xS_1^{-1},\operatorname{Im} P^x)\leqq\|S_1Q^xS_1^{-1}-P^x\|\leqq 3d(1+\|P^x\|)<\frac{\eta}{2}.$$

Let $\Pi_0(\Pi)$ be the projection of X along $\operatorname{Im} P(\operatorname{Im} Q)$ onto $\operatorname{Im} P^x(\operatorname{Im} Q^x)$, and put $\tilde{\Pi}=S_1\Pi S_1^{-1}$. Then $\tilde{\Pi}$ is a projection of X and $\operatorname{Ker}\tilde{\Pi}=\operatorname{Ker}\Pi_0$. Further, $\operatorname{Im}\tilde{\Pi}=\operatorname{Im} S_1Q^xS_1^{-1}$, and so we have

$$\rho(\operatorname{Im}\tilde{\Pi},\operatorname{Im}\Pi_0)<\frac{\eta}{2}=\frac{1}{2}\eta(\operatorname{Ker}\Pi_0,\operatorname{Im}\Pi_0).$$

Hence, if R denotes the angular operator of $\operatorname{Im}\tilde{\Pi}$ with respect to Π_0, then because of Lemma 3.2,

$$\|R\|\leqq\frac{2}{\eta}\rho(\operatorname{Im}\tilde{\Pi},\operatorname{Im}\Pi_0).$$

Since $\rho(\operatorname{Im}\tilde{\Pi},\operatorname{Im}\Pi_0)\leqq 3d(1+\|P^x\|)$, this implies that $\|R\|\leqq d\alpha_0^{-1}$.

Next, put $S_2=I-R\Pi_0$, and take $S=S_2S_1$. Clearly, S_2 is invertible; in fact $S_2^{-1}=I+R\Pi_0$. It follows that S is invertible too. From the properties of the angular operator one easily sees that with this choice of S statement (i) holds true. It remains to prove (ii).

From $S=(I-R\Pi_0)(I+P-Q)$ and the fact that $\|P-Q\|<\frac{1}{4}$, one deduces that $\|S-I\|\leqq\|P-Q\|+\frac{5}{4}\|R\|\cdot\|\Pi_0\|$. Moreover $\|R\|\leqq d\alpha_0^{-1}$, and from (3.6) we know that $\|\Pi_0\|\leqq\eta^{-1}$. It follows that

$$\|S-I\|\leqq d+\frac{5d}{4\alpha_0\eta}. \tag{3.9}$$

Recall that $S_1^{-1}=I+V$ with $\|V\|\leqq\frac{4}{3}\|P-Q\|<\frac{1}{3}$. This can be used to show that

$$\|S^{-1}-I\|\leqq\frac{4d}{3}+\frac{4d}{3\alpha_0\eta}. \tag{3.10}$$

Statement (ii) is now an easy consequence of (3.9), (3.10), and the fact that $6\alpha_0\eta\leqq 1$.

3.3. Stability of spectral divisors. Let $\theta=(A,B,C;X,Y)$ and $\theta_0=(A_0,B_0,C_0;X,Y)$ be systems. The *distance* between θ and θ_0 is defined to be the number

$$\|\theta-\theta_0\|=\|A-A_0\|+\|B-B_0\|+\|C-C_0\|.$$

We also put $\|\theta\|=\|A\|+\|B\|+\|C\|$. If $W(\lambda)$ and $W_0(\lambda)$ are the transfer functions of θ and θ_0, respectively, then

$$\|W(\lambda)-W_0(\lambda)\|\leqq\frac{\|\theta-\theta_0\|\cdot\|\theta\|\cdot\|\theta_0\|}{\|A\|\cdot\|A_0\|},$$

provided that $|\lambda|>2\max\{\|A\|,\|A_0\|\}$.

THEOREM 3.4. *Let* Π_0 *be a supporting projection for the system* $\theta_0=(A_0, B_0, C_0; X, Y)$, *and assume that*

$$\operatorname{Ker}\Pi_0=\operatorname{Im}P,\qquad \operatorname{Im}\Pi_0=\operatorname{Im}P^x,$$

where P *and* P^x *are projections of* X. *Put*

$$\alpha_0=\tfrac{1}{6}\eta(\operatorname{Ker}\Pi_0,\operatorname{Im}\Pi_0)\cdot(1+\|P^x\|)^{-1}.$$

Let $\theta=(A, B, C; X, Y)$ *be another system, and let* Q *and* Q^x *be projections of* X *such that*

$$A[\operatorname{Im}Q]\subset\operatorname{Im}Q,\qquad A^x[\operatorname{Im}Q^x]\subset\operatorname{Im}Q^x,\tag{3.11}$$
$$\|P-Q\|+\|P^x-Q^x\|<\alpha_0.$$

Then $X=\operatorname{Im}Q\oplus\operatorname{Im}Q^x$, *there exists an invertible operator* $S: X\to X$ *such that* $S^{-1}\Pi_0S$ *is the projection* Π *of* X *onto* $\operatorname{Im}Q^x$ *along* $\operatorname{Im}Q$, *and the projection* Π_0 *is a supporting projection for the system* $\hat\theta=(SAS^{-1}, SB, CS^{-1}; X, Y)$, *while for the corresponding factors we have*

$$\max\{\|\mathrm{pr}_{I-\Pi_0}(\theta_0)-\mathrm{pr}_{I-\Pi_0}(\hat\theta)\|,\|\mathrm{pr}_{\Pi_0}(\theta_0)-\mathrm{pr}_{\Pi_0}(\hat\theta)\|\}$$
$$\leqq\frac{9}{\eta(\operatorname{Im}P,\operatorname{Im}P^x)^3}\left[\|\theta-\theta_0\|+\frac{\|\theta_0\|}{\alpha_0}(\|P-Q\|+\|P^x-Q^x\|)\right].$$

Proof. From Lemma 3.3 we know that $X=\operatorname{Im}Q\oplus\operatorname{Im}Q^x$. Take S as in Lemma 3.3. Then $S^{-1}\Pi_0S$ is the projection Π of X onto $\operatorname{Im}Q^x$ along $\operatorname{Im}Q$. It is now clear from formula (3.11) that $S^{-1}\Pi_0S$ is a supporting projection for the system $\theta=(A, B, C; X, Y)$. But then Π_0 is a supporting projection for $\hat\theta=(SAS^{-1}, SB, CS^{-1}; X, Y)$.

Let θ_{01} and $\hat\theta_1$ be the left factors of θ_0 and $\hat\theta$ associated with Π_0, and let θ_{02} and $\hat\theta_2$ be the right factors. From the definition of the factors (see § 1.1) it is clear that $\|\theta_{01}-\hat\theta_1\|\leqq\|I-\Pi_0\|\cdot\|\theta_0-\hat\theta\|$ and $\|\theta_{02}-\hat\theta_2\|\leqq\|\Pi_0\|\cdot\|\theta_0-\hat\theta\|$. Using (3.6), we obtain

$$\max\|\theta_{0i}-\hat\theta_i\|\leqq\frac{1}{\eta}\|\theta_0-\hat\theta\|,\qquad i=1,2,\tag{3.12}$$

where $\eta=\eta(\operatorname{Im}P,\operatorname{Im}P^x)$. As $\|\theta_0-\hat\theta\|\leqq\|\theta_0-\theta\|+\|\theta-\hat\theta\|$, it remains to compute a suitable upper bound for $\|\theta-\hat\theta\|$.

Put $S=I+V$ and $S^{-1}=I+W$. Note that $\|\theta-\hat\theta\|\leqq\|A\|(\|V\|+\|W\|+\|V\|\cdot\|W\|)+\|B\|\cdot\|V\|+\|C\|\cdot\|W\|$. By Lemma 3.3, we have $\max\{\|V\|,\|W\|\}\leqq2d(\alpha_0\eta)^{-1}$, where $d=\|P-Q\|+\|P^x-Q^x\|$. It follows that

$$\|\theta-\hat\theta\|\leqq\frac{4d}{\alpha_0\eta}\left(1+\frac{d}{\alpha_0\eta}\right)\|\theta\|.\tag{3.13}$$

Since $d\alpha_0^{-1} < 1$ and $\eta \leqq 1$, one can use formula (3.13) to show that

$$\|\theta_0 - \hat{\theta}\| \leqq \frac{9}{\eta^3}\left[\|\theta - \theta_0\| + \frac{d}{\alpha_0}\|\theta_0\|\right].$$

This, together with formula (3.12), gives the desired result.

THEOREM 3.5. *Let Π_0 be a supporting projection for the system $\theta_0 = (A_0, B_0, C_0; X, Y)$. Assume that*

$$\operatorname{Ker} \Pi_0 = \operatorname{Im} P(A_0; \Gamma), \qquad \operatorname{Im} \Pi_0 = \operatorname{Im} P(A_0^x; \Gamma^x),$$

where Γ and Γ^x are contours such that A_0 and A_0^x have no eigenvalues on Γ and Γ^x, respectively. Then there exist positive constants α, β_1 and β_2 such that the following holds. If $\theta = (A, B, C; X, Y)$ is a system such that $\|\theta - \theta_0\| < \alpha$, then A has no eigenvalues on Γ, A^x has no eigenvalues on Γ^x,

$$X = \operatorname{Im} P(A; \Gamma) \oplus \operatorname{Im} P(A^x; \Gamma^x),$$

the projection Π of X along $\operatorname{Im} P(A; \Gamma)$ onto $\operatorname{Im} P(A^x; \Gamma^x)$ is a supporting projection for θ, and there exists a similarity transformation S such that

$$\|S - I\| \leqq \beta_1 \|\theta - \theta_0\|,$$

$\Pi_0 = S\Pi S^{-1}$, Π_0 is a supporting projection for the system $\hat{\theta} = (SAS^{-1}, SB, CS^{-1}; X, Y)$ and for the corresponding divisors we have

$$\|\mathrm{pr}_{I-\Pi_0}(\theta_0) - \mathrm{pr}_{I-\Pi_0}(\hat{\theta})\| \leqq \beta_2 \|\theta - \theta_0\|,$$
$$\|\mathrm{pr}_{\Pi_0}(\theta_0) - \mathrm{pr}_{\Pi_0}(\hat{\theta})\| \leqq \beta_2 \|\theta - \theta_0\|.$$

Furthermore, if θ_0 is minimal, then α can be chosen such that θ is minimal whenever $\|\theta - \theta_0\| < \alpha$.

Proof. Let ℓ be the maximum of the lengths of the curves Γ and Γ^x,

$$\gamma = \max\left\{\max_{\lambda \in \Gamma} \|(\lambda I - A_0)^{-1}\|, \max_{\lambda \in \Gamma^x} \|(\lambda I - A_0^x)^{-1}\|\right\},$$

and

$$\alpha_0 = \tfrac{1}{6}\eta(\operatorname{Ker} \Pi_0, \operatorname{Im} \Pi_0) \cdot (1 + \|P(A_0^x; \Gamma^x)\|)^{-1}.$$

Put

$$\alpha = (1 + \|\theta_0\|)^{-1} \min\left\{1, \frac{1}{2\gamma}, \frac{\alpha_0 \pi}{2\gamma^2 \ell}\right\},$$

$$\beta_1 = 4(1 + \|\theta_0\|)\gamma^2 \ell [\pi \alpha_0 \eta(\operatorname{Ker} \Pi_0, \operatorname{Im} \Pi_0)]^{-1}, \tag{3.14}$$

$$\beta_2 = \frac{9}{\eta(\operatorname{Ker} \Pi_0, \operatorname{Im} \Pi_0)^3}\left[1 + \frac{2\gamma^2 \ell}{\pi \alpha_0}\|\theta_0\|(1 + \|\theta_0\|)\right].$$

We shall prove that α, β_1 and β_2 have the properties mentioned in the first part of the theorem. For convenience we write $\eta = \eta(\operatorname{Ker} \Pi_0, \operatorname{Im} \Pi_0)$, $P = P(A_0; \Gamma)$ and $P^x = P(A_0^x; \Gamma^x)$.

Suppose $\theta = (A, B, C; X, Y)$ is a system such that $\|\theta - \theta_0\| < \alpha$. Then, in particular, $\|\theta - \theta_0\| < 1$. Since $\|A^x - A_0^x\| \leqq \|\theta - \theta_0\| \cdot (1 + \|\theta_0\|)$, it follows that

$$\max\{\|A - A_0\|, \|A^x - A_0^x\|\} < \frac{1}{2\gamma}.$$

Using elementary spectral theory, we may conclude that A has no eigenvalues on Γ, $A^\times$ has no eigenvalues on $\Gamma^\times$, while further,

$$\|(\lambda I-A)^{-1}-(\lambda I-A_0)^{-1}\|\leqq 2\gamma^2\|\theta-\theta_0\|\cdot(1+\|\theta_0\|),\qquad \lambda\in\Gamma,$$
$$\|(\lambda I-A^\times)^{-1}-(\lambda I-A_0^\times)^{-1}\|\leqq 2\gamma^2\|\theta-\theta_0\|\cdot(1+\|\theta_0\|),\qquad \lambda\in\Gamma^\times.$$

Hence for the corresponding Riesz projections $Q=P(A;\Gamma)$ and $Q^\times=P(A^\times;\Gamma^\times)$, we have

$$\|P-P^\times\|+\|Q-Q^\times\|\leqq\frac{2\gamma^2\ell}{\pi}\|\theta-\theta_0\|\cdot(1+\|\theta_0\|)<\alpha_0.$$

The fact that α, β_1, and β_2 meet the requirements of the first part of the theorem is now an easy consequence of Lemma 3.3 and Theorem 3.4.

Assume that θ_0 is minimal. We want to define the constant α in such a way that it also has the property that θ is minimal whenever $\|\theta-\theta_0\|<\alpha$. Let n be the dimension of X. Since θ_0 is minimal, the operator col $(C_0A_0^j)_{j=0}^{n-1}$ is left invertible, say with left inverse L, and the operator row $(A_0^jB_0)_{j=0}^{n-1}$ is right invertible, say with right inverse R. If $E:X\to Y^n$ is an operator satisfying

$$\|E-\operatorname{col}(C_0A_0^j)_{j=0}^{n-1}\|<\|L\|^{-1},$$

then E is also left invertible. A similar remark can be made involving row $(A_0^jB_0)_{j=0}^{n-1}$. Hence there exists a positive number $\omega(\theta_0)$ such that $\|\theta-\theta_0\|<\omega(\theta_0)$ implies that θ is minimal. The new α may now be defined as

$$\alpha=\min\left[\omega(\theta_0),(1+\|\theta_0\|)^{-1}\min\left\{1,\frac{1}{2\gamma},\frac{\alpha_0\pi}{2\gamma^2\ell}\right\}\right].\tag{3.15}$$

This completes the proof of the theorem.

We now come to the proof of Theorem 3.1.

Proof of Theorem 3.1. The matrices appearing in Theorem 3.1, will be identified with their canonical action as operators. Put $\theta_0=(A_0,B_0,C_0;\mathbb{C}^\delta,\mathbb{C}^n)$ and

$$\theta_{0i}=(A_{0i},B_{0i},C_{0i};\mathbb{C}^{\delta_i},\mathbb{C}^n),\qquad i=1,2.$$

Since the factorization $W_0(\lambda)=W_{01}(\lambda)W_{02}(\lambda)$ is minimal and $\delta_1+\delta_2=\delta$, the realizations (3.3) are minimal. Hence θ_0 is similar to the product $\bar\theta_0=\theta_{01}\theta_{02}$, say with system similarity $T:\mathbb{C}^\delta\to\mathbb{C}^{\delta_1}\oplus\mathbb{C}^{\delta_2}$.

With respect to the direct sum $\mathbb{C}^{\delta_1}\oplus\mathbb{C}^{\delta_2}$, the main operator $\bar A_0$ and the associate main operator $\bar A_0^\times$ of the system $\bar\theta_0=\theta_{01}\theta_{02}$ have the form

$$\bar A_0=\begin{bmatrix}A_{01} & B_{01}C_{02}\\ 0 & A_{02}\end{bmatrix},\qquad \bar A_0^\times=\begin{bmatrix}A_{01}^\times & 0\\ -B_{02}C_{01} & A_{02}^\times\end{bmatrix}.$$

The hypothesis of Theorem 3.1 concerning the poles and zeros of $W_{01}(\lambda)$ and $W_{02}(\lambda)$ just means that A_{01} and A_{02} have no common eigenvalues and that $A_{01}^\times$ and $A_{02}^\times$ have no common eigenvalues. Let Γ be a contour that separates the eigenvalues of A_{01} from those of A_{02}. Similarly, let $\Gamma^\times$ be a contour that separates the eigenvalues of $A_{01}^\times$ from those of $A_{02}^\times$. Then

$$\operatorname{Im}P(\bar A_0;\Gamma)=\mathbb{C}^{\delta_1}\oplus(0),\qquad \operatorname{Im}P(\bar A_0^\times;\Gamma^\times)=(0)\oplus\mathbb{C}^{\delta_2}.$$

It follows that we may apply Theorem 3.5 to the system $\bar\theta_0$.

Let α and β_2 be the positive numbers that according to Theorem 3.5 correspond to the system $\bar{\theta}_0$ (cf. (3.14) and (3.15)), and put

$$\omega = \alpha[\|T\| \cdot \|T^{-1}\| + \|T\| + \|T^{-1}\|]^{-1},$$

$$\varepsilon = \beta_2[\|T\| \cdot \|T^{-1}\| + \|T\| + \|T^{-1}\|]^{-1}.$$

Suppose (3.4) holds and write $\bar{\theta} = (TAT^{-1}, TB, CT^{-1}; \mathbb{C}^{\delta_1} \oplus \mathbb{C}^{\delta_2}, \mathbb{C}^n)$. Then

$$\|\bar{\theta} - \bar{\theta}_0\| \leqq \|\theta - \theta_0\| \cdot (\|T\| \cdot \|T^{-1}\| + \|T\| + \|T^{-1}\|) < \alpha.$$

Hence $\bar{\theta}$ is minimal. This means that the realization $W(\lambda) = I_n + C(\lambda I_\delta - A)^{-1}B$ is minimal. Moreover, there exists a similarity transformation S such that for the system

$$\tilde{\theta} = (STAT^{-1}S^{-1}, STB, CT^{-1}S^{-1}; \mathbb{C}^{\delta_1} \oplus \mathbb{C}^{\delta_2}, \mathbb{C}^n),$$

the projection Π_0 of $\mathbb{C}^{\delta_1} \oplus \mathbb{C}^{\delta_2}$ along $\mathbb{C}^{\delta_1} \oplus (0)$ onto $(0) \oplus \mathbb{C}^{\delta_2}$ is a supporting projection. This shows that $W(\lambda)$ admits a minimal factorization $W(\lambda) = W_1(\lambda)W_2(\lambda)$, with $W_1(\lambda)$ and $W_2(\lambda)$ of the form (3.5). We also know that

$$\|\mathrm{pr}_{I-\Pi_0}(\bar{\theta}_0) - \mathrm{pr}_{I-\Pi_0}(\tilde{\theta})\| \leqq \beta_2 \|\bar{\theta}_0 - \bar{\theta}\|,$$

$$\|\mathrm{pr}_{\Pi_0}(\bar{\theta}_0) - \mathrm{pr}_{\Pi_0}(\tilde{\theta})\| \leqq \beta_2 \|\bar{\theta}_0 - \bar{\theta}\|.$$

But this is the same as

$$\|A_i - A_{0i}\| + \|B_i - B_{0i}\| + \|C_i - C_{0i}\| \leqq \beta_2 \|\bar{\theta}_0 - \bar{\theta}\|$$

$$\leqq \varepsilon \|\theta - \theta_0\| = \varepsilon(\|A - A_0\| + \|B - B_0\| + \|C - C_0\|).$$

Let $\bar{A}$ be the main operator of $\bar{\theta}$, and let $\bar{A}^\times$ be the main operator of the associate system $\bar{\theta}^\times$. As $\|\bar{\theta} - \bar{\theta}_0\| < \alpha$, we can apply Theorem 3.5 to show that $\bar{A}$ has no eigenvalues on Γ, $\bar{A}^\times$ has no eigenvalues on $\Gamma^\times$, and

$$\mathbb{C}^{\delta_1} \oplus \mathbb{C}^{\delta_2} = \operatorname{Im} P(\bar{A}; \Gamma) \oplus \operatorname{Im} P(\bar{A}^\times; \Gamma^\times).$$

Let Π be the projection of $\mathbb{C}^{\delta_1} \oplus \mathbb{C}^{\delta_2}$ along $\operatorname{Im} P(\bar{A}; \Gamma)$ onto $\operatorname{Im} P(\bar{A}^\times; \Gamma^\times)$. Then $\Pi_0 = S\Pi S^{-1}$. It follows that the eigenvalues of A_1 are inside and those of A_2 are outside the contour Γ. Similarly the eigenvalues of $A_2^\times$ are inside and those of $A_1^\times$ are outside $\Gamma^\times$. Thus $W_1(\lambda)$ and $W_2(\lambda)$ have no common zeros and no common poles. This completes the proof.

3.4. Application to the Riccati equation. In this section we show that the method of § 3.3 also can be used to prove stability theorems for solutions of the operator Riccati equation. Here we restrict ourselves to "spectral" solutions (cf. Theorem 3.6 below). The general case has been investigated in [4], [8].

Throughout this section, X_1 and X_2 are linear spaces. We use the symbol $\mathscr{L}(X_j, X_i)$ to indicate the space of all linear operators from X_j into X_i.

THEOREM 3.6. *Let $T_{ij} \in \mathscr{L}(X_j, X_i)$, $i, j = 1, 2$, and let $R \in \mathscr{L}(X_2, X_1)$ be a solution of the Riccati equation*

$$ZT_{21}Z + ZT_{22} - T_{11}Z - T_{12} = 0.$$

Assume that $T_{11} - RT_{21}$ and $T_{22} + T_{21}R$ have no common eigenvalues, and let Γ be a contour whose interior domain contains all eigenvalues of $T_{22} + T_{21}R$ and whose exterior domain contains all eigenvalues of $T_{11} - RT_{21}$. Then there exist positive constants ω and ε such that the following holds. If $S_{ij} \in \mathscr{L}(X_j, X_i)$ and

$$(3.16) \qquad \|S_{ij} - T_{ij}\| < \omega, \qquad i, j = 1, 2,$$

then the equation

$$ZS_{21}Z + ZS_{22} - S_{11}Z - S_{12} = 0 \tag{3.17}$$

has a solution $Q \in \mathcal{L}(X_2, X_1)$ *such that all eigenvalues of* $S_{22} + S_{21}Q$ *are inside* Γ, *all eigenvalues of* $S_{11} - QS_{21}$ *are outside* Γ *and*

$$\|R - Q\| \leqq \varepsilon \max_{i,j=1,2} \|T_{ij} - S_{ij}\|. \tag{3.18}$$

Proof. Consider the operators

$$T = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix}, \qquad S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{12} \end{bmatrix},$$

on $X = X_1 \oplus X_2$. Assume that X is endowed with the norm $\|(x_1, x_2)\| = \|x_1\| + \|x_2\|$. Then

$$\|T - S\| \leqq 2 \max_{i,j=1,2} \|T_{ij} - S_{ij}\|. \tag{3.19}$$

From Proposition 1.5 we know that $N_R = \{(Rx, x) | x \in X_2\}$ is a spectral subspace for T. In fact, if Γ is as in the statement of the theorem, then T has no eigenvalues on Γ and $N_R = \operatorname{Im} P(T; \Gamma)$.

Let ℓ be the length of Γ, and put $\gamma = \max_{\lambda \in \Gamma} \|(\lambda I - T)^{-1}\|$. Take $\|T - S\| < (2\gamma)^{-1}$. By elementary spectral theory this implies that S has no eigenvalues on Γ and

$$\|(\lambda I - T)^{-1} - (\lambda I - S)^{-1}\| \leqq 2\gamma^2 \|S - T\|, \qquad \lambda \in \Gamma.$$

But then $\|P(T; \Gamma) - P(S; \Gamma)\| \leqq \pi^{-1}\gamma^2 \ell \|S - T\|$.

As $X = X_1 \oplus N_R$, the number $\eta(X_1, N_R)$ is positive. Put

$$\omega = \min\left\{\frac{1}{4\gamma}, \frac{\pi}{4\gamma^2 \ell} \eta(X_1, N_R)\right\},$$

and assume that (3.16) holds true. By (3.19) this implies that $\|T - S\| < 2\omega \leqq (2\gamma)^{-1}$, and we can apply the result of the previous paragraph to show that

$$\|P(T; \Gamma) - P(S; \Gamma)\| < \tfrac{1}{2}\eta(X_1, N_R).$$

In particular we see that

$$\operatorname{gap}(N_R, \operatorname{Im} P(S; \Gamma)) < \tfrac{1}{2}\eta(X_1, N_R). \tag{3.20}$$

By Theorem 2 in [14] this implies that

$$X = X_1 \oplus \operatorname{Im} P(S; \Gamma).$$

It follows that there exists $Q \in \mathcal{L}(X_2, X_1)$ such that

$$N_Q = \{(Qz, z) | z \in X_2\} = \operatorname{Im} P(S; \Gamma).$$

By the results of § 1.3, this operator Q is a solution of (3.17), the eigenvalues of $S_{22} + S_{21}Q$ are inside Γ, and the eigenvalues of $S_{11} - QS_{21}$ are outside Γ.

By (3.20), we have $\operatorname{gap}(N_R, N_Q) < \frac{1}{2}\eta(X_1, N_R)$. So we can apply Lemma 3.2 to show that

$$\|R - Q\| \leqq \frac{2(1 + \|R\|)}{\eta(X_1, N_R)} \operatorname{gap}(N_R, N_Q). \tag{3.21}$$

But

$$\text{(3.22)} \qquad \operatorname{gap}(N_R, N_Q) \leqq \|P(T, \Gamma) - P(S; \Gamma)\| \leqq \frac{\gamma^2\ell}{\pi}\|T-S\| \leqq 2\frac{\gamma^2\ell}{\pi}\max_{i,j=1,2}\|T_{ij}-S_{ij}\|.$$

Put

$$\varepsilon = 4(1+\|R\|)\frac{\gamma^2\ell}{\pi\eta(X_1, N_R)}.$$

Then we see from (3.21) and (3.22) that (3.18) holds true. This completes the proof of the theorem.

IV. Numerical and computational aspects. In this chapter we shall discuss some of the practical numerical aspects of minimal factorization of rational matrix functions. In contrast with the results obtained in earlier sections, the coordinate system becomes here of crucial importance. Indeed, for computational problems, the matrices A, B, and C determining the transfer function

$$\text{(4.1)} \qquad W(\lambda) = I_n + C(\lambda I_\delta - A)^{-1}B,$$

are known with a certain relative accuracy. Any coordinate transformation T, required to construct a factorization, causes a loss of accuracy which is proportional to

$$\operatorname{cond}(T) = \|T\| \cdot \|T^{-1}\|,$$

(cf. [24]). The number cond (T) is called the *condition number* of T.

THEOREM 3.7. *Let* $\mathbb{C}^\delta = \mathbb{C}^{\delta_1} \oplus \mathbb{C}^{\delta_2}$ *be the* (*Euclidean*) *direct sum of* $\mathbb{C}^{\delta_1}$ *and* $\mathbb{C}^{\delta_2}$, *and let* $T_1: \mathbb{C}^{\delta_1} \to \mathbb{C}^\delta$ *and* $T_2: \mathbb{C}^{\delta_2} \to \mathbb{C}^\delta$ *be operators. If* T_1 *and* T_2 *are isometries, then*

$$\text{(4.2)} \qquad \|T_1^* T_2\| = \cos\phi_{\min},$$

where $\phi_{\min}$ *is the minimal angle between* Im T_1 *and* Im T_2. *Moreover, if*

$$T = [T_1 T_2]: \mathbb{C}^{\delta_1} \oplus \mathbb{C}^{\delta_2} \to \mathbb{C}^\delta$$

is invertible, then

$$\operatorname{cond}(T) \geqq \frac{1+\cos\phi_{\min}}{\sin\phi_{\min}},$$

equality occurring when T_1 *and* T_2 *are isometries.*

Proof. First assume that T_1 and T_2 are isometries. Put $Q_1 = T_1 T_1^*$. Then Q_1 is the orthogonal projection of $\mathbb{C}^\delta$ onto $M_1 = \operatorname{Im} T_1$. It is not difficult to prove that

$$\|T_1^* T_2\| = \sup_{0 \neq x \in M_2} \frac{\|Q_1 x\|}{\|x\|},$$

where $M_2 = \operatorname{Im} T_2$. Hence, by formula (3.7),

$$\|T_1^* T_2\| = [1 - \eta(M_1, M_2)^2]^{1/2}.$$

The equality (4.2) is now immediate from the definition of $\phi_{\min}$.

Suppose that T is invertible and that T_1 and T_2 are isometries. In order to determine $\|T\|$ and $\|T^{-1}\|$, we compute the spectrum of T^*T. With respect to the decomposition $\mathbb{C}^\delta = \mathbb{C}^{\delta_1} \oplus \mathbb{C}^{\delta_2}$, we have

$$\lambda I - T^*T = \begin{bmatrix} (\lambda-1)I & -T_1^*T_2 \\ -T_2^*T_1 & (\lambda-1)I \end{bmatrix}.$$

For $\lambda \neq 1$, one can write the right-hand side as a product of three operator matrices as follows:

$$\begin{bmatrix} \frac{1}{\lambda-1}I & -\frac{1}{(\lambda-1)^2}T_1^*T_2 \\ 0 & \frac{1}{\lambda-1}I \end{bmatrix} \cdot \begin{bmatrix} (\lambda-1)^2 I - T_1^*T_2T_2^*T_1 & 0 \\ 0 & (\lambda-1)^2 I \end{bmatrix} \cdot \tag{4.3}$$
$$\cdot \begin{bmatrix} I & 0 \\ -\frac{1}{\lambda-1}T_2^*T_1 & I \end{bmatrix}.$$

In this way one sees that for $\lambda \neq 1$, the operator $\lambda I - T^*T$ is invertible if and only if $(\lambda-1)^2 I - T_1^*T_2T_2^*T_1$ is invertible. It follows that

$$\|T\|^2 = 1 + \|T_1^*T_2\|, \qquad \|T^{-1}\|^{-2} = 1 - \|T_1^*T_2\|.$$

But $\|T_1^*T_2\| = \cos\phi_{\min}$, and hence

$$\operatorname{cond}(T)^2 = \frac{1+\cos\phi_{\min}}{1-\cos\phi_{\min}} = \frac{(1+\cos\phi_{\min})^2}{\sin^2\phi_{\min}}.$$

This proves the theorem for the case when T_1 and T_2 are isometries.

Finally we consider the general case, where T_1 and T_2 are arbitrary operators such that $T = [T_1 T_2]$ is invertible. Using polar decomposition, we may write $T_1 = U_1R_1$ and $T_2 = U_2R_2$, where U_1 and U_2 are isometries and R_1 and R_2 are strictly positive selfadjoint operators acting on $\mathbb{C}^{\delta_1}$ and $\mathbb{C}^{\delta_2}$, respectively. Put $S = [U_1U_2]$, and $R = \operatorname{diag}(R_1, R_2)$. Then R is invertible and $T^*T = RS^*SR$.

Set $\alpha = \cos\phi_{\min}$. Then $\alpha = \|U_1^*U_2\|$, and there exists $x \in \mathbb{C}^{\delta_1}$ such that $\|x\| = 1$ and $U_1^*U_2U_2^*U_1x = \alpha^2x$. Put

$$z_i = \begin{bmatrix} x \\ \frac{(-1)^i}{\alpha}U_2^*U_1x \end{bmatrix}, \qquad \lambda_i = 1 + (-1)^i\alpha, \quad i = 1, 2.$$

For $\lambda \neq 1$, we know that $\lambda I - S^*S$ is equal to the product (4.3), provided the operators T_1 and T_2 are replaced by U_1 and U_2, respectively. It follows that

$$S^*Sz_i = \lambda_i z_i, \qquad i = 1, 2.$$

Note that $\|R^{-1}z_1\| = \|R^{-1}z_2\| \geqq \|R\|^{-1} > 0$. So

$$\operatorname{cond}(T)^2 \geqq \frac{\|T^*TR^{-1}z_2\|}{\|T^*TR^{-1}z_1\|} = \frac{\|RS^*Sz_2\|}{\|RS^*Sz_1\|}$$
$$\geqq \frac{\lambda_2\|Rz_2\|}{\lambda_1\|Rz_1\|} = \frac{\lambda_2}{\lambda_1} = \frac{1+\cos\phi_{\min}}{1-\cos\phi_{\min}} = \frac{(1+\cos\phi_{\min})^2}{\sin^2\phi_{\min}},$$

and the proof is complete.

The preceding theorem sheds some light on the numerical problem of computing minimal factorizations of rational matrix functions. Consider the realization (4.1). We assume that the realization is minimal. From Theorem 2.2 we know that there is a one-to-one correspondence between the minimal factorizations of $W(\lambda)$ and the supporting projections of the (minimal) system $\theta = (A, B, C; \mathbb{C}^\delta, \mathbb{C}^n)$. In turn these supporting projections are completely determined by pairs of subspaces M, M^x satisfying

$$(4.4)\qquad AM\subset M,\quad A^xM^x\subset M^x,\quad \mathbb{C}^\delta = M\oplus M^x.$$

Here, as usual, $A^x = A - BC$.

For the computation of invariant subspaces of a matrix, reliable algorithms are available in the literature [25]. A common way to proceed is to construct a unitary matrix Q_1 such that

$$(4.5)\qquad A_S = Q_1^*AQ_1 = \begin{bmatrix} \alpha_1 & * & \cdots & * \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & * \\ 0 & \cdots & 0 & \alpha_\delta \end{bmatrix}$$

is in upper Schur form. The diagonal elements $\alpha_1, \cdots, \alpha_\delta$ of A_S are the poles of $W(\lambda)$. Similarly, one can construct a unitary matrix Q_2 which transforms A^x to lower Schur form:

$$(4.6)\qquad Q_2^*A^xQ_2 = A_S^x = \begin{bmatrix} \beta_1 & 0 & \cdots & 0 \\ * & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ * & \cdots & * & \beta_\delta \end{bmatrix}.$$

Here $\beta_1, \cdots, \beta_\delta$ are the zeros of $W(\lambda)$. Algorithms that perform these decompositions are known as the QR and QL algorithms [25].

Given natural numbers δ_1 and δ_2 for which $\delta = \delta_1 + \delta_2$, we partition Q_1 and Q_2 as follows,

$$Q_1 = [\underbrace{V_1}_{\delta_1} \vdots \underbrace{W_1}_{\delta_2}],\qquad Q_2 = [\underbrace{V_2}_{\delta_1} \vdots \underbrace{W_2}_{\delta_2}].$$

From (4.5) and (4.6) it is clear that the columns of V_1 and W_2 form orthonormal bases for invariant subspaces M and M^x of A and A^x, respectively. Now $\mathbb{C}^\delta = M\oplus M^x$ if and only if the minimal angle $\phi_{\min}$ between M and M^x is nonzero. Thus M and M^x satisfy (4.4) if and only if $\phi_{\min} > 0$. By Theorem 3.7 we have $\cos\phi_{\min} = \|V_1^*W_2\|$. Therefore, defining the matrix Q by $Q = Q_1^*Q_2$ and partitioning it as follows

$$Q = \begin{array}{c} \begin{bmatrix} Q_{11} & \vdots & Q_{12} \\ \cdots & & \cdots \\ \underbrace{Q_{21}}_{\delta_1} & \vdots & \underbrace{Q_{22}}_{\delta_2} \end{bmatrix} \end{array} \begin{array}{l} \}\delta_1 \\ \\ \}\delta_2, \end{array}$$

one can measure $\phi_{\min}$ from the block $Q_{12} = V_1^*W_2$. Indeed, whenever the norm of Q_{12} is smaller than one, the spaces M and M^x yield a supporting projection of the system θ, and, consequently, a minimal factorization $W(\lambda) = W_1(\lambda)W_2(\lambda)$ of $W(\lambda)$. Observe that δ_1 and δ_2 are the McMillan degrees of W_1 and W_2, respectively.

In order to determine the factors W_1 and W_2 we put $T = [V_1 \vdots W_2]$. If $\mathbb{C}^\delta = M\oplus M^x$, the matrix T is invertible. But then the system $(T^{-1}AT, T^{-1}B, CT; \mathbb{C}^\delta, \mathbb{C}^n)$ is similar to the system θ and has $W(\lambda)$ as its transfer function. One easily verifies that the matrices $T^{-1}AT$, $T^{-1}B$, and CT admit a partitioning of the following type

$$T^{-1}AT = \begin{bmatrix} A_1 & \vdots & B_1C_2 \\ \cdots & & \cdots \\ \underbrace{0}_{\delta_1} & \vdots & \underbrace{A_2}_{\delta_2} \end{bmatrix} \begin{array}{l} \}\delta_1 \\ \\ \}\delta_2, \end{array}\quad T^{-1}B = \begin{bmatrix} B_1 \\ \cdots \\ B_2 \end{bmatrix} \begin{array}{l} \}\delta_1 \\ \\ \}\delta_2, \end{array}\quad CT = [\underbrace{C_1}_{\delta_1} \vdots \underbrace{C_2}_{\delta_2}].$$

Put $\theta_1 = (A_1, B_1, C_1; \mathbb{C}^{\delta_1}, \mathbb{C}^n)$ and $\theta_2 = (A_2, B_2, C_2; \mathbb{C}^{\delta_2}, \mathbb{C}^n)$. Then $\theta = \theta_1\theta_2$. The factors W_1 and W_2 are now the transfer functions of θ_1 and θ_2, respectively. The poles of W_1 are the first δ_1 diagonal entries in A_S; the zeros of W_1 are the first δ_1 diagonal entries in $A_S^\times$. A similar remark can be made about W_2.

The smaller cond (T) is, the more accurate the constructed factors W_1 and W_2 will be. This shows the significance of Theorem 3.7. Indeed, the similarity transformation T constructed above maps $\mathbb{C}^{\delta_1} \oplus (0)$ onto M and $(0) \oplus \mathbb{C}^{\delta_2}$ onto $M^\times$. By Theorem 3.7, a lower bound for the condition number of a transformation having this property is given by the number

$$\frac{1 + \cos \phi_{\min}}{\sin \phi_{\min}}.$$

In the present situation cond (T) is actually equal to this bound, for V_1 and W_2 are isometries. So in this respect T is optimal. On the other hand, for a very small angle $\phi_{\min}$, the condition number of T will be very large. In that case one can expect a very bad relative accuracy that may even exceed 1. This will occur whenever $\phi_{\min}$ is smaller than a certain threshold ϕ_0 which depends on the accuracy of the data. Therefore the spaces M and $M^\times$ cannot be used when their minimal angle is too small. If that happens, one can try different choices of δ_1 and δ_2, while using the same matrices Q_1 and Q_2. Also one can try other Schur decompositions of A and $A^\times$.

For the amount of computations involved in the construction of a factorization of the transfer function (4.1), we can give the following rough estimates, where 1 operation stands for 1 multiplication plus 1 addition:

$\delta^2 n$	operations for constructing $A^\times$,
$20\delta^3$	operations for each Schur decomposition,
δ^3	operations for the product $Q = Q_1^* Q_2$,
$\frac{10}{3}\delta_1^2\delta_2$	operations (if $\delta_1 < \delta_2$) for calculating $\cos \phi_{\min}$,
$2\delta^2(n + \delta)$	operations for computing A_1, B_1, C_1, A_2, B_2 and C_2 if $\phi_{\min} > \phi_0$.

In general, the Schur decompositions constitute the most time-consuming step, but for $\delta = 100$, e.g., experiments have yielded run times that are still within the orders of seconds [9].

As we have seen, the determination of minimal factorizations is closely related to that of pairs of "matching" invariant subspaces. The number of invariant subspaces involved may be very large or even infinite. In practice this may lead to very cumbersome combinatorial problems. For more details, the reader is referred to [21].

REFERENCES

[1] N. I. AHIEZER AND I. M. GLAZMAN, *Theorie der Linearen Operatoren im Hilbertraum*, Akademie Verlag, Berlin, 1960.

[2] S. BARNETT, *Introduction to Mathematical Control Theory*, Clarendon Press, Oxford, 1975.

[3] H. BART, I. GOHBERG AND M. A. KAASHOEK, *Operator polynomials as inverses of characteristic functions*, Integral Equations Operator Theory, 1 (1978), pp. 1–12.

[4] ———, *Stable factorizations of monic matrix polynomials and stable invariant subspaces*, Integral Equations Operator Theory, 1 (1978), pp. 496–517.

[5] ———, *Minimal factorizations of matrix and operator functions*, to appear.

[6] V. BELEVITCH, *Classical Network Theory*, Holden-Day, San Francisco–Cambridge–Amsterdam, 1968.

[7] M. S. BRODSKIĬ, *Triangular and Jordan representations of linear operators*, Transl. Math. Monographs, Vol. 32, American Mathematical Society, Providence RI, 1970.

[8] S. CAMPBELL AND J. DAUGHTRY, *The stable solutions of quadratic matrix equations*, Proc. Amer. Math. Soc., 74 (1979), pp. 19–23.

[9] EISPACK (Guide), *Matrix eigensystem routines*, 2nd Ed., Springer-Verlag, Berlin, 1976.

[10] I. C. GOHBERG AND A. I. FELDMAN, *Convolution equations and projection methods for their solution*, Transl. Math. Monographs, Vol. 41, American Mathematical Society, Providence RI, 1974.

[11] I. C. GOHBERG AND M. G. KREIN, *Systems of integral equations on a half line with kernels depending on the difference of arguments*, Uspehi Mat. Nauk, 13 (1958), pp. 3–72 (in Russian) translated as Amer. Math. Soc. Transl., (2) 14 (1960), pp. 217–287.

[12] I. GOHBERG, P. LANCASTER AND L. RODMAN, *Spectral analysis of matrix polynomials I. Canonical forms and divisors*, Linear Algebra Appl., 20 (1978), pp. 1–44.

[13] ——, *Spectral analysis of matrix polynomials II. The resolvent form and spectral divisors*, Linear Algebra Appl., 21 (1978), pp. 65–88.

[14] I. C. GOHBERG AND A. S. MARKUS, *Two theorems on the gap between subspaces of a Banach space*, Upsehi Mat. Nauk, 14 (1959), pp. 135–140 (in Russian).

[15] I. C. GOHBERG AND E. I. SIGAL, *An operator generalization of the logarithmic residue theorem and the theorem of Rouché*, Mat. Sbornik, 84 (1971), pp. 607–629 (in Russian), transl. as Math. USSR Sbornik, 13 (1971), pp. 603–625.

[16] R. E. KALMAN, P. L. FALB AND M. A. ARBIB, *Topics in Mathematical System Theory*, McGraw-Hill, New York, 1966.

[17] M. G. KREIN, *Introduction to the geometry of indefinite J-spaces and to the theory of operators in these spaces*, Amer. Math. Soc. Transl., (2) Vol. 93, 1970.

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

[19] B. P. MOLINARI, *Equivalence relations for the algebraic Riccati equation*, SIAM J. Control, 11 (1973), pp. 272–285.

[20] L. A. SAHNOVIČ, *On the factorization of an operator-valued transfer function*, Soviet Math. Dokl., 17 (1976), pp. 203–207.

[21] P. VAN DOOREN AND P. DEWILDE, *Minimal factorization of rational matrices*, Proceedings of the 17th IEEE Conference on Decision Control, 1979, pp. 170–172.

[22] P. DEWILDE, *Cascade Scattering Matrix Synthesis*, Tech. Rep. 6560-21. Information System Lab., Stanford University, Stanford CA, 1970.

[23] P. DEWILDE AND J. VANDEWALLE, *On the factorization of a non-singular rational matrix*, IEEE Trans. Circuits and Systems, CAS-22 (1975), pp. 637–645.

[24] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.

[25] J. H. WILKINSON AND C. REINSCH, *Handbook for Automatic Computation, Vol. II: Linear Algebra*, Springer-Verlag, New York, 1971.

[26] D. C. YOULA AND P. TISSI, *An explicit formula for the degree of a rational matrix*, Electrophysics Memo. PIBM RI-1273-65, Polytechnic Institute of Brooklyn, Electrophysics Department, Brooklyn NY, 1965.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 6, November 1980

0363-0129/80/1806-0010 $01.00/0

ON THE OPTIMAL STOPPING TIME PROBLEM FOR DEGENERATE DIFFUSIONS*

J. L. MENALDI†

Abstract. In this paper we give a characterization of the optimal cost of a stopping time problem as the maximum solution of a variational inequality without coercivity. Some properties of continuity for the optimal cost are also given.

Introduction. Summary of main results. This article develops the proofs of results obtained in Note [12].

A. Bensoussan and J. L. Lions [3] have introduced the variational inequality approach in order to solve the optimal stopping time problem in the case of nondegenerate diffusions. A. Friedman [8] treated the same case, M. Robin [18] the optimal stopping time problem for Feller processes, and J. M. Bismut [4] the same problem for a class of more general processes. C. Bardos [1] studied partial differential equations of first order, M. I. Freidlin [7] degenerate elliptic equations, and N. V. Krylov [9] nonlinear degenerate elliptic equations.

In [14] and [17] the variational inequality associated with the deterministic optimal stopping time problem is considered, and in [11] the degenerate nonlinear variational equalities are also studied.

In this paper, the case of degenerate variational inequality associated with the optimal stopping time problem for diffusion processes is developed combining analytic and probabilistic methods.

Let $(\Omega, \mathscr{F}, P)$ be a probability space and $\{\mathscr{F}^t\}_{t\geqq 0}$ be a nondecreasing right-continuous family of completed sub-σ-fields of $\mathscr{F}$.

Now let $y(t)=y_x(t,\omega)$, $t\geqq 0$, $\omega\in\Omega$ be the diffusion on $\mathbb{R}^N$ with Lipschitz continuous coefficients $g(\cdot)$ and $\sigma(\cdot)$, starting at x.

Suppose that $\mathcal{O}$ is an open subset of $\mathbb{R}^N$, and that $\tau=\tau_x(\omega)$ is the first exit time of process $y(t)$ from $\bar{\mathcal{O}}$.

Next, let $f(\cdot)$, $\psi(\cdot)$ be real bounded measurable functions on $\bar{\mathcal{O}}$, and θ be any stopping time. The cost functional $J_x(\theta)$ is given by

$$J_x(\theta)=E\left\{\int_0^{\theta\wedge\tau} f(y(t))\,e^{-\alpha t}\,dt+1_{\theta<\tau}\psi(y(\theta))\,e^{-\alpha\theta}\right\}, \tag{0.1}$$

where α is a positive constant.

Our purpose is to characterize the optimal cost

$$\hat{u}(x)=\inf\{J_x(\theta)/\theta \text{ stopping time}\}, \tag{0.2}$$

and to obtain an optimal stopping time.

We denote by A_0 the second order differential operator associated with the Ito equation[1]

$$A_0=-\tfrac{1}{2}\operatorname{tr}\left(\sigma\sigma^*\frac{\partial^2}{\partial x}\right)-g\frac{\partial}{\partial x}, \tag{0.3}$$

and $A=A_0+\alpha$.

* Received by the editors June 15, 1979, and in final revised form March 3, 1980.

† Université de Paris IX (Dauphine), Paris, France.

[1] If B is a matrix, then B^* denotes the transpose of B and tr (B) the trace of B.

We define Γ_0 as the set of regular points

$$\Gamma_0=\{x\in\partial\mathcal{O}/P(\tau_x>0)=0\}, \tag{0.4}$$

and we give the following integral formulation of the operator A, inspired by D. W. Stroock and S. R. S. Varadhan [19], for any real bounded measurable function on $\bar{\mathcal{O}}$, u and v.

$Au\leqq v$ in $\bar{\mathcal{O}}\backslash\Gamma_0$ if the process

$$X_t=\int_0^{t\wedge\tau} v(y(s))\,e^{-\alpha s}\,ds+u(y(t\wedge\tau))\,e^{-\alpha(t\wedge\tau)} \tag{0.5^2}$$

is a strong submartingale for each $x\in\bar{\mathcal{O}}\backslash\Gamma_0$.

Finally, we introduce the problem: To find a real bounded measurable function on $\bar{\mathcal{O}}$, $u(x)$ such that

$$\begin{aligned} u&=0 \quad \text{on } \Gamma_0,\\ u&\leqq\psi \quad \text{in } \bar{\mathcal{O}}\backslash\Gamma_0,\\ Au&\leqq f \quad \text{in } \bar{\mathcal{O}}\backslash\Gamma_0. \end{aligned} \tag{0.6}$$

We obtain the following characterization.

THEOREM 0.1. *Assume that g, σ are Lipschitz continuous and that f, ψ are Borel measurable and bounded. Suppose also*

$$\psi(x)\geqq 0 \quad \forall x\in\Gamma_0, \quad \psi \text{ upper semicontinuous.} \tag{0.7}$$

Then, the problem (0.6) *has a maximum solution $\hat{u}$ given explicitly by* (0.2). *Moreover, if ψ is continuous, the stopping time $\hat{\theta}=\hat{\theta}_x$ defined by*

$$\hat{\theta}=\inf\{t\in[0,\tau]/\hat{u}(y(t))=\psi(y(t))\} \tag{0.8}$$

is optimal.

We have also the following regularity result.

THEOREM 0.2. *Let the assumptions be as in Theorem* 0.1. *Suppose that*

$$\Gamma_0 \text{ is a closed set.} \tag{0.9}$$

Then if the functions f and ψ are upper semicontinuous (*or continuous*) *the optimal cost $\hat{u}$ is also upper semicontinuous* (*or continuous*).

Now in order to use the variational inequality approach, we assume that the open set $\mathcal{O}$ is bounded, with smooth boundary Γ verifying

$$\Gamma=\{x\in\Gamma/|\sigma(x)n(x)|>0\}\cup\{x\in\Gamma/2g(x)n(x)<-\operatorname{tr}[\sigma\sigma^*(x)]\}, \tag{0.10}$$

where $n(x)$ denotes the inner normal. We remark that (0.10) implies $\Gamma_0=\Gamma$ (cf. D. W. Stroock and S. R. S. Varadhan [19]).

Denote by $(\cdot,\cdot)$ the duality between $H^{-1}(\mathcal{O})$ and $H_0^1(\mathcal{O})$, and by A the differential operator (0.3).

Let us consider the following degenerate variational inequality associated with the stopping time problem

$$\begin{aligned} &u\in H_0^1(u), \qquad u\leqq\psi\\ &(Au, v-u)\geqq(f, v-u) \quad \forall v\in H_0^1(\mathcal{O}), \quad v\geqq\psi. \end{aligned} \tag{0.11}$$

[2] $t\wedge\tau$ denotes the minimum between t and τ.

We have

THEOREM 0.3. *Let the assumptions be as in Theorem* 0.1. *Suppose that f, ψ are Lipschitz continuous, and that conditions* (0.10),

$$A\psi \in L^\infty(\mathcal{O}), \tag{0.12}$$

are satisfied.[3] *Then, there exists one and only one solution u of the variational inequality* (0.11) *which is given as the optimal cost* (0.2). *Moreover, the solution u is Lipschitz continuous and verifies* (0.12).

Remark 0.1. A weak formulation of the variational inequality (0.11) is also considered, and the case of an unbounded domain $\mathcal{O}$ is studied.

This work is divided into four sections. The first section gives some useful lemmas. In § 2 we study the penalized problem, and in § 3 we solve the initial problem. Finally, in the last section, we treat the variational inequality.

1. Preliminary lemmas. Let $(\Omega, \mathcal{F}, P)$ be a probability space, $\{\mathcal{F}^t\}_{t\geqq 0}$ be a nondecreasing right continuous family of completed sub-σ-fields of $\mathcal{F}$, and $w(t)$ be a Brownian motion in $\mathbb{R}^N$ with respect to $\mathcal{F}^t$.

Suppose we are given two Lipschitz continuous functions $g(x)$ and $\sigma(x)$ on $\mathbb{R}^N$, taking values in $\mathbb{R}^N$ and $\mathbb{R}^N \otimes \mathbb{R}^N$ respectively, $g=(g_i)$, $\sigma=(\sigma_{ij})$,

$$\frac{\partial g_i}{\partial x_k}, \frac{\partial \sigma_{ij}}{\partial x_k} \in B(\mathbb{R}^N), \qquad i, j, k = 1, \cdots, N. \tag{1.1^4}$$

We consider the diffusion $y(t) = y_x(t, \omega)$, $t \geqq 0$, $\omega \in \Omega$, and $x \in \mathbb{R}^N$ described by the Ito equation

$$\begin{aligned} &dy(t) = g(y(t))\,dt + \sigma(y(t))\,dw(t), \qquad t \geqq 0,\\ &y(0) = x. \end{aligned} \tag{1.2}$$

We have

LEMMA 1.1. *Suppose* (1.1), *and let θ be any stopping time with respect to $\mathcal{F}^t$. Then there exists a constant γ depending only on the Lipschitz constants of g and σ such that*

$$E\{|y_x(\theta) - y_{x'}(\theta)|^2 e^{-\gamma\theta}\} \leqq |x - x'|^2 \quad \forall x, x' \in \mathbb{R}^N. \tag{1.3}$$

Proof. We set

$$\begin{aligned} \gamma = \sup\Bigg\{ &\operatorname{tr}\left[\frac{(\sigma(x)-\sigma(x'))(\sigma(x)-\sigma(x'))^*}{|x-x'|^2}\right]\\ &+ \frac{2(x-x')(g(x)-g(x'))}{|x-x'|^2} \Bigg/ x, x' \in \mathbb{R}^N\Bigg\}. \end{aligned} \tag{1.4}$$

Then Ito's formula applied to the function $|x|^2 e^{-\gamma t}$ and the process $y_x(t) - y_{x'}(t)$ gives

$$\begin{aligned} |y_x(t) - y_{x'}(t)|^2 e^{-\gamma t} \leqq |x - x'|^2 + 2\int_0^t &(y_x(s) - y_{x'}(s))\\ &\cdot [\sigma(y_x(s)) - \sigma(y_{x'}(s))]\, e^{-\gamma s}\, dw(s). \end{aligned} \tag{1.5}$$

Hence (1.3) follows. □

[3] We also assume α large enough, $\mathcal{O}$ bounded, and Γ smooth.

[4] $B(\mathbb{R}^N)$ denotes the set of all Borel measurable and bounded functions on $\mathbb{R}^N$ taking values in $\mathbb{R}$.

Remark 1.1. Using the martingale inequality

$$E\left\{\sup_{t \geqq 0} \left| \int_0^t \phi(s)\, dw(s) \right| \right\} \leqq 3E\left\{ \sqrt{\int_0^t \phi^2(s)\, ds} \right\}, \tag{1.6}$$

and the same technique as in Lemma 1.1, we can obtain

$$E\{\sup_{t \geqq 0} |y_x(t) - y_{x'}(t)|^k e^{-\gamma t}\} \leqq C|x - x'|^k \quad \forall x, x' \in \mathbb{R}^N, \tag{1.7}$$

where the constants γ and C depend only on $k > 0$ and on the Lipschitz constants of $g(x)$ and $\sigma(x)$. □

Now let $\tau = \tau_x(\omega)$ and $\tau'_x = \tau'(\omega)$ be the first exit time of the process $y(t)$ from the closed set $\bar{\mathcal{O}}$ and the open set $\mathcal{O}$ respectively,

(1.8)[5]
$$\tau = \inf\{t \geqq 0 / y(t) \notin \bar{\mathcal{O}}\},$$

and a similar definition for τ' with $\mathcal{O}$ instead of $\bar{\mathcal{O}}$.

We have

LEMMA 1.2. *Suppose* (1.1). *Then, for any constant* $T > 0$ *and* $x \in \mathbb{R}^N$, *we have*

(1.9)[6]
$$\lim_{z \to x} E\{(T \wedge \tau'_x - T \wedge \tau'_z)^+\} = 0,$$

$$\lim_{z \to x} E\{(T \wedge \tau_z - T \wedge \tau_x)^+\} = 0. \tag{1.10}$$

Proof. Let z_n be a sequence, $z_n \to x$, and let us consider the diffusions $y_n(t)$, $y(t)$ starting respectively at z_n, x. Using Lemma 1.1, we can assume that

$$\lim \sup_{0 \leqq t \leqq T} |y_n(t) - y(t)| = 0 \quad \text{a.s.}$$

In order to obtain (1.9), we will prove

$$\underline{\lim}\, \tau'_n \geqq \tau' \quad \text{a.s.} \tag{1.11}$$

We assume $\omega \in \Omega$ fixed. Then, if $\tau' = 0$, (1.11) is clearly verified, and so we can suppose $0 < \delta < \tau'$ and define the set $K_\omega = \{y(t)/t \in [0, \delta]\}$ which is a compact subset of $\mathcal{O}$. Hence for n large enough, $n \geqq N_\omega$,

$$\{y_n(t)/t \in [0, \delta]\} \subset \mathcal{O}.$$

Thus $\tau'_n \geqq \delta$ and taking the limit,

$$\underline{\lim}\, \tau'_n \geqq \delta;$$

since δ is arbitrary, we deduce (1.11).

Now we are going to prove

$$\overline{\lim}\, \tau_n \leqq \tau \quad \text{a.s.}, \tag{1.12}$$

so that (1.10) holds.

We assume $\omega \in \Omega$ fixed. Then, if $\tau = \infty$, (1.12) is clearly verified, so we can assume $\delta > \tau$. Hence, there exists $s < \delta$ such that $y(s) \notin \bar{\mathcal{O}}$. Thus for n large enough, $y_n(s) \notin \bar{\mathcal{O}}$, so $\tau_n \leqq s < \delta$, and taking the limit

$$\underline{\lim}\, \tau_n \leqq \delta;$$

since δ is arbitrary, we deduce (1.12). □

[5] $\tau = +\infty$ if $y(t) \in \bar{\mathcal{O}}\ \forall t \geqq 0$.

[6] If $a \in \mathbb{R}$ we denote by a^+ the maximum between a and zero.

Remark 1.2. From E. B. Dynkin [6, Theorem 10.2, p. 302] it follows that either the process $y_x(t)$ stopped at the exit of $\mathcal{O}$, or $\bar{\mathcal{O}}$ is a strong Markov process. Also observe that τ and τ' are stopping times with respect to the family $\mathcal{F}^t$.

D. W. Stroock and S. R. S. Varadhan [19] proved Lemma 1.2 in a different way.

Remark 1.3. We recall the following martingale property: Let $a(t)$ and $b(t)$ be measurable adapted and bounded processes, such that

$$M_t = a(t) + \int_0^t b(s)\,ds \quad \text{is a martingale.}$$

Then, for any arbitrary measurable adapted and bounded process $c(t)$, the process

$$a(t) \exp\left(-\int_0^t c(s)\,ds\right) + \int_0^t (b(s)+c(s)a(s)) \exp\left(-\int_0^s c(r)\,dr\right) ds$$

is the martingale

$$M_0 + \int_0^t \exp\left(-\int_0^s c(r)\,dr\right) dM_s.$$

Now, we define the set Γ_0 of regular points (cf. D. W. Stroock and S. R. S. Varadhan [19]), $\Gamma = \partial\mathcal{O}$,

$$\Gamma_0 = \{x \in \Gamma / P(\tau_x > 0) = 0\}. \tag{1.13}$$

We have

LEMMA 1.3. *Assume* (1.1) *and that*

$$\Gamma_0 \text{ is a closed set.} \tag{1.14}$$

Then for any constant $T > 0$, *and* $x \in \mathcal{O}$, *we have*

$$\lim_{z \to x} E\{|T \wedge \tau_z - T \wedge \tau_x|\} = 0, \qquad z \in \bar{\mathcal{O}}. \tag{1.15}$$

Proof. Let $\tilde{\tau} = \tilde{\tau}_x(\omega)$ be the first exit time from $\bar{\mathcal{O}}\backslash\Gamma_0$ of the process $y(t)$. From the strong Markov property of the process $y(t)$ stopped at the exit of $\bar{\mathcal{O}}$, we easily deduce

$$P(\tau \neq \tilde{\tau}) = 0. \tag{1.16}$$

Later on, we will show

$$\lim_{z \to x} E\{(T \wedge \tilde{\tau}_x - T \wedge \tilde{\tau}_z)^+\} = 0, \qquad z \in \bar{\mathcal{O}}. \tag{1.17}$$

Indeed, we assume $\omega \in \Omega$ fixed and the notations of Lemma 1.2 with $\tilde{\tau}$ instead of τ'. Then, without loss of generality, we suppose $0 < \delta < \tilde{\tau}$, and we define the set $K_\omega = \{y(t)/t \in [0, \delta]\}$, which is a compact subset of $\mathcal{O}$ such that $K_\omega \cap \Gamma_0 = \emptyset$. Because of (1.14), for n sufficiently large, $n \geqq N_\omega$, we have

$$\{y_n(t)/t \in [0, \delta]\} \subset \bar{\mathcal{O}}\backslash\Gamma_0.$$

Thus $\tilde{\tau}_n \geqq \delta$, and taking the limit we obtain

$$\underline{\lim}\, \tilde{\tau}_n \geqq \tilde{\tau} \quad \text{a.s.}$$

So, (1.17) follows.

Finally, by combining (1.16), (1.17) and (1.10) the lemma is proved. □

Remark 1.4. In D. W. Stroock and S. R. S. Varadhan [19] it is proved that, assuming (1.14), we have $\tau_x = \tau'_x$ a.s. for each $x \in \mathcal{O} \cup \Gamma_0$. Then we deduce Lemma 1.3 for the particular case $x \in \mathcal{O} \cup \Gamma_0$. Notice that Lemma 1.3 implies that the process $y(t \wedge \tau)$ is Feller continuous on the whole domain $\bar{\mathcal{O}}$.

Let us consider the differential operator A given by (0.3) where α is a constant large enough, $2\alpha \geqq \gamma$, defined in Lemma 1.1.

LEMMA 1.4. *Suppose* (1.1). *Let* $f(x)$, $\psi(x)$, *and* $\bar{u}(x)$ *be continuous real function on* $\bar{\mathcal{O}}$ *such that*

$$\begin{aligned} &\bar{u} \in C(\bar{\mathcal{O}}), \qquad \frac{\partial \bar{u}}{\partial x_i} \in L^\infty(\mathcal{O}), \qquad i = 1, \cdots, N, \\ &\bar{u} \leqq \psi \textit{ in } \mathcal{O}, \qquad \bar{u}(x) = \mathcal{O} \quad \forall x \in \Gamma_0, \\ &A\bar{u} \leqq -|f| \quad \textit{in } \mathcal{D}'(\mathcal{O}). \end{aligned} \tag{1.18}$$

Then for any nonnegative, bounded and adapted process $\delta(t) = \delta(t, \omega)$, *the following estimate holds*

$$\begin{aligned} &E\left\{ \int_{\tau_x \wedge \tau_{x'}}^{\tau_x} (|f(y_x(t))| - \delta(t)\psi(y_x(t))) \exp\left(-\int_0^t (\alpha - \delta(s))\, ds\right) dt \right\} \\ &\qquad \leqq \left\| \frac{\partial \bar{u}}{\partial x} \right\| |x - x'| \quad \forall x, x' \in \bar{\mathcal{O}} \end{aligned} \tag{1.19}$$

where $\|\partial \bar{u}/\partial x\|$ *denotes the smallest Lipschitz continuous constant of the function* $\bar{u}$.

Proof. First suppose $\bar{u} \in C^2(\mathcal{O})$. Ito's formula applied to the function $\bar{u}(x)$ and the process $y_x(t)$ gives

$$\begin{aligned} E\Big\{ &\bar{u}(y_x(\tau_x)) \exp\left(-\int_0^{\tau_x} (\alpha + \delta(t))\, dt\right) - \bar{u}(y_x(\tau_x \wedge \tau_{x'})) \\ &\cdot \exp\left(-\int_0^{\tau_x \wedge \tau_{x'}} (\alpha + \delta(t))\, dt\right)\Big\} \\ &= -E\left\{ \int_{\tau_x \wedge \tau_{x'}}^{\tau_x} [(A\bar{u})(y_x(t)) + \delta(t)\bar{u}(y_x(t))] \exp\left(-\int_0^\tau (\alpha + \delta(s))\, ds\right) dt \right\}. \end{aligned} \tag{1.20}$$

Using

$$\bar{u}(y_x(\tau_x)) = 0 = \bar{u}(y_{x'}(\tau_x \wedge \tau_{x'})) \quad \text{a.s. in } [\tau_{x'} \leqq \tau_x < \infty],$$

from (1.20) we have

$$\begin{aligned} &E\left\{ \int_{\tau_x \wedge \tau_{x'}}^{\tau_x} (|f(y(t))| - \delta(t)\psi(y_x(t))) \exp\left(-\int_0^t (\alpha + \delta(s))\, ds\right) dt \right\} \\ &\qquad \leqq E\{|\bar{u}(y_x(\tau_x \wedge \tau_{x'})) - \bar{u}(y_{x'}(\tau_x \wedge \tau_{x'}))|\, e^{-\alpha(\tau_x \wedge \tau_{x'})}\}. \end{aligned} \tag{1.21}$$

Next, choosing $\theta = \tau_x \wedge \tau_{x'}$ in Lemma 1.1, we deduce from (1.21) the estimate (1.19).

Finally, if $\bar{u} \notin C^2(\mathcal{O})$, by approximating $\bar{u}$ by regular functions the lemma is proved. □

Remark 1.5. Clearly, Lemma 1.4 implies

$$E\{|e^{-\alpha\tau_x} - e^{-\alpha\tau_{x'}}|\} \leqq 2\alpha \left\| \frac{\partial u_0}{\partial x} \right\| |x - x'|, \qquad x, x' \in \bar{\mathcal{O}}, \tag{1.22}$$

if $2\alpha \geqq \gamma$ and u_0 is a Lipschitz continuous function on $\bar{\mathcal{O}}$, vanishing on Γ_0, such that $Au_0 \leqq -1$ in $\mathcal{D}'(\mathcal{O})$.

Remark 1.6. For instance, suppose $\mathcal{O}$ is a bounded domain given by

$$(1.23)\qquad \begin{aligned} &\mathcal{O} = \{x \in \mathbb{R}^N / \phi(x) < 0\}, &&\phi \in C^2, \\ &\Gamma = \{x \in \mathbb{R}^N / \phi(x) = 0\}, &&|\nabla \phi(x)| \geqq 1 \quad \forall x \in \Gamma, \end{aligned}$$

and assume

$$(1.24)\qquad A_0 \phi \leqq -1 \quad \text{on } \Gamma.$$

Then for any continuous functions f and ψ on $\bar{\mathcal{O}}$, $\psi \in C^2(\bar{\mathcal{O}})$, $\psi \geqq 0$ on Γ; we can take $\bar{u} = \lambda \phi$, which verifies (1.18) for α and λ large enough. Clearly, applying Ito's formula to the function $\bar{u}$ and process $y(t)$ between τ' and τ, we deduce $\Gamma_0 = \Gamma$.

Now, some sufficient conditions for the existence of a Lipschitz continuous subsolution are given using barrier functions as in [11].

LEMMA 1.5. *Assume* (1.1). *Suppose also that $\mathcal{O}$ is bounded, has the uniform exterior sphere property*[7]

(1.25) *There exist $\rho > 0$ such that for each point $\xi \in \Gamma$ there is a ball $B = B(\xi^*, \rho)$ of radius ρ and center ξ^* verifying $\bar{\mathcal{O}} \cap B = \{\xi\}$,*

and

$$(1.26)\qquad \begin{aligned} &\Gamma = \{x \in \Gamma / |\sigma(x) n(x)| > 0\} \cup \{x \in \Gamma / 2g(x) n(x) < -\operatorname{tr}[\sigma\sigma^*(x)]\}, \\ &n(x) \textit{ is the inner normal of modulus } \rho. \end{aligned}$$

Then $\Gamma_0 = \Gamma$, *and there exists a Lipschitz continuous subsolution* $u_0(x)$

$$(1.27)\qquad \begin{aligned} &u_0 \in C(\bar{\mathcal{O}}), \qquad \frac{\partial u_0}{\partial x_i} \in L^\infty(\bar{\mathcal{O}}), \qquad i = 1, \cdots, N, \\ &Au_0 \leqq -1 \textit{ in } \mathcal{D}'(\mathcal{O}), \qquad u_0 = 0 \quad \textit{on } \Gamma. \end{aligned}$$

Proof. It is necessary to prove only (1.27).

Introducing the barrier functions $k > 0$, $\xi \in \Gamma$, $x \in \bar{\mathcal{O}}$,

$$v(x, \xi) = \exp(-k|x - \xi^*|^2) - \exp(-k\rho^2),$$

we have from (1.26) $A_0 v(x, \xi) \leqq -2\beta < 0$, if $x = \xi$ and k is sufficiently large independent of ξ. Hence, by continuity, we have for some $\delta > 0$,

$$(1.28)\qquad A_0 v(x, \xi) \leqq -\beta < 0 \quad \forall x \in U_\xi = \{x \in \mathcal{O} / |x - \xi| < \delta\};$$

now using the fact that $v(x, \xi) \leqq -\gamma < 0$ $\forall x \in \bar{\mathcal{O}} \backslash U_\xi$, er deduce, for α large enough,

$$(1.29)\qquad Av(x, \xi) \leqq -\beta < 0 \quad \forall x \in \bar{\mathcal{O}}.$$

Finally, remarking that $v(x, \xi)$ are equi-Lipschitz continuous in $x \in \bar{\mathcal{O}}$, we set

$$(1.30)\qquad u_0(x) = \frac{1}{\beta} \sup \{v(x, \xi) / \xi \in \Gamma\}.$$

Hence, $Au_0 \leqq -1$ in the martingale sense (0.5) and in the distribution sense. □

Remark 1.7. Suppose u_0 given as in Lemma 1.5. Then for any f, $\psi \in C(\bar{\mathcal{O}})$ such that

$$(1.31)\qquad \psi \geqq 0 \text{ on } \Gamma \quad \text{and} \quad A\psi \in L^\infty(\mathcal{O}),$$

and taking $\bar{u} = \lambda u_0$, where $\lambda \geqq \|f\| + \|A\psi\|$,[8] we deduce (1.18).

[7] The constant α is supposed large enough.

[8] $\|\cdot\|$ denotes the L^∞-norm in $\mathcal{O}$.

Remark 1.8. Clearly, using other barrier functions, different sufficient conditions for the existence of a Lipschitz continuous subsolutions may be obtained.

2. Penalized problem. Before studying the stopping time problem we will start with an intermediate stochastic control problem.

We call an admissible control ν a scalar measurable adapted process such that $0 \leqq \nu(t;\omega) \leqq 1, t \geqq 0$.

Let $f(x)$, $\psi(x)$ be functions such that

$$f, \psi \in B(\bar{\mathcal{O}}), \tag{2.1}$$

and let α be a positive constant. We define the functional J_x^ε, $\varepsilon > 0$,

$$J_x^\varepsilon(\nu) = E\Big\{\Big[\Big[\int_0^\tau f(y(t)) + \frac{1}{\varepsilon}\nu(t)\psi(y(t))\Big] \exp\Big(-\int_0^t \Big(\alpha + \frac{1}{\varepsilon}\nu(s)\Big)\, ds\Big)\, dt\Big\}, \tag{2.2}$$

and we wish to characterize the optimal penalized cost,

$$u_\varepsilon(x) = \inf\{J_x^\varepsilon(\nu)/\nu \text{ any admissible control}\}. \tag{2.3}$$

The integral formulation of operator A (cf. D. W. Stroock and S. R. S. Varadhan [19]) is given for $u, v \in B(\bar{\mathcal{O}})$ by

$Au = v$ in $\bar{\mathcal{O}} \backslash \Gamma_0$ if the process

$$X_t = \int_0^{t \wedge \tau} v(y(s))\, e^{-\alpha s}\, ds + u(y(t \wedge \tau))\, e^{-\alpha(t \wedge \tau)} \tag{2.4}$$

is a martingale for each $x \in \bar{\mathcal{O}} \backslash \Gamma_0$.

We remark that if $Au = v$ in the sense of (2.4), then we also have $Au = v$ in the distribution in $\mathcal{O}$ for σ smooth.

Next, the following problem is considered: To find a function $u_\varepsilon(x)$ such that

$$u_\varepsilon \in B(\bar{\mathcal{O}}), \qquad u_\varepsilon(x) = 0 \quad \forall x \in \Gamma_0, \tag{2.5}$$

$$Au_\varepsilon = f - \frac{1}{\varepsilon}(u_\varepsilon - \psi)^+ \quad \text{in } \bar{\mathcal{O}} \backslash \Gamma_0 \text{ [in the martingale sense].} \tag{2.6}$$

Remark 2.1. Let $\Phi(t)$ be the semigroup in $B(\bar{\mathcal{O}})$ given by

$$\Phi(t)h = E\{h(y(t \wedge \tau))\, e^{-\alpha(t \wedge \tau)}\}, \tag{2.7}$$

and $\mathscr{X}$ be the characteristic function of the set $\bar{\mathcal{O}} \backslash \Gamma_0$.

Then, using the strong Markov property of process $y(t)$ stopped at the exit of $\mathcal{O}$, we show that (2.5) and (2.6) and the condition $u_\varepsilon \in B(\bar{\mathcal{O}})$,

$$u_\varepsilon = \int_0^t \Phi(s)\Big(f\mathscr{X} - \frac{1}{\varepsilon}(u_\varepsilon - \psi)^+\mathscr{X}\Big)\, ds + \Phi(t)(u_\varepsilon \mathscr{X}) \quad \forall t \geqq 0, \tag{2.8}$$

are equivalent. Moreover, the condition

$$u_\varepsilon = \int_0^\infty \Phi(t)\Big(f\mathscr{X} - \frac{1}{\varepsilon}(u_\varepsilon - \psi)^+\mathscr{X}\Big)\, dt \tag{2.9}$$

is also equivalent to (2.8).

Remark 2.2. The semigroup formulation (2.8) is used by A. Bensoussan [2] for the nondegenerate case. Here, if we assume that the set of regular points Γ_0 is closed, the stopping process is Feller continuous (because of Lemma 1.3). Then, a semigroup formulation can also be studied as in M. Robin [18].

This section is divided in three parts. First we solve problem (2.5), (2.6). Next, we consider the case where the set of regular points Γ_0 is closed. Finally, we give some complementary results.

2.1. Existence and semicontinuity results. We have

THEOREM 2.1. *Assume* (1.1) *and* (2.1). *Then problem* (2.5), (2.6) *has one and only one solution* u_ε *which is given by* (2.3).

Proof. First we prove that problem (2.5), (2.6) has one and only one solution $w_\varepsilon(x)$. Indeed, from the equality

$$-\frac{1}{\varepsilon}(w_\varepsilon-\psi)^+=-\frac{1}{\varepsilon}w_\varepsilon+\frac{1}{\varepsilon}(w_\varepsilon\wedge\psi),$$

and applying Remark 1.3 for

$$a(t)=w_\varepsilon(y(\tau\wedge\tau))\,e^{-\alpha(t\wedge\tau)},\qquad c(t)=\frac{1}{\varepsilon},$$

$$b(t)=\begin{cases} f(y(t))\,e^{-\alpha t} & \text{if } t\leq\tau,\\ 0 & \text{otherwise,}\end{cases}$$

we deduce that the conditions (2.5), (2.6) are equivalent to (2.5),

$$\text{(2.10)}[9]\qquad \left(A+\frac{1}{\varepsilon}\right)w_\varepsilon=f+\frac{1}{\varepsilon}(w_\varepsilon\wedge\psi).$$

So, using the strong Markov property, we only need to find a unique solution of the equation,

$$\text{(2.11)}\qquad w_\varepsilon=E\left\{\int_0^\tau\left[f(y(t))+\frac{1}{\varepsilon}(w_\varepsilon\wedge\psi)(y(t))\right]\exp\left(-\alpha t-\frac{1}{\varepsilon}t\right)dt\right\}.$$

Thus, we define the operator T_ε in $B(\bar{\mathcal{O}})$ by

$$\text{(2.12)}\qquad T_\varepsilon w=E\left\{\int_0^\tau\left[f(y(t))+\frac{1}{\varepsilon}(w\wedge\psi)(y(t))\right]\exp\left(-\alpha t-\frac{1}{\varepsilon}t\right)dt\right\},$$

and we have[10]

$$\left\|T_\varepsilon v-T_\varepsilon w\right\|\leqq\frac{1}{1+\alpha\varepsilon}\left\|v-w\right\|.$$

Hence, T_ε is a contraction in $B(\bar{\mathcal{O}})$ and so the equation (2.11) has one and only one solution.

Next, we are going to show that the solution of problem (2.5), (2.6) is given by (2.3). Indeed, let w_ε be the solution for (2.5), (2.6). Then using Remark 1.3 with $\delta(t)=(1/\varepsilon)\nu(t)$, $\nu(t)$ any admissible control, we obtain

$$w_\varepsilon=E\left\{\int_0^\tau\left[f-\frac{1}{\varepsilon}(w_\varepsilon-\psi)^++\frac{1}{\varepsilon}\nu(t)w_\varepsilon\right](y(t))\exp\left(-\int_0^t\left(\alpha+\frac{1}{\varepsilon}\nu(s)\right)ds\right)dt\right\}.$$

[9] w_ε instead of u_ε.

[10] $\|\cdot\|$ denotes the supremum norm in $\bar{\mathcal{O}}$.

Since

$$-(w_\varepsilon - \psi)^+ + \nu w_\varepsilon \leqq \nu\psi \quad \text{if } 0 \leqq \nu \leqq 1,$$

we have

$$w_\varepsilon(x) \leqq J_x^\varepsilon(\nu), \quad \nu \text{ any admissible control,} \tag{2.13}$$

and for

$$\hat\nu(t) = \begin{cases} 1 & \text{if } w_\varepsilon(y(t)) > \psi(y(t)), \\ 0 & \text{if } w_\varepsilon(y(t)) \leqq \psi(y(t)), \end{cases}$$

$$w_\varepsilon(x) = J_x^\varepsilon(\hat\nu). \tag{2.14}$$

Thus, (2.13) and (2.14) give $w_\varepsilon = u_\varepsilon$. □

Remark 2.3. If u_ε and $\tilde u_\varepsilon$ denote the functions given by (2.3) with f, ψ and $\tilde f$, $\tilde\psi$ respectively, the following estimate is true,

$$\|u_\varepsilon - \tilde u_\varepsilon\| \leqq \frac{1}{\alpha}\|f - \tilde f\| + \|\psi - \tilde\psi\|, \tag{2.15}$$

where $\|\cdot\|$ denotes the norm of supremum over $\bar{\mathcal{O}}$.

It is possible to consider the case with τ' instead of τ and to obtain analogous results.

Now we study properties of continuity on u_ε. We have

THEOREM 2.2. *Let the conditions* (1.1), (2.1) *hold. Then if f and ψ are nonnegative upper semicontinuous on $\bar{\mathcal{O}}$, so is u_ε defined by* (2.3).

Proof. Letting T be a positive constant, we define

$$J_x^\varepsilon(\nu, T) = E\left\{\int_0^{T\wedge\tau} \left[f(y)(t)) + \frac{1}{\varepsilon}\nu(t)\psi(y(t))\right] \exp\left(-\int_0^t \left(\alpha + \frac{1}{\varepsilon}\nu(s)\right) ds\right) dt\right\} \tag{2.16}$$

and

$$u_\varepsilon^T(x) = \inf\{J_x^\varepsilon(\nu, T)/\nu \text{ any admissible control}\}. \tag{2.17}$$

We have the estimate

$$\|u_\varepsilon^T - u_\varepsilon\| \leqq \left(\frac{1}{\alpha}\|f\| + \|\psi\|\right) e^{-\alpha T}. \tag{2.18}$$

So it is sufficient to consider u_ε^T instead of u_ε.

Then, we start with

$$u_\varepsilon^T(z) - u_\varepsilon^T(x) \leqq \sup\{[J_z^\varepsilon(\nu, T) - J_x^\varepsilon(\nu, T)]/\nu \text{ any admissible control}\}.$$

Next it follows that

$$\begin{aligned} u_\varepsilon^T(z) - u_\varepsilon^T(x) \leqq {} & \left(\|f\| + \frac{1}{\varepsilon}\|\psi\|\right) E\{(T\wedge\tau_z - T\wedge\tau_x)^+\} \\ & + E\left\{\int_0^{T\wedge\tau_z\wedge\tau_x} [f(y_z(t)) - f(y_x(t))]^+ e^{-\alpha t}\, dt\right\} \\ & + \frac{1}{\varepsilon} E\left\{\int_0^{T\wedge\tau_z\wedge\tau_x} [\psi(y_z(t)) - \psi(y_x(t))]^+ e^{-\alpha t}\, dt\right\}. \end{aligned} \tag{2.19}$$

Thus taking the limit in (2.19) and using (1.3), (1.10), the theorem is proved. □

Remark 2.4. Let $u'_\varepsilon(x)$ be the optimal cost in the open set $\mathcal{O}$; that is, u'_ε is defined by (2.3) with τ' instead of τ. Then a similar theorem of regularity is proved: If f and ψ are nonnegative lower semicontinuous on $\bar{\mathcal{O}}$, so is u'_ε. Notice that the function u'_ε is the solution of (2.5), (2.6) with Γ, τ' instead of Γ_0, τ.

2.2. Regular case. In this part we assume that

$$\text{(2.20)} \qquad \Gamma_0 \text{ given by (1.13) is a closed set,}$$

so we have

THEOREM 2.3. *Suppose* (1.1), (2.1), *and* (2.20) *hold. Then if f and ψ are upper* (*lower*) *semicontinuous on $\bar{\mathcal{O}}$, so is u_ε given by* (2.3).

Proof. The proof is similar to Theorem 2.2 from

$$\begin{aligned} u_\varepsilon^T(z)-u_\varepsilon^T(x) \leqq & \left\{\|f\|+\frac{1}{\varepsilon}\|\psi\|\right\}E\{|T\wedge\tau_z-T\wedge\tau_x|\} \\ & +E\left\{\int_0^{T\wedge\tau_z\wedge\tau_x}[f(y_z(t))-f(y_x(t))]^+\,e^{-\alpha t}\,dt\right\} \\ & +\frac{1}{\varepsilon}E\left\{\int_0^{T\wedge\tau_z\wedge\tau_x}[\psi(y_z(t))-\psi(y_z(t))]^+\,e^{-\alpha t}\,dt\right\}; \end{aligned}$$

using (1.3) and (1.15) gives the result. □

Remark 2.5. Let $\mathcal{O}$ be smooth and $n(x)$ be the inner normal of boundary $\Gamma=\partial\mathcal{O}$. Suppose that

$$\text{(2.21)} \qquad \begin{aligned} &\sigma(x)=0 \qquad \forall x\in\mathbb{R}^N\backslash\mathcal{O}, \\ &g(x)n(x)\geqq 0 \quad \forall x\in\Gamma; \end{aligned}$$

then $\Gamma_0=\varnothing$, so (2.20) is true. Clearly, if $\mathcal{O}=\mathbb{R}^N$, (2.20) can be removed.

Now we are going to obtain some a priori estimates.

THEOREM 2.4. *Assume* (1.1), (2.1), $\mathcal{O}=\mathbb{R}^N$, *and*

$$\text{(2.22)} \qquad \frac{\partial f}{\partial x_i}, \frac{\partial \psi}{\partial x_i}\in L^\infty(\mathbb{R}^N), \qquad i=1,\cdots,N.$$

Then u_ε is Lipschitz continuous and verifies

$$\text{(2.23)[11]} \qquad \left\|\frac{\partial u_\varepsilon}{\partial x}\right\|\leqq\frac{1}{\alpha-\gamma_0}\left\|\frac{\partial f}{\partial x}\right\|+\left\|\frac{\partial \psi}{\partial x}\right\|.$$

Proof. Let T_ε be the operator defined by (2.12). From Theorem 2.1, u_ε is the fixed point of the contraction T_ε. Suppose w is a Lipschitz continuous function on $\mathbb{R}^N$, and denote $\alpha_0=\alpha-\frac{1}{2}\gamma>0$; then from (1.3) it follows that

$$\text{(2.24)[12]} \qquad \left\|\frac{\partial T_\varepsilon w}{\partial x}\right\|\leqq\frac{\varepsilon}{1+\varepsilon\alpha_0}\left\|\frac{\partial f}{\partial x}\right\|+\frac{1}{1+\varepsilon\alpha_0}\left(\left\|\frac{\partial w}{\partial x}\right\|\vee\left|\frac{\partial \psi}{\partial x}\right\|\right).$$

Thus, (2.24) implies

$$\left\|\frac{\partial T_\varepsilon^k w}{\partial x}\right\|\leqq\varepsilon\left\|\frac{\partial f}{\partial x}\right\|\sum_{i=1}^k(1+\varepsilon\alpha_0)^{-i}+\frac{1}{1+\varepsilon\alpha_0}\left(\left\|\frac{\partial w}{\partial x}\right\|\vee\left\|\frac{\partial \psi}{\partial x}\right\|\right).$$

[11] γ is given by (1.4), and $\|\partial f/\partial x\|$ denotes the smallest Lipschitz of f.

[12] If $a, b\in R$, then $a\vee b$ denotes the maximum between a and b.

Hence

$$\left\|\frac{\partial T_\varepsilon^k w}{\partial x}\right\| \leqq \frac{1}{\alpha_0}\left\|\frac{\partial f}{\partial x}\right\| + \left(\left\|\frac{\partial w}{\partial x}\right\| \vee \left\|\frac{\partial \psi}{\partial x}\right\|\right), \tag{2.25}$$

and taking $w = 0$ and letting $k \to \infty$ in (2.25) we prove (2.23). □

THEOREM 2.5. *Let the assumptions* (1.1) *and* (2.1) *hold. Suppose that there exists a Lipschitz continuous subsolution, i.e.,*

$$\begin{aligned} &\bar{u} \in C(\bar{\mathcal{O}}), \qquad \frac{\partial u}{\partial x_i} \in L^\infty(\mathcal{O}), \qquad i = 1, \cdots, N, \\ &\bar{u} \leqq \psi \text{ in } \mathcal{O}, \qquad \bar{u}(x) = 0 \quad \forall x \in \Gamma_0, \\ &A\bar{u} \leqq -|f| \quad \mathit{in}\ \mathcal{D}'(\mathcal{O}), \end{aligned} \tag{2.26}$$

and

$$\frac{\partial f}{\partial x_i}, \ \frac{\partial \psi}{\partial x_i} \in L^\infty(\mathcal{O}), \qquad i = 1, \cdots, N. \tag{2.27}$$

Then u_ε is Lipschitz continuous on $\bar{\mathcal{O}}$, and verifies

$$\left\|\frac{\partial u_\varepsilon}{\partial x}\right\| \leqq \frac{1}{\alpha - \gamma}\left\|\frac{\partial f}{\partial x}\right\| + \left\|\frac{\partial \psi}{\partial x}\right\| + \left\|\frac{\partial \bar{u}}{\partial x}\right\|. \tag{2.28}$$

Proof. Starting at

$$u_\varepsilon(x) - u_\varepsilon(x') = \sup_{\nu'} \inf_{\nu} \, [J_x^\varepsilon(\nu) - J_{x'}^\varepsilon(\nu)],$$

and taking

$$\nu(t) = \begin{cases} \nu'(t) & \text{if } t \in [0, \tau_x \wedge \tau_{x'}], \\ 0 & \text{otherwise,} \end{cases}$$

we have

$$\begin{aligned} u_\varepsilon(x) - u_\varepsilon(x') \leqq E\left\{\int_0^{\tau_x \wedge \tau_{x'}} |f(y_x(t)) - f(y_{x'}(t))|\, e^{-\alpha t}\, dt\right\} \\ + \sup_{\nu'} E\left\{\int_0^{\tau_x \wedge \tau_{x'}} \frac{1}{\varepsilon}\, \nu'(t) |\psi(y_x(t)) - \psi(y_{x'}(t))| \right. \\ \exp\left(-\int_0^t \left(\alpha + \frac{1}{\varepsilon}\, \nu'(s)\right) ds\right) dt \\ + \int_{\tau_x \wedge \tau_{x'}}^{\tau_{x'}} \left[f^-(y_{x'}(t)) - \frac{1}{\varepsilon}\nu'(t)\psi(y_{x'}(t))\right] \exp\left(-\int_0^t \left(\alpha + \frac{1}{\varepsilon}\nu'(s)\right) ds\right) dt \\ \left. + \int_{\tau_x \wedge \tau_x}^{\tau_x} f^+(y_x(t)) \exp\left(-\int_0^t \left(\alpha + \frac{1}{\varepsilon}\, \nu'(s)\right) ds\right) dt\right\}. \end{aligned}$$

Next, using Lemmas 1.1 and 1.4 we obtain

$$u_\varepsilon(x) - u_\varepsilon(x') \leqq \left[\frac{1}{\alpha - \gamma_0}\left\|\frac{\partial f}{\partial x}\right\| + \left\|\frac{\partial \psi}{\partial x}\right\| + \left\|\frac{\partial \bar{u}}{\partial x}\right\|\right] |x - x'|. \tag{2.29}$$

Clearly, from (2.29) the theorem is proved. □

Remark 2.6. Notice that condition (2.26) implies (2.20). Indeed, from Remark 1.5, the function $x \to E\{\exp(-\alpha\tau_x)\}$ is continuous on $\bar{\mathcal{O}}$. Then, using the fact that $\Gamma_0 = \{x \in \bar{\mathcal{O}}/E\{\exp(-\alpha\tau_x)\} = 1\}$, we reach our conclusions.

2.3. Complementary results. Now, we consider u_ε as a distribution in $\mathcal{O}$.

Let A be the differential operator

$$A = -\tfrac{1}{2}\operatorname{tr}\left(\sigma\sigma^* \frac{\partial^2}{\partial x^2}\right) - g\frac{\partial}{\partial x} + \alpha. \tag{2.30}$$

Assume

$$\frac{\partial^2 \sigma\sigma^*}{\partial x^2} \in L^1_{\mathrm{loc}}(\mathcal{O}). \tag{2.31}$$

So we can define Au for $u \in B(\bar{\mathcal{O}})$, as a distribution in $\mathcal{O}$, by

$$\langle Au, \phi\rangle = \int_{\mathcal{O}} uA^*\phi \, dx \quad \forall \phi \in \mathcal{D}(\mathcal{O}), \tag{2.32}$$

where A^* is the operator

$$A^*\phi = -\tfrac{1}{2}\operatorname{tr}\left(\frac{\partial^2 \sigma\sigma^*\phi}{\partial x}\right) + g\frac{\partial\phi}{\partial x} + \alpha\phi. \tag{2.33}$$

Then we have

THEOREM 2.6. *Let the conditions* (1.1), (2.1), *and* (2.3) *hold. Suppose that the boundary* Γ *is smooth. Then the optimal cost* u_ε *given by* (2.3) *satisfies*

$$Au_\varepsilon + \frac{1}{\varepsilon}(u_\varepsilon - \psi)^+ = f \quad \text{in } \mathcal{D}'(\mathcal{O}). \tag{2.34}$$

Moreover, if

(2.35)[13] *there exists* $w \in B(\bar{\mathcal{O}})$ *such that* $A\psi = w$ *in* $\bar{\mathcal{O}}\backslash\Gamma_0$,

$$\psi(x) \geqq 0 \quad \forall x \in \Gamma_0, \tag{2.36}$$

the following estimate is true.

$$\|Au_\varepsilon\| \leqq \|f\| + \|(f - A\psi)^+\|. \tag{2.37}$$

Proof. Equation (2.34) is obtained by regularization, or as in D. W. Stroock and S. R. S. Varadhan [19] using an argument of monotone class. In order to get (2.37) we will show that

$$\|(u_\varepsilon - \psi)^+\| \leqq \varepsilon\|(f - A\psi)^+\|. \tag{2.38}$$

Indeed, from (2.35) and Remark 1.3 we have

$$\psi = E\left\{\int_0^\tau \left[A\psi(y(t)) + \frac{1}{\varepsilon}\psi(y(t))\right] \exp\left(-\alpha t - \frac{1}{\varepsilon}t\right) dt\right\}$$
$$+ E\left\{1_{\tau<\infty}\psi(y(\tau)) \exp\left(-\alpha\tau - \frac{1}{\varepsilon}\tau\right)\right\}.$$

[13] In the martingale sense of (2.4).

Since

$$u_\varepsilon-\psi=E\left\{\int_0^\tau[f(y(t))-A\psi(y(t))]\exp\left(-\alpha t-\frac{1}{\varepsilon}t\right)dt\right\}$$
$$-\frac{1}{\varepsilon}E\left\{\int_0^\tau[\psi(y(t))-u_\varepsilon(y(t))]^+\exp\left(-\alpha t-\frac{1}{\varepsilon}t\right)dt\right\}$$
$$-E\left\{1_{\tau<\infty}\psi(y(\tau))\exp\left(-\alpha\tau-\frac{1}{\varepsilon}\tau\right)\right\},$$

and because $y(\tau)\in\Gamma_0$ a.s. if $\tau<\infty$, we obtain

$$u_\varepsilon-\psi\leqq\|(f-A\psi)^+\|E\left\{\int_0^\tau\exp\left(-\alpha t-\frac{1}{\varepsilon}t\right)dt\right\}.$$

Hence, (2.38) follows. □

Remark 2.7. Notice that (2.38) remains true even if Γ is not smooth. Also, if, for instance, $\psi\in C(\Gamma_0)$ and $A\psi\in L^\infty(\mathcal{O})$, then from D. W. Stroock and S. R. S. Varadhan [19] the assumption (2.35) is satisfied.

Remark 2.8. A result analogous to Theorem 2.6 can be proved for the optimal cost u'_ε in the open set $\mathcal{O}$.

We also have monotonicity in ε.

THEOREM 2.7. *Assume* (1.1) *and* (2.1). *Then if* $0<\varepsilon\leqq\varepsilon'$ *we obtain*

$$u_\varepsilon\leqq u_{\varepsilon'}. \tag{2.39}$$

Proof. Let T_ε be the operator introduced in Theorem 2.1 by (2.12). First, we are going to prove that

$$T_\varepsilon u_{\varepsilon'}\leqq u_{\varepsilon'}\quad\text{if } 0<\varepsilon\leqq\varepsilon'. \tag{2.40}$$

Indeed, as in Theorem 2.6, we obtain for any $u\in B(\bar{\mathcal{O}})$ which satisfies (2.35)[14] and vanishes on Γ_0,

$$T_\varepsilon u-u=E\left\{\int_0^\tau\left[f-Au-\frac{1}{\varepsilon}(u-\psi)^+\right](y(t))\exp\left(-\alpha t-\frac{1}{\varepsilon}t\right)dt\right\}. \tag{2.41}$$

So using the equality

$$f-Au_{\varepsilon'}-\frac{1}{\varepsilon}(u_{\varepsilon'}-\psi)^+=\left(\frac{1}{\varepsilon'}-\frac{1}{\varepsilon}\right)(u_{\varepsilon'}-\psi)^+,$$

and taking $u=u_\varepsilon$ in (2.41), we deduce (2.40).

Further, knowing that T_ε has the monotone property (if $u\leqq u'$ then $T_\varepsilon u\leqq T_\varepsilon u'$), from (2.40) we obtain

$$T_\varepsilon^k u_{\varepsilon'}\leqq u_{\varepsilon'}. \tag{2.42}$$

Hence, taking the limit in (2.42) as $k\to\infty$, we prove (2.39). □

Remark 2.8. As for Theorem 2.7, an analogous property is obtained for the optimal cost u'_ε in the open case.

Remark 2.9. Approximating u_ε by regular functions (cf. D. W. Stroock and S. R. S. Varadhan [19, Coroll. 8.1], we have

$$t\to u_\varepsilon(y(t\wedge\tau))\text{ is a.s. continuous.} \tag{2.43}$$

The same argument holds for functions ψ satisfying (2.35).

[14] Clearly, with u instead of ψ.

Also, using the semigroup associated with the process $y(t)$ stopped at the exit from the open set $\mathcal{O}$, we prove

$$t \to u'_\varepsilon(y(t \wedge \tau')) \text{ is a.s. right continuous,} \tag{2.44}$$

where u'_ε denote the optimal cost in the open case.

3. Integral formulation. Recall that Γ_0 denotes the set of regular points given by (0.4) and that if $u, v \in B(\bar{\mathcal{O}})$ we set

$Au \leqq v$ in $\bar{\mathcal{O}} \backslash \Gamma_0$ if the process

$$X_t = \int_0^{\tau \wedge \tau} v(y(s))\, e^{-\alpha s}\, ds + u(y(\tau \wedge \tau))\, e^{-\alpha(\tau \wedge \tau)} \tag{3.1}$$

is a strong submartingale[15] for each $x \in \bar{\mathcal{O}} \backslash \Gamma_0$.

The following problem is considered: Find $u(x)$ such that

$$u \in B(\bar{\mathcal{O}}), \qquad u(x) = 0 \quad \forall x \in \Gamma_0, \tag{3.2}$$

$$Au \leqq f \quad \text{in } \bar{\mathcal{O}} \backslash \Gamma_0 \text{ [in the martingale sense (3.1)],} \tag{3.3}$$

$$u \leqq \psi \quad \text{in } \bar{\mathcal{O}} \backslash \Gamma_0. \tag{3.4}$$

In order to find solutions of problem (3.2), (3.3), (3.4) which have some continuity property, it is necessary to assume that

$$\psi(x) \geqq 0 \quad \forall x \in \Gamma_0. \tag{3.5}$$

This section is divided into three parts. First, we consider the case where ψ is regular. Next, we extend the results for ψ continuous or upper semicontinuous. Finally, we give some complementary results.

3.1. Regular case. We have

THEOREM 3.1. *Let the conditions* (1.1), (2.1), (3.5) *hold. We also assume that*

(3.6) *there exists* $w \in B(\bar{\mathcal{O}})$ *such that* $A\psi = w$ *in* $\bar{\mathcal{O}} \backslash \Gamma_0$ [*martingale sense*].

Then the problem (3.2), (3.3), (3.4) *admits a maximum solution* u *which is given by the decreasing limit*

$$u(x) = \lim_{\varepsilon \downarrow 0} u_\varepsilon(x) \quad \forall x \in \bar{\mathcal{O}}, \tag{3.7}$$

where u_ε *is the solution of problem* (2.5), (2.6).

Proof. Using Theorem 2.7 we can define a function $u(x)$ by the limit (3.7).

First we are going to prove that u, given by (3.7), is a solution or problem (3.2), (3.3), (3.4). Indeed, assertion (3.2) is trival from (2.5) and Remark 2.1. Condition (3.3) is obtained taking the limit in the martingale expression of (2.6), and (3.4) follows from the estimate (2.38).

Next, in order to show that u is the maximum solution, it is only necessary to prove that each solution v of problem (3.2), (3.3), (3.4) satisfies

$$v \leqq u_\varepsilon \quad \text{in } \bar{\mathcal{O}} \quad \forall \varepsilon > 0. \tag{3.8}$$

But, as in Theorem 2.7, (3.8) follows from

$$v \leqq T_\varepsilon v \quad \text{in } \bar{\mathcal{O}}. \tag{3.9}$$

[15] That is, X_t satisfies the Doob optional sampling theorem.

Thus, using Remark 1.3 as in Theorem 2.7, we obtain (3.9), and so the theorem is proved. □

Now, the optimal stopping time problem is considered.

THEOREM 3.2. *Under assumptions* (1.1), (2.1), (3.5), *and* (3.6) *the maximum solution* $\hat{u}$ *of problem* (3.2), (3.3), (3.4) *is also given as the optimal cost* (0.2), *and the estimate*

$$\|u_\varepsilon - \hat{u}\| \leqq \varepsilon \|(f - A\psi)^+\| \quad \forall \varepsilon > 0, \tag{3.10}$$

holds. Moreover, the stopping time $\hat{\theta} = \hat{\theta}_x$ *defined by*

(3.11)[16] $$\hat{\theta} = \inf\{t \in [0, \tau] / \hat{u}(y(t)) = \psi(y(t))\}$$

is optimal; i.e.,

$$\hat{u}(x) = J_x(\hat{\theta}). \tag{3.12}$$

Proof. Denote by $\hat{u}$ the optimal cost (0.2), and by u_ε the solution of the penalized problem (2.5), (2.6).

First we are going to show that

$$u_\varepsilon \geqq \hat{u} \quad \forall \varepsilon > 0, \tag{3.13}$$

(3.14)[16] $$\hat{\theta}^\varepsilon = \inf\{t \in [0, \tau] / u_\varepsilon(y(t)) \geqq \psi(y(t))\}$$

satisfies

(3.15)[17] $$\begin{aligned} &1_{\hat{\theta}^\varepsilon \wedge \tau < \infty}\, u_\varepsilon(y(\hat{\theta}^\varepsilon \wedge \tau)) = 1_{\hat{\theta}^\varepsilon < \tau} \psi(y(\hat{\theta}^\varepsilon)), \\ &u_\varepsilon(y(t)) < \psi(y(t)) \quad \text{if } t \in [0, \hat{\theta}^\varepsilon]. \end{aligned}$$

Note that (2.6) implies

$$u_\varepsilon = E\left\{\int_0^{\theta \wedge \tau} \left[f - \frac{1}{\varepsilon}(u_\varepsilon - \psi)^+\right](y(t))\, e^{-\alpha t}\, dt + 1_{\theta \wedge \tau < \infty} u_\varepsilon(y(\theta \wedge \tau))\, e^{-\alpha(\theta \wedge \tau)}\right\}, \tag{3.16}$$

for any stopping time θ. Thus, taking $\theta = \hat{\theta}_\varepsilon$ in (3.16) and regarding (3.15), we deduce

$$u_\varepsilon(x) = J_x(\hat{\theta}^\varepsilon), \tag{3.17}$$

and so (3.13) follows.

Next we are going to prove

$$u_\varepsilon - \hat{u} \leqq \varepsilon \|(f - A\psi)^+\|. \tag{3.18}$$

Indeed, starting at

$$u_\varepsilon(x) - \hat{u}(x) = \sup_\theta \inf_\nu [J_x^\varepsilon(\nu) - J_x(\theta)], \tag{3.19}$$

and setting

$$\nu_\theta(s) = \begin{cases} 1 & \text{if } s > \theta, \\ 0 & \text{if } s \leqq \theta, \end{cases}$$

[16] With $\hat{\theta} = \tau$ or $\hat{\theta}^\varepsilon = \tau$ if the corresponding set is empty.

[17] $1_{a<b}$ denotes the function $= 1$ if $a < b$ and $= 0$ otherwise.

we deduce, as in Theorem 2.6,

$$\begin{aligned} J_x^\varepsilon(\nu_\theta) - J_x(\theta) = -E\left\{1_{\tau<\infty}1_{\theta<\tau}\psi(y(\tau)) \exp\left(-\alpha\tau - \frac{\tau-\theta}{\varepsilon}\right)\right\} \\ +E\left\{\int_{\theta_\wedge\tau}^{\tau} (f-A\psi)(y(t)) \exp\left(-\alpha t - \frac{t-\theta\wedge\tau}{\varepsilon}\right) dt\right\}. \end{aligned} \tag{3.20}$$

Hence, using (3.5) from (3.20) and (3.19), we have (3.18).

Clearly, (3.18) and (3.13) imply (3.10). So we obtain from (2.43),

$$t \to \hat{u}(y(t \wedge \tau)) \text{ is a.s. continuous.} \tag{3.21}$$

Further, from Theorem 2.7, (3.21), and estimate (3.10), we have

$$\lim_{\varepsilon \downarrow 0} \hat{\theta}^\varepsilon = \hat{\theta} \quad \text{a.s.,} \tag{3.22}$$

where the limit is increasing.

Finally, choosing $\theta = \hat{\theta}_{\varepsilon'}$, $\varepsilon' > \varepsilon > 0$ in (3.16), and letting $\varepsilon \to 0$ and then $\varepsilon' \to 0$, and using the convergence (3.10), (3.22) we establish (3.12). □

3.2. Nonregular case. Now, we relax the regularity assumptions on ψ. Without assuming (3.6), ψ will be only continuous or upper semicontinuous. We have

THEOREM 3.3. *Under assumptions* (1.1), (2.1), (3.5), *and*

$$\psi \textit{ is uniformly continuous on } \bar{\mathcal{O}}, \tag{3.23}$$

the problem (3.2), (3.3), *and* (3.4) *admits a maximum solution $\hat{u}$ which is given as the optimal cost* (0.2). *Moreover,*

$$\lim_{\varepsilon \downarrow 0} \|u_\varepsilon - \hat{u}\| = 0, \tag{3.24}$$

and the relation (3.12) *is true.*

Proof. First we remark that if u_i denotes the optimal cost (0.2) corresponding to f_i, ψ_i for $i = 1, 2$, we immediately obtain the estimate,

$$\|\hat{u}_1 - \hat{u}_2\| \leqq \frac{1}{\alpha}\|f_1 - f_2\| + \|\psi_1 - \psi_2\|. \tag{3.25}$$

Next, notice that in Theorem 3.1 the assumption (3.6) was used only in order to prove (3.4). Also, the same arguments as in Theorem 3.2 show that provided (3.25) and (3.24) hold, we can deduce (3.12). So, using the fact that $\hat{u}$ defined by (0.2) satisfies (3.4), we just need to prove the convergence (3.24). Then, approximating ψ by a sequence of smooth functions and using the estimates (3.25) and (2.15) the convergence (3.24) is established. □

Remark 3.1. If the obstacle ψ is only continuous, the assertions of Theorem 3.3 remain true but the convergence (3.24) holds only on compact sets of $\bar{\mathcal{O}}$,

THEOREM 3.4. *Let the conditions* (1.1), (2.1), (3.5), *and*

$$\psi \textit{ upper semicontinuous on } \bar{\mathcal{O}} \tag{3.26}$$

hold. The problem (3.2), (3.3), (3.4) *admits a maximum solution $\hat{u}$ which is given as the optimal cost* (0.2). *Moreover, given any constant $\varepsilon > 0$ there exists a function $\hat{\theta}^\varepsilon = \hat{\theta}^\varepsilon(x)$ such that*

$$\begin{aligned} &\hat{\theta}^\varepsilon : \bar{\mathcal{O}} \times \Omega \to [0, \infty] \textit{ is measurable,} \\ &\forall x \in \bar{\mathcal{O}}, \quad \hat{\theta}^\varepsilon(x) \textit{ is a stopping time,} \end{aligned} \tag{3.27}$$

and

$$\hat{u}(x)+\varepsilon \geqq J_x(\tilde{\theta}^\varepsilon(x) \quad \forall x \in \bar{\mathcal{O}}. \tag{3.28}$$

Proof. Since ψ is bounded and upper semicontinuous on $\bar{\mathcal{O}}$, there exists a sequence $\{\psi_k\}_{k=1}^{\infty}$ of bounded and continuous functions on $\bar{\mathcal{O}}$ decreasing to ψ (cf. Bourbaki [5, p. 30]). Let $\hat{u}$ and $\hat{u}_k$ be the optimal costs according to ψ and ψ_k respectively; then clearly, $\hat{u}_k$ is decreasing to $\hat{u}$.

Next, from Theorem 3.4 and Remark 3.1, the functions $\hat{u}_k$ verify (3.2), (3.3), and

$$\hat{u}_k \leqq \psi_k. \tag{3.29}$$

So, if we let $k \to \infty$, the function $\hat{u}$ satisfies (3.4). Moreover, from monotonicity, $\hat{u}$ is the maximum solution of (3.2), (3.3), (3.4).

Finally, we set

$$k_\varepsilon(x) = \inf\{k \geqq 1/\hat{u}_k(x) \leqq \hat{u}(x)+\varepsilon\}, \tag{3.30}$$

and

$$\tilde{\theta}^\varepsilon = \inf\{t \in [0, \tau]/\hat{u}_{k_\varepsilon}(y(t)) = \psi_{k_\varepsilon}(y(t))\}. \tag{3.31}$$

It is easy to check that $\tilde{\theta}^\varepsilon$ satisfies (3.27), (3.28), and the proof is completed. □

Now, using Theorem 3.4, Theorem 2.7 and Theorem 2.2, we obtain

COROLLARY 3.1. *Let the conditions* (1.1), (2.1), *and* (3.5) *hold. Then if f and ψ are nonnegative upper semicontinuous on $\bar{\mathcal{O}}$, so is the optimal cost $\hat{u}$ defined by* (0.2).

Next, using Remark 3.1, Theorem 3.4, and Theorem 2.3, we obtain

COROLLARY 3.2. *Assume* (1.1), (2.1), (2.20), *and* (3.5). *Then if f and ψ are upper semicontinuous or* (*continuous*) *on $\bar{\mathcal{O}}$, so is the optimal cost $\hat{u}$ defined by* (0.2).

Remark 3.2. With suitable modification in the proofs, results similar to Theorem 3.1, Theorem 3.2, Theorem 3.3 and Corollary 3.1 are obtained for the optimal cost u' in the case of the open set $\mathcal{O}$.

3.3. Complementary results. A relation between the two problems, in the closed set $\bar{\mathcal{O}}$ and in the open set $\bar{\mathcal{O}}$, is given by

THEOREM 3.5. *Let the conditions* (1.1) *and* (2.1) *hold. Then the following estimates hold,*

$$\|(\hat{u}'-\hat{u})^+\| \leqq \frac{1}{\alpha}\|f^-\| + \|\psi^-\|, \tag{3.32}$$

$$\|(\hat{u}'-\hat{u})^-\| \leqq \|1_{\Gamma\setminus\Gamma_0}\psi^+\|, \tag{3.33}$$

where $\hat{u}'$ and $\hat{u}$ denote the optimal cost corresponding to the problem in the open subset $\mathcal{O}$ and closed set $\bar{\mathcal{O}}$ respectively.

Proof. Recall that τ' denotes the first exit time of process $y(t)$ from the open subset $\mathcal{O}$, and $J'_x(\theta)$ the functional cost given by (0.1) with τ' instead of τ.

Starting at

$$\hat{u}(x)-\hat{u}(x) = \sup_{\theta} \inf_{\theta'} [J'_x(\theta')-J_x(\theta)], \tag{3.34}$$

and choosing for the infimum $\theta'=\theta$ in (3.34), we deduce

$$\text{(3.35} \qquad \hat{u}'(x)-\hat{u}(x) \leqq E\left\{\int_{\tau'}^{\tau} f^-(y(t)\, e^{-\alpha t}\, dt\right\} + \sup_{\theta} E\{1_{\tau' \leqq \theta < \tau}\psi^-(y(\theta))\, e^{-\alpha\theta}\},$$

and (3.32) follows.

Further, taking from the supremum $\theta = \theta' \wedge \tau'$ in (3.34), we have

$$\hat{u}'(x) - \hat{u}(x) \geqq -E\{1_{\tau'<\tau}\psi^+(y(\tau'))\, e^{-\alpha\tau'}\}. \tag{3.36}$$

Hence (3.33) is proved. □

Next, combining Theorem 3.5, Corollary 3.1 and Remark 3.2, we obtain

COROLLARY 3.3. *Assume* (1.1), (2.1), (3.5), *and*

$$\psi(x) = 0 \quad \forall x \in \Gamma \backslash \Gamma_0. \tag{3.37}$$

Then if f and ψ are nonnegative continuous on $\bar{\mathcal{O}}$, the two optimal costs $\hat{u}'$ and $\hat{u}$ coincide. It follows from Theorem 2.2 *and Remark* 2.4 *that the optimal cost $\hat{u}$ given by* (0.2) *is continuous on $\bar{\mathcal{O}}$.*

Now, $\hat{u}$ is regarded as a distribution in $\mathcal{O}$. Recalling that A represents the differential operator given by (2.30), we have

THEOREM 3.6. *Suppose that the boundary Γ is smooth and the conditions* (1.1), (2.1), (2.31), (3.5), *and*

$$\psi \text{ continuous on } \bar{\mathcal{O}} \tag{3.38}$$

hold. Then the optimal cost $\hat{u}$ satisfies

$$A\hat{u} \leqq f \quad \text{in } \mathcal{D}'(\mathcal{O}), \tag{3.39}$$

$$A\hat{u} = f \quad \text{in } \mathcal{D}'([\hat{u} < \psi]). \tag{3.40}$$ [18]

Furthermore, if ψ verifies (3.6), *the following estimate is true*

$$\|A\hat{u}\| \leqq \|f\| + \|(f - A\psi)^+\|. \tag{3.41}$$

So $A\hat{u} \in L^\infty(\mathcal{O})$.

Proof. First we recall that the condition (3.40) has meaning if the subset $[\hat{u} < \psi]$ is open. Using Corollary 3.1 and Corollary 3.2 this fact can be deduced.

Next the conditions (3.39) and (3.41) are immediate from Theorems 3.4 and 2.6.

Finally, if $\phi \in \mathcal{D}([\hat{u} \in \psi])$, using the uniform convergence (3.24) we obtain

$$(\hat{u}_\varepsilon - \psi)^+ \phi = 0 \quad \text{if } \varepsilon \text{ is small enough.} \tag{3.42}$$

Therefore, from (3.42) and (2.34) the equality (3.40) is proved. □

Remark 3.3. Let U be the subset of $\mathcal{O}$ where $\sigma(x)$ is nondegenerate. Suppose that $\hat{u}$ is continuous (see Corollary 3.2). Then, from (3.41), $\hat{u}$ can be regarded as the unique solution of a Dirichlet problem on U. This fact leads to a $W^{2,p}_{\text{loc}}(U)$, $1 < p < \infty$, regularity for the optimal cost $\hat{u}$ given by (0.2).

Remark 3.4. All these results can be extended for f and ψ with polynomial growth.

Remark 3.5. It is possible to consider a more general case of a cost functional $J_x(\theta)$, exchanging the term $\exp(-\alpha t)$ with

$$\exp\left(-\int_0^\tau c(y(s))\, ds\right),$$

and adding a final cost

$$1_{\tau<\infty} 1_{\theta \geqq \tau} h(y(\tau)) \exp\left(-\int_0^\tau c(y(t))\, dt\right),$$

provided $c(y) \geqq \alpha_0 > 0$.

[18] $[\hat{u} < \psi]$ denotes the subset of points $x \in \bar{\mathcal{O}}$ such that $\hat{u}(x) < \psi(x)$.

Remark 3.6. A result analogous to Theorem 3.6 is given for the problem in the open set $\mathcal{O}$.

Remark 3.7. All these results can be extended to the parabolic case.

4. Variational inequality. Let $a_{ij}(x)$, $a_i(x)$ be functions for $i, j = 1, \cdots, N$, such that

$$(a_{ij})_{ij} \text{ is a nonnegative symmetric matrix and}$$

$$a_{ij} \in C^1(\mathbb{R}^N), \qquad \frac{\partial^2 a_{ij}}{\partial x_k\, \partial x_l} \in L^\infty(\mathbb{R}^N) \quad \forall i, j, k, l = 1, \cdots, N, \tag{4.1}$$

$$a_i \in C(\mathbb{R}^N), \qquad \frac{\partial a_i}{\partial x_k} \in L^\infty(\mathbb{R}^N) \quad \forall i, k = 1, \cdots, N. \tag{4.2}$$

Define the following differential operator A,

$$A = -\sum_{i,j=1}^N \frac{\partial}{\partial x_i} a_{ij} \frac{\partial}{\partial x_j} + \sum_{i=1}^N a_i \frac{\partial}{\partial x_i} + \alpha, \tag{4.3}$$

where α is a positive constant.

We always identify g and σ given by (1.1) as

$$\begin{aligned} (a_{ij})_{ij} &= \tfrac{1}{2}\sigma\sigma^*, \\ a_i &= \sum_{j=1}^N \frac{\partial a_{ij}}{\partial x_j} - g_i. \end{aligned} \tag{4.4}$$

Let $\beta_0(x)$ and $\beta_1(x)$ be the weight functions $(1+|x|^2)^{-(\lambda+1)/2}$ and $(1+|x|^2)^{-\lambda/2}$, $\lambda > N/2$, respectively. Introduce the following Hilbert spaces:

$$H = \{v / \beta_0 v \in L^2(\mathcal{O})\}, \tag{4.5}$$

with the inner product

$$(u, v) = \int_{\mathcal{O}} (\beta_0 u)(\beta_0 v)\, dx \tag{4.6}$$

and the norm $|\cdot|$;

$$V = \left\{ v \in H / \beta_1 \frac{\partial v}{\partial x_k} \in L^2(\mathcal{O}), \forall k = 1, \cdots, N \right\}, \tag{4.7}$$

with the norm

$$\|v\| = \left(|v|^2 + \sum_{k=1}^N \int_{\mathcal{O}} \left| \beta_1 \frac{\partial v}{\partial x_k} \right|^2 dx \right)^{1/2}. \tag{4.8}$$

V' denotes the dual space of V, and $\langle \cdot, \cdot \rangle$ the duality between V' and V.

We have

$$\begin{gathered} V \subset H \subset V'; L^\infty(\mathcal{O}) \subset H; \\ \left\{ v \middle/ \frac{\partial v}{\partial x_i} \in L^\infty(\mathcal{O}) \ \forall i = 1, \cdots, N \right\} \subset V. \end{gathered} \tag{4.9}$$

Let $a(\cdot, \cdot)$ be the bilinear form associated to the operator A,

$$a(u, v) = \sum_{i,j=1}^N \int_{\mathcal{O}} \tilde{a}_{ij} \left(\beta_1 \frac{\partial u}{\partial x_j} \right) \left(\beta_1 \frac{\partial v}{\partial x_i} \right) dx + \sum_{i=1}^N \int_{\mathcal{O}} \tilde{a}_i \left(\beta_1 \frac{\partial u}{\partial x_i} \right) (\beta_0 v)\, dx + \alpha(u, v), \tag{4.10}$$

where

$$a_{ij}(x)=(1+|x|^2)^{-1}a_{ij}(x),$$
$$\tilde{a}_i(x)=(1+|x|^2)^{-1/2}a_i(x)-2(\lambda+1)(1+|x|^2)^{-3/2}\sum_{j=1}^{N}a_{ij}(x)x_j. \tag{4.11}$$

Notice that a_{ij}, a_i are not supposed to be bounded, but a_{ij} is at most of quadratic growth, and a_i of linear growth. Then $\tilde{a}_{ij}$, $\tilde{a}_i$ in (4.11) are bounded.

This section is divided into three parts. First, we consider the case where $\mathcal{O}=\mathbb{R}^N$. Next, we give a weak formulation. Finally, we study the general case.

4.1. Case $\mathcal{O}=\mathbb{R}^N$. Assume $\mathcal{O}=\mathbb{R}^N$. After some computation we deduce

$$a(u,v)=(Au,v)\quad \forall u,v\in V, Au\in H, \tag{4.12}$$

(4.13)[19] $$|a(u,v)|\leqq C\|u\|\,\|v\|\quad \forall u,v\in V,$$

and if α is large enough there exists $\alpha_0>0$ such that

$$a(u,u)\geqq\alpha_0(u,u)\quad \forall u\in V. \tag{4.14}$$

Next, from (4.12) and (4.13) it follows that

$$a(u,v)=\langle Au,v\rangle,\qquad u,v\in V. \tag{4.15}$$

Now, let K be the following closed cone in V:

$$K=\{v\in V/v(x)\leqq\psi(x)\text{ a.e. in }\mathbb{R}^N\}, \tag{4.16}$$

and let us consider the variational inequality

$$\text{Find } u\in K \text{ such that } a(u,v-u)\geqq(f,v-u)\quad \forall v\in K. \tag{4.17}$$

Recalling the cost functional

$$J_x(\theta)=E\left\{\int_0^\theta f(y(t))\,e^{-\alpha t}\,dt+1_{\theta<\infty}\psi(y(\theta))\,e^{-\alpha\theta}\right\}, \tag{4.18}$$

we have

THEOREM 4.1. *Let the assumptions* (4.1), (4.2), *and*[20]

$$\frac{\partial f}{\partial x_k},\quad \frac{\partial\psi}{\partial x_k}\in L^\infty(\mathbb{R}^N),\qquad k=1,\cdots,N, \tag{4.19}$$

hold. Then there exists one and only one solution u of the variational inequality (4.17). *This solution u is given as the optimal cost,*

$$u(x)=\inf\{J_x(\theta)/\theta \textit{ is a stopping time}\}. \tag{4.20}$$

Moreover, the following estimate is true:

$$\left\|\frac{\partial u}{\partial x}\right\|_{L_\infty}\leqq\frac{1}{\alpha-\gamma_0}\left\|\frac{\partial f}{\partial x}\right\|_{L_\infty}+\left\|\frac{\partial\psi}{\partial x}\right\|_{L_\infty}, \tag{4.21}$$

where $\|\partial u/\partial x\|_{L_\infty}$ *denotes the smallest Lipschitz constant of the function u.*[21]

[19] C denotes a constant.

[20] α is assumed large enough, and f, ψ are not necessarily bounded.

[21] There exists also an optimal stopping time (Theorem 3.2).

Proof. Without loss of generality, we may assume that f, ψ are bounded (Remark 3.4). From (4.14) the uniqueness of the variational inequality (4.17) is obtained by classic methods (cf. A. Bensoussan and J. L. Lions [3]).

Using Theorem 2.6, we have for the optimal penalized cost u_ε given by (2.2),

$$Au_\varepsilon+\frac{1}{\varepsilon}(u_\varepsilon-\psi)^+=f \quad \text{in } \mathcal{D}'(\mathbb{R}^N). \tag{4.22}$$

Thus, from the convergence (3.24) and the estimate (2.23), we can take limits when $\varepsilon\to 0$ in (4.22) for the weak convergence in V, and using the monotonicity of operator A, we obtain (4.17); so the theorem is proved. □

4.2. Weak formulation. In order to give a weak formulation of the variational inequality (4.17) we introduce the Hilbert space D_A which is the closure of the set

(4.23)[22] $$\{v\in V/Av\in H\},$$

with the graph norm

$$\|v\|_{D_A}=(|v|^2+|Av|^2)^{1/2}. \tag{4.24}$$

Using density arguments we also have

$$(Au,u)\geqq \alpha_0(u,u) \quad \forall u\in D_A. \tag{4.25}$$

The following problem is considered,

$$\begin{aligned}&\text{Find } u\in D_A \text{ such that } u\leqq\psi \text{ a.e., and}\\ &(Au,v-u)\geqq(f,v-u) \quad \forall v\in D_A,\ v\leqq\psi \quad \text{a.e.}\end{aligned} \tag{4.26}$$

THEOREM 4.2. *Assume* (4.1), (4.2) *and*[23]

$$f,\psi\in C(\mathbb{R}^N)\cap L^\infty(\mathbb{R}^N), \tag{4.27}$$

$$A\psi\in L^\infty(\mathbb{R}^N). \tag{4.28}$$

Then problem (4.26) *has one and only one solution u which is given as the optimal cost* (4.20). *Moreover, the function u is bounded and continuous, and the following estimate holds*:

$$\|Au\|_{L^\infty}\leqq\|f\|_{L^\infty}+\|(f-A\psi)^+\|_{L^\infty}. \tag{4.29}$$

Proof. Notice that (4.27) and (4.28) imply (Remark 2.7) that

(4.30)[24] There exists $w\in B(\mathbb{R}^N)$ such that $A\psi=w$ in the martingale sense.

So, using Theorem 2.6, we have

$$\|Au_\varepsilon\|_{L^\infty}\leqq\|f\|_{L^\infty}+\|(f-A\psi)^+\|_{L^\infty}, \tag{4.31}$$

and also (Remark 2.3)

$$\|u_\varepsilon\|_{L^\infty}\leqq\frac{1}{\alpha}\|f\|_{L^\infty}+\|\psi\|_{L^\infty}. \tag{4.32}$$

Then we take limits when $\varepsilon\to 0$ in (4.22) as in Theorem 4.1, and the proof is complete. □

[22] A denotes the differential operator (4.3).

[23] α is assumed large enough in order to have (4.25).

[24] In the sense of (2.4), $\mathcal{O}=\mathbb{R}^N$.

Remark 4.1. Under assumption (4.30), Theorem 4.2 remains true for f and ψ upper semicontinuous and bounded instead of (4.27).

Remark 4.2. The problem (4.26) can be interpreted as

$$\begin{aligned} &u \in D_A, \qquad u \leqq \psi \quad \text{a.e.},\\ &Au \leqq f \quad \text{a.e.},\\ &(Au - f)(u - \psi) = 0 \quad \text{a.e.}, \end{aligned} \tag{4.33}$$

using standard methods. Clearly, under assumptions (4.28), (4.19), the weak formulation (4.26) implies the strong formulation (4.17).

4.3. General case. We come back to the general case. Now, $\mathcal{O}$ is an open subset of $\mathbb{R}^N$ with boundary Γ smooth enough. Recalling that the subset of regular point Γ_0 is given by (0.4), we have (cf. D. Stroock and S. R. S. Varadhan [19, p. 686]).

$$\sum_{i=1}^{N} a_i(x) n_i(x) \leqq 0 \quad \forall x \in \Gamma \backslash \Gamma_0, \tag{4.34}$$

where $n(x) = (n_i(x))$ is the inner normal of $\mathcal{O}$.

Next, define the closed subspace of V,

$$V_0 = \{v \in V / v = 0 \text{ on } \Gamma_0\}. \tag{4.35}$$

Then, as in the case $\mathcal{O} = \mathbb{R}^N$, if α is large enough, using (4.34) it is possible to find a constant $\alpha_0 > 0$ such that

$$a(u, u) \geqq \alpha_0(u, u) \quad \forall u \in V_0. \tag{4.36}$$

Furthermore, assuming

$$\sum_{i=1}^{N} a_{ij}(x) n_i(x) = 0 \quad \forall x \in \Gamma \backslash \Gamma_0, \quad j = 1, \cdots, N, \tag{4.37}$$

we deduce

$$a(u, v) = \langle Au, v \rangle \quad \forall u, v \in V_0. \tag{4.38}$$

Remark 4.3. If we assume

$$\sum_{i,j=1}^{N} a_{ij}(x) n_i(x) n_j(x) + \left(\sum_{i=1}^{N} a_i(x) n_i(x) \right)^+ > 0 \quad \forall x \in \Gamma, \tag{4.39}$$

the condition (4.37) is true and $\Gamma = \Gamma_0$.

Setting K_0 the closed cone in V_0,

$$K_0 = \{v \in V_0 / v(x) \leqq \psi(x) \text{ a.e. in } \mathcal{O}\}, \tag{4.40}$$

we consider the variational inequality

$$\text{Find } u \in K_0 \text{ such that } a(u, v - u) \geqq (f, v - u) \quad \forall v \in K_0. \tag{4.41}$$

THEOREM 4.3. *Under assumptions* (4.1), (4.2), (2.26), *and* (2.27)[25] *the variational inequality* (4.41) *has exactly one solution u which is given as the optimal cost* (0.2).

[25] α is assumed large enough.

Moreover, the function u is Lipschitz continuous and verifies

$$\left\|\frac{\partial u}{\partial x}\right\|_{L^\infty} \leqq \frac{1}{\alpha-\gamma_0}\left\|\frac{\partial f}{\partial x}\right\|_{L^\infty} + \left\|\frac{\partial \psi}{\partial x}\right\|_{L^\infty} + \left\|\frac{\partial u}{\partial x}\right\|_{L^\infty}, \tag{4.42}$$

where $\|(\partial u/\partial x)\|_{L^\infty}$ *denotes the smallest Lipschitz constant of u.*

Proof. We just need to use the estimate (2.28) and the technique of Theorem 4.1. □

Remark 4.4. Clearly, combining Lemma 1.5 and Remark 1.7, we obtain a sufficient condition in order to have a Lipschitz continuous subsolution u, i.e., assumption (2.26).

Remark 4.5. Provided (4.37) holds, a weak formulation of the variational inequality (4.41) as (4.26) also can be considered.

Remark 4.6. All these results can be extended for f and ψ with polynomial growth, and we can also consider a function $a_0(x)$ instead of the constant α for the definition of operator A. Using the same technique, we can treat the parabolic case.

Remark 4.7. An application to the optimal stopping time problem with partial information is given in [16].

Remark 4.8. In the particular case, where the operator $A = A_1(x_1) + A_2(x_2)$, $x = (x_1, x_2)$ with A_1 coercive and A_2 of first order, a weak formulation (4.41) is obtained using only analytic methods (cf. M. Langlais [10]).

Final Remark. In a separate article in this issue [15], a degenerate quasi-variational inequality corresponding to the impulse control problem is studied (cf. [13]).

REFERENCES

[1] C. BARDOS, *Problèmes aux limites pour les equations aux deriveès partielles du premier ordre*, Ann. Sci. Ecole Norm. Sup. (4), Serie 3 (1970), pp. 185–233.

[2] A. BENSOUSSAN, *On the semigroup formulation of variational inequalities and quasi-variational inequalities*, First Franco-Southeast Asian Mathematical Conference, Singapore, May, 1979.

[3] A. BENSOUSSAN AND J. L. LIONS, *Applications des inéquations variationnelles en contrôle-stochastique*, Dunod, Paris, 1978.

[4] J. M. BISMUT, *Le problème de temps d'arrêt optimal*, C.R. Acad. Sci. Paris, Sér. A, 281 (1975), pp. 989–992.

[5] BOURBAKI, *XVI, Topologie generale*, Fascicule de Resultats, Hermann, Paris.

[6] E. B. DYNKIN, *Markov Processes*, Vol. I. Springer-Verlag, Berlin, 1965.

[7] M. I. FREIDLIN, *On the formulation of boundary value problems for degenerate elliptic equations*, Soviet Math. Dokl., 7 (1966), pp. 1204–1207.

[8] A. FRIEDMAN, *Stochastic Differential Equations and Applications*, Vol. I, Vol. II, Academic Press, New York, 1976.

[9] N. V. KRYLOV, *On control of a solution of a stochastic equation with degeneration*, Math. USSR-Izv., 6 (1972), pp. 249–262.

[10] M. LANGLAIS, *Solutions fortes et régularité pour des problèmes aux limites du second dégré dégénérés*, Cahier d'Analyse Appliquée et Informatique No 7703, Université de Bordeux I, Bordeaux, 1977.

[11] P. L. LIONS AND J. L. MENALDI, *Problèmes de Bellman avec le contrôle dans le coefficients de plus haut dégré*, C.R. Acad. Sci. Paris, Sér. A, 287 (1978), pp. 409–412.

[12] J. L. MENALDI, *Sur le problème de temps d'arrêt d'inéquation variationnelle dégénéré associée*, C.R. Acad. Sci. Paris, Sér. A, 284 (1977), pp. 1443–1446.

[13] ———, *Sur le problème de contrôle impulsionnel et l'inéquation quasi-variationnelle dégénérée associée*, C.R. Acad. Sci. Paris, Sér. A, 284 (1977), pp. 1499–1502.

[14] ———, *Le problème de temps d'arrêt optimal deterministe et l'inequation variationnelle de premier ordre associée*, Appl. Math. Optim., to appear.

[15] ———, *On the optimal impulse control problem for degenerate diffusions*, this Journal, this issue, pp. 722–739.

[16] ———, *Le principe de separation pour le problème de temps d'arrêt optimal*, Stochastics, 3 (1979), pp. 47–59.

[17] J. L. MENALDI AND E. ROFMAN, *On the final stopping time problem*, Lecture Notes in Computer Science, 27, Springer-Verlag, Berlin, 1975, pp. 210–211.

[18] M. ROBIN, *Contrôle impulsionnel des processus de Markov*, Thèse d'Etat, Université de Paris IX, Paris, 1977.

[19] D. W. STROOCK AND S. R. S. VARADHAN, *On degenerate elliptic-parabolic operators of second order and their associated diffusions*, Comm. Pure Appl. Math., 25 (1972), pp. 651–713.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 18, No. 6, November 1980

0363-0129/80/1806-0009 $01.00/0

ON THE OPTIMAL IMPULSE CONTROL PROBLEM FOR DEGENERATE DIFFUSIONS*

J. L. MENALDI†

Abstract. In this paper, we give a characterization of the optimal cost of an impulse control problem as the maximum solution of a quasi-variational inequality without assuming nondegeneracy. An estimate of the velocity of uniform convergence of the sequence of stopping time problems associated with the impulse control problem is given.

Introduction. Summary of main results. In this article, we develop the proofs of results announced in Note [5].

The impulse control problem has been studied by several authors. A. Bensoussan and J. L. Lions [2] treated nondegenerate diffusions, M. Robin [11] developed the case of Feller processes, and J. P. Lepeltier and B. Marchal [4] investigated a similar problem for a more general kind of Markov processes. In a purely analytical framework, L. Tartar [13] considered an abstract coercive quasi-variational inequality and F. Mignot and J. P. Puel [10] a first order quasi-variational inequality.

We study here the case of degenerate diffusions which lead to a second order noncoercive quasi-variational inequality. The deterministic case leading to a first order quasi-variational inequality is treated in [6].

Let $(\Omega, \mathcal{F}, P)$ be a probability space and $\{\mathcal{F}^t\}_{t\geqq 0}$ be a nondecreasing right-continuous family of completed sub-σ-fields of $\mathcal{F}$.

Let ν be any admissible[1] impulse control and $y(t)=y_x(t,\nu,\omega)$, $t\geqq 0$, $\omega\in\Omega$ be the diffusion with jumps on $\mathbb{R}^N$ starting at x, with Lipschitz continuous coefficients $g(\cdot)$ and $\sigma(\cdot)$.

Suppose $\mathcal{O}$ is an open subset of $\mathbb{R}^N$, and $\tau=\tau_x(\nu,\omega)$ the first exit time of process $y(t)$ from $\bar{\mathcal{O}}$.

Next, let $f(x)$ be a bounded upper semicontinuous nonnegative real function on $\bar{\mathcal{O}}$, and $k(\xi)$ be a continuous real function on $\mathbb{R}^N_+$ such that

$$k(\xi)\geqq k_0>0\ \forall\xi\geqq 0, \quad \text{and} \quad k(\xi)\to\infty \text{ if } |\xi|\to\infty. \tag{0.1}$$

Given $x\in\bar{\mathcal{O}}$ and an admissible impulse control $\nu=\{\theta_1,\xi_1;\cdots;\theta_i,\xi_i;\cdots\}$, the functional cost is defined by

$$J_x(\nu)=E\left\{\int_0^\tau f(y(t))\,e^{-\alpha t}\,dt+\sum_{i=1}^{\infty}k(\xi_i)1_{\theta_i<\infty}\,e^{-\alpha\theta_i}\right\}, \tag{0.2}$$

where α is a positive constant.

Our purpose is to characterize the optimal cost

$$\hat{u}(x)=\inf\{J_x(\nu)/\nu \text{ an admissible impulse control}\}, \tag{0.3}$$

and to obtain an optimal admissible impulse control.

* Received by the editors July 10, 1979, and in final revised form March 3, 1980.

† Université de Paris IX (Dauphine), Paris, France.

[1] See Def. (1.7).

We denote by A_0 the second order differential operator associated with the Ito equation[2]

$$A_0 = -\tfrac{1}{2}\operatorname{tr}\left(\sigma\sigma^*\frac{\partial^2}{\partial x^2}\right) - g\frac{\partial}{\partial x} \tag{0.4}$$

and $A = A_0 + \alpha$.

Let $\Gamma_0 \subset \partial\mathcal{O}$ be the set of regular points, and let us use the integral formulation of A[3].

We define by M the operator

$$[M\phi](x) = \inf\{k(\xi) + \phi(x+\xi)/\xi \geqq 0,\ x+\xi \in \bar{\mathcal{O}}\}. \tag{0.5}$$

Assume that $\mathcal{O}$ is sufficiently smooth such that M maps continuous functions ϕ into continuous functions $M\phi$. We will give conditions below (Lemma 1.3), so that M has the proposed regularity.

Finally, we introduce the problem: To find a real bounded measurable function on $\bar{\mathcal{O}}$, $u(x)$ such that

$$\begin{aligned} u &= 0 && \text{on } \Gamma_0,\\ u &\leqq Mu && \text{in } \bar{\mathcal{O}}\backslash\Gamma_0,\\ Au &\leqq f && \text{in the martingale sense on } \bar{\mathcal{O}}\backslash\Gamma_0. \end{aligned} \tag{0.6}$$

Now, we consider the following sequence of variational inequalities corresponding to optimal stopping time problems (cf. [7]).

Let $\hat{u}^0(x)$ be the bounded upper semicontinuous nonnegative real function on $\bar{\mathcal{O}}$ such that

$$\begin{aligned} \hat{u}^0 &= 0 && \text{on } \Gamma_0,\\ A\hat{u}^0 &= f && \text{in the martingale sense on } \bar{\mathcal{O}}\backslash\Gamma_0, \end{aligned} \tag{0.7}$$

and given $\hat{u}^{n-1}(x)$, let $\hat{u}^n(x)$ be the bounded upper semicontinuous nonnegative real function on $\bar{\mathcal{O}}$ which is the maximum solution of

$$\begin{aligned} u^n &= 0 && \text{on } \Gamma_0,\\ u^n &\leqq M\hat{u}^{n-1} && \text{in } \bar{\mathcal{O}}\backslash\Gamma_0,\\ Au^n &\leqq f && \text{in the martingale sense on } \bar{\mathcal{O}}\backslash\Gamma_0. \end{aligned} \tag{0.8}$$

We have the following characterization.

THEOREM 0.1. *Assume that g, σ are Lipschitz continuous,* (0.1), *and that f is bounded upper semicontinuous and nonnegative. Then problem* (0.6) *admits a maximum solution $\hat{u}$ which is upper semicontinuous and given as the optimal cost* (0.3). *Moreover, the following assertions are true.*

(0.9)[4]
$$\|\hat{u}\| \leqq \frac{1}{\alpha}\|f\|,$$

$$\hat{u}^n(x) \to \hat{u}(x)(n \to \infty) \quad \textit{uniformly in } x \in \bar{\mathcal{O}}. \tag{0.10}$$

[2] If B is a matrix, then B^* denotes the transpose of B and tr (B) the trace of B.

[3] See Def. (1.13).

[4] $\|\cdot\|$ denotes the supremum norm on $\bar{\mathcal{O}}$.

Furthermore, if Γ_0 *is closed and f continuous, the function* $\hat{u}$ *is also continuous on* $\bar{\mathcal{O}}$ *and there exists an optimal admissible impulse control.*

Regarding $\hat{u}$ as a distribution in $\mathcal{O}$, we have

THEOREM 0.2. *Let the assumptions be the same as in Theorem* 0.1. *Suppose*

$$\frac{\partial^2}{\partial x^2}\sigma\sigma^* \in L^1_{\text{loc}}(\mathcal{O}). \tag{0.11}$$

Then the optimal cost $\hat{u}$ *verifies*

$$A\hat{u} \leqq f \quad \text{in } \mathcal{D}'(\mathcal{O}). \tag{0.12}$$

Moreover, if Γ_0 *is closed and f continuous, the following equation*

$$A\hat{u} = f \quad \text{in } \mathcal{D}'([\hat{u} < M\hat{u}]) \tag{0.13}$$

is also true.

Now, a quasi-variational formulation is given.

Let $\beta_0(x)$, $\beta_1(x)$ be the weight functions $(1+|x|^2)^{-(\lambda+1)/2}$, $(1+|x|^2)^{-\lambda/2}$, $\lambda > N/2$ respectively. Introduce the following Hilbert spaces, $H = \{v/\beta_0 v \in L^2(\mathcal{O})\}$ with scalar product $(\cdot\,,\cdot)$, and $V = \{v \in H/\beta_1(\partial v/\partial x_i) \in L^2(\mathcal{O}), \forall i = 1, \cdots, N \text{ and } v = 0 \text{ on } \Gamma\}$. The space V' is the dual of V, and $\langle\cdot\,,\cdot\rangle$ denotes the duality between V' and V.

Consider the following quasi-variational inequality:

$$\begin{aligned} &u \in V, \qquad u \leqq Mu, \\ &\langle Au, v-u\rangle \geqq (f, v-u) \quad \forall v \in V, \quad v \leqq Mu. \end{aligned} \tag{0.14}$$

Assume

$$\frac{\partial^2}{\partial x^2}\sigma\sigma^* \in L^\infty(\mathcal{O}), \tag{0.15}$$

and that there exists a Lipschitz continuous subsolution $\bar{w}$, i.e.,

$$\bar{w} \in W_0^{1,\infty}(\mathcal{O}) \quad \text{and} \quad A\bar{w} \leqq -f \text{ in } \mathcal{D}'(\mathcal{O}), \tag{0.16}$$ [5]

where the constant α is assumed large enough.

For instance, if $\mathcal{O} = \mathbb{R}^N$ or $\sigma\sigma^*$ is coercive on Γ, then the assumption (0.16) is satisfied.

THEOREM 0.3. *Let the conditions of Theorem* 0.1, (0.15), *and* (0.16) *hold. Suppose that f is Lipschitz continuous; then the quasi-variational inequality* (0.14) *has a maximum solution* $\hat{u}$ *which is Lipschitz continuous and explicitly given as the optimal cost* (0.2).

This work is divided into three sections. The first section establishes several useful lemmas. In § 2, the integral formulation of the impulse control problem is studied, and in the last section, the associated quasi-variational inequality is treated.

In this paper, we will use extensively the results of [7].

1. Preliminary results. Let $(\Omega, \mathcal{F}, P)$ be a probability space, $\{\mathcal{F}^t\}_{t\geqq 0}$ a nondecreasing right-continuous family of completed sub-σ-fields of $\mathcal{F}$, and $w(t)$ a standard Brownian motion in $\mathbb{R}^N$ with respect to $\mathcal{F}^t$.

[5] Also in the martingale sense.

Suppose we are given two Lipschitz continuous functions $g(x)$ and $\sigma(x)$ on $\mathbb{R}^N$, taking values in $\mathbb{R}^N$ and $\mathbb{R}^N \otimes \mathbb{R}^N$, respectively, $g = (g_i)$, $\sigma = (\sigma_{ij})$,

$$(1.1)^6 \qquad \frac{\partial g_i}{\partial x_k}, \frac{\partial \sigma_{ij}}{\partial x_k} \in B(\mathbb{R}^N), \qquad i, j, k = 1, \cdots, N.$$

We consider the diffusion $y^0(t) = y_x^0(t, \omega)$, $t \geqq 0$, $\omega \in \Omega$ and $x \in \mathbb{R}^N$, described by the Ito equation

$$(1.2) \qquad \begin{aligned} &dy^0(t) = g(y^0(t))\,dt + \sigma(y^0(t))\,dw(t), \qquad t \geqq 0, \\ &y^0(0) = x. \end{aligned}$$

Let Λ be a closed subset of $\mathbb{R}^N$, convex with respect to zero[7]. An impulse control ν is a set $\{\theta_1, \xi_1; \cdots; \theta_i, \xi_i; \cdots\}$ where $\{\theta_i\}_{i=1}^{\infty}$ is an increasing sequence of stopping times with respect to $\mathcal{F}^t$ convergent to infinity ($\theta_i \leqq \theta_{i+1}$, $\theta_i \to \infty$) and $\{\xi_i\}_{i=1}^{\infty}$ is a sequence of random variables taking values on Λ, adapted with respect to $\{\theta_i\}_{i=1}^{\infty}$ ($\xi_i : \Omega \to \Lambda$, $\mathcal{F}^{\theta_i}$ measurable).

Now, we define the sequence of diffusions with jumps $\{y^n(t)\}_{n=1}^{\infty}$, $y^n(t) = y_x^n(t, \nu, \omega)$, $t \geqq 0$, $\omega \in \Omega$, $x \in \mathbb{R}^N$, and ν any impulse control, by the Ito equation

$$(1.3) \qquad \begin{aligned} &dy^n(t) = g(y^n(t))\,dt + \sigma(y^n(t))\,dw(t), \qquad &t \geqq \theta_n, \\ &y^n(t) = y^{n-1}(t) + 1_{\theta_n = t}\xi_n, \qquad &t \leqq \theta_n. \end{aligned}$$

We have

$$(1.4) \qquad y^n(t) = y^i(t) \text{ on } [0, \theta_n] \quad \forall_i \geqq n.$$

So, if we define

$$(1.5) \qquad y(t, \nu) = \lim_{n \to \infty} y^n(t), \qquad t \geqq 0,$$

the process $y(t) = y_x(t, \nu, \omega)$, which is right-continuous[8], satisfies the stochastic equation,

$$(1.6) \qquad \begin{aligned} &dy(t) = g(y(t))\,dt + \sigma(y(t))\,dw(t) + \sum_{i=1}^{\infty} \xi_i \delta(t - \theta_i)\,dt, \qquad t \geqq 0, \\ &y(0) = x, \end{aligned}$$

where $\delta(t)$ is the Dirac measure.

Suppose $\mathcal{O}$ an open subset of $\mathbb{R}^N$, and $\tau = \tau_x(\nu, \omega)$, $\tau^0 = \tau_x^0(\omega)$ the first exit time of processes $y(t)$, $y^0(t)$ respectively, from $\bar{\mathcal{O}}$.

We call $\nu = \{\theta_1, \xi_1; \cdots; \theta_i, \xi_i; \cdots\}$ an admissible impulse control if it satisfies

$$(1.7) \qquad y(\tau) \in \bar{\mathcal{O}} \quad \text{a.s. on } [\tau < \infty];$$

that is, no jump of the process $y(t)$ is outside of $\bar{\mathcal{O}}$ before τ.

Denote by Γ_0 the set of regular points (cf. D. W. Stroock and S. R. S. Varadhan [12]),

$$(1.8) \qquad \Gamma_0 = \{x \in \Gamma = \partial\mathcal{O} / P(\tau_x^0 > 0) = 0\}.$$

[6] $B(\mathbb{R}^N)$ denotes the set of all Borel measurable and bounded functions on $\mathbb{R}^N$ taking values in $\mathbb{R}$.

[7] i.e., $\lambda\xi \in \Lambda$, $\forall \lambda \in [0, 1]$, $\forall \xi \in \Lambda$. Generally, we take $\Lambda = \mathbb{R}_+^N$.

[8] $y(t)$ has also left limits.

LEMMA 1.1. *Assume* (1.1). *Let ν be any admissible impulse control, and θ be any stopping time; then the following assertions are true.*

$$P(y(\tau,\nu)\notin\Gamma_0,\ \tau<\infty)=0, \tag{1.9}$$

$$E\{|y_x(\theta)-y_{x'}(\theta)|^2 e^{-\gamma\theta}\}\leqq |x-x'|^2 \quad \forall x, x'\in\mathbb{R}^N, \tag{1.10}$$

where the positive constant γ depends on the Lipschitz constant of functions g and σ.

Proof. Setting

$$\begin{aligned}\gamma=\sup\Big\{\operatorname{tr}\Big[\frac{(\sigma(x)-\sigma(x'))(\sigma(x)-\sigma(x'))^*}{|x-x'|^2}\Big]\\ +\frac{2(x-x')(g(x)-g(x'))}{|x-x'|^2}\Big/ x, x'\in\mathbb{R}^N\Big\},\end{aligned} \tag{1.11}$$

and recalling that the process $y_x(t)-y_{x'}(t)$ is a diffusion (from Ito's formula) to the function $|x|^2 e^{-\gamma t}$, we obtain (1.10) as Lemma 1.1 in [7].

Finally, using (1.7) from Markov's property we get

$$P(y^n(\tau^n)\notin\Gamma_0,\ \tau^n<\infty)=0, \tag{1.12}$$

where τ^n is the first exit time of process $y^n(t)$ from $\bar{\mathcal{O}}$. So regarding (1.4), we deduce (1.9). □

Let u, v be real bounded[9] upper semicontinuous functions on $\bar{\mathcal{O}}$. Then the integral formulation of operation A (cf. [7]) is given by

$Au\leqq v$ in $\bar{\mathcal{O}}\backslash\Gamma_0$ if the process

(1.13)[10]
$$X_t=\int_0^{\theta\wedge\tau^0} v(y^0(s))\,e^{-\alpha s}\,ds+u(y^0(t\wedge\tau^0))\,e^{-\alpha(t\wedge\tau^0)}$$

is a submartingale for each $x\in\bar{\mathcal{O}}\backslash\Gamma_0$.

LEMMA 1.2. *Assume* (1.1) *and $\mathcal{O}$ smooth*[11]. *Let $f(x)$ be a real bounded continuous function on $\bar{\mathcal{O}}$. Suppose that there exists $\bar{w}$ such that*

$$\begin{gathered}\bar{w}\in C(\bar{\mathcal{O}}),\qquad \bar{w},\frac{\partial\bar{w}}{\partial x_i}\in B(\bar{\mathcal{O}}),\qquad i=1,\cdots,N,\\ A\bar{w}\leqq -f \quad in\ \mathcal{D}'(\mathcal{O}),\qquad \bar{w}(x)=0\quad \forall x\in\Gamma.\end{gathered} \tag{1.14}$$

Then, for any admissible[12] *impulse control $\nu=\{\theta_1,\xi_1;\cdots;\theta_i,\xi_i;\cdots\}$ such that*

(1.15)[13]
$$\theta_i\notin[\tau_x\wedge\tau_{x'},\tau_x[\quad \forall i=1,2,\cdots,$$

the following estimation is true:

$$E\Big\{\int_{\tau_x\wedge\tau_{x'}}^{\tau_x} f(y_x(t))\,e^{-\alpha t}\,dt\Big\}\leqq\Big\|\frac{\partial\bar{w}}{\partial x}\Big\||x-x'|\qquad \forall x, x'\in\bar{\mathcal{O}}, \tag{1.16}$$

where $\|\partial\bar{w}/\partial x\|$ denotes the smallest Lipschitz continuous constant of $\bar{w}$.

[9] u and v may have polynomial growth if $\mathcal{O}$ is not bounded.

[10] We say $Au\leqq v$ in the martingale sense.

[11] We also assume α large enough.

[12] Clearly, admissible for x.

[13] $\tau_x\wedge\tau_{x'}$ denotes the minimum between τ_x and $\tau_{x'}$.

Proof. First, assume $\bar{w} \in C^2(\mathbb{R}^N)$; $\bar{w}$, $\partial \bar{w}/\partial x_i \in B(\mathbb{R}^N)$, $i = 1, \cdots, N$. Ito's formula applied to function $\bar{w}(x)$ and process $y_x(t)$ gives

$$E\{\bar{w}(y_x(\tau_x))\, e^{-\alpha\tau_x} - \bar{w}(y_x(\tau_x \wedge \tau_{x'}))\, e^{-\alpha(\tau_x \wedge \tau_{x'})}\} \tag{1.17}$$
$$= -E\left\{\int_{\tau_x \wedge \tau_{x'}}^{\tau_x} A\bar{w}(y_x(t))\, e^{-\alpha t}\, dt\right\}.$$

Since

$$\bar{w}(y_x(\tau_x)) = 0 = \bar{w}(y_{x'}(\tau_x \wedge \tau_{x'})) \quad \text{a.s. in } (\tau_{x'} \leqq \tau_x < \infty],$$

from (1.17), we deduce

$$E\left\{\int_{\tau_x \wedge \tau_{x'}}^{\tau_x} f(y_x(t))\, e^{-\alpha t}\, dt\right\} \tag{1.18}$$
$$\leqq E\{|\bar{w}(y_x(\tau_x \wedge \tau_{x'})) - \bar{w}(y_{x'}(\tau_x \wedge \tau_{x'}))|\, e^{-\alpha(\tau_x \wedge \tau_{x'})}\}.$$

Next, defining

$$\gamma_0 = \sup\left\{\frac{1}{2}\operatorname{tr}\left[\frac{(\sigma(x) - \sigma(x'))(\sigma(x) - \sigma(x'))^*}{|x - x|^2}\right]\right. \tag{1.19}$$
$$\left. + \frac{(x - x')(g(x) - g(x'))}{|x - x'|^2} \Big/ x, x' \in \mathbb{R}^N\right\},$$

and assuming $\alpha \geqq \gamma_0$, from Lemma 1.1 and (1.18) we obtain (1.16). Finally, if $\bar{w} \notin C^2(\mathcal{O})$, by approximating $\bar{w}$ by regular functions the lemma is proved. □

Remark 1.1. Assume $\bar{w} \in W^{1,\infty}(\mathcal{O})$, $f \in C(\bar{\mathcal{O}}) \cap B(\bar{\mathcal{O}})$. Approximating $\bar{w}$ by regular functions, we deduce that [$A\bar{w} \leqq f$ in $\mathcal{D}'(\mathcal{O})$] is equivalent to [$A\bar{w} \leqq f$ in the martingale sense of (1.13)]. This fact will be used several times.

Suppose we are given a continuous real function $k(\xi)$ on Λ, such that

$$\begin{aligned} &k(\xi) \geqq k_0 > 0 \quad \forall \xi \in \Lambda, \\ &k(\xi) \to \infty \quad \text{if } |\xi| \to \infty \quad \text{with } \xi \in \Lambda. \end{aligned} \tag{1.20}$$

We define the operator $M: B(\bar{\mathcal{O}}) \to B(\bar{\mathcal{O}})$ by

$$[M\phi](x) = \inf\{k(\xi) + \phi(x + \xi) / \xi \in \Lambda, x + \xi \in \bar{\mathcal{O}}\}. \tag{1.21}$$

We always assume $\mathcal{O}$ and Λ smooth enough, such that

There exists $P: \bar{\mathcal{O}} \times \Lambda \to \Lambda$ measurable and uniformly continuous in $x \in \bar{\mathcal{O}}$ verifying

$$\begin{aligned} &x + P(x, \xi) \in \bar{\mathcal{O}} \quad \forall x \in \bar{\mathcal{O}}, \quad \forall \xi \in \Lambda, \\ &P(x, \xi) = \xi \quad \text{if } x + \xi \in \bar{\mathcal{O}}. \end{aligned} \tag{1.22}$$

For instance, if $\Lambda = \mathbb{R}_+^N$ and $\mathcal{O}$ convex with regular boundary, we can take $P(x, \xi)$ as the projection of ξ on $\Lambda \cap (\bar{\mathcal{O}} - x)$.

LEMMA 1.3. *Assume* (1.20) *and* (1.22). *Then if ϕ is upper semicontinuous* (*or continuous*) *on $\bar{\mathcal{O}}$, so is $M\phi$.*

Proof. Starting at

$$[M\phi](x) - [M\phi](x') = \sup_{\xi'} \inf_{\xi} [(k(\xi) - k(\xi') + (\phi(x + \xi) - \phi(x' + \xi'))],$$

and choosing $\xi = P(x, \xi')$, we get

$$[M\phi](x) - [M\phi](x') \leqq \sup_{\xi'} [k(P(x, \xi')) - k(P(x', \xi'))] \tag{1.23}$$
$$+ \sup_{\xi'} [\phi(x + P(x, \xi')) - \phi(x' + P(x', \xi'))].$$

So, from (1.23) and the uniform continuity of function $P(x, \xi)$, the lemma is proved. □

LEMMA 1.4. *Suppose* (1.20), (1.22) *and*

$$\text{(1.24)} \qquad \phi \textit{ bounded and upper semicontinuous on } \bar{\mathcal{O}}.$$

Then, for each $\varepsilon > 0$ *there exists a function* $\xi_\varepsilon(x)$ *such that*

$$\text{(1.25)} \qquad \begin{array}{l} \xi_\varepsilon : \bar{\mathcal{O}} \to \Lambda \textit{ bounded and Borel measurable,} \\ x + \hat{\xi}_\varepsilon(x) \in \bar{\mathcal{O}} \quad \forall x \in \bar{\mathcal{O}}, \end{array}$$

$$\text{(1.26)} \qquad [M\phi](x) + \varepsilon \geqq [k(\hat{\xi}_\varepsilon(x)) + \phi(x + \hat{\xi}_\varepsilon(x))] \quad \forall x \in \bar{\mathcal{O}}.$$

Moreover, if ϕ *is continuous, there exists* $\hat{\xi}(x)$ *verifying* (1.25) *and* (1.26) *with* $\varepsilon = 0$.

Proof. First, if ϕ is continuous, the classical theorems of selection imply the result.

Next, if ϕ is only upper semicontinuous, there exists a decreasing sequence $\{\phi_n\}_{n=1}^\infty$ of continuous functions convergent to ϕ. So, we also have $M\phi_n$ decreasing to $M\phi$.

Let $\hat{\xi}^n(x)$ be a function which satisfies (1.25) and

$$[M\phi_n](x) = [k(\hat{\xi}^n(x)) + \phi_n(x + \hat{\xi}^n(x))] \quad \forall x \in \bar{\mathcal{O}},$$

and let $n_\varepsilon(x)$ be the function

$$n_\varepsilon(x) = \min \{n \geqq 1/[M\phi_n](x) \leqq [M\phi](x) + \varepsilon\}.$$

Then, if we set

$$\text{(1.27)} \qquad \hat{\xi}_\varepsilon(x) = \hat{\xi}^n(x) \quad \text{if } n = n_\varepsilon(x),$$

the lemma is proved. □

2. Integral formulation. Let Γ_0 be the set of regular points (1.8) and A be the operator given by (1.13). Assume $f(x)$ an upper semicontinuous function on $\bar{\mathcal{O}}$ such that

$$\text{(2.1)} \qquad f \in B(\bar{\mathcal{O}}), \qquad f \geqq 0.$$

Consider the following problem: To find $u(x)$ such that

$$\text{(2.2)} \qquad u \in B(\bar{\mathcal{O}}), \qquad u(x) = 0 \quad \forall x \in \Gamma_0,$$

$$\text{(2.3)} \qquad Au \leqq f \text{ in } \bar{\mathcal{O}} \backslash \Gamma_0 \quad \text{[martingale sense (1.13)]},$$

$$\text{(2.4)} \qquad u \leqq Mu \quad \text{on } \bar{\mathcal{O}} \backslash \Gamma_0.$$

Let us define the sequence $\{\hat{u}^n\}_{n-1}^\infty$ of solutions to variational inequalities corresponding to optimal stopping time problems (cf. [7]). Starting with $\hat{u}^0(x)$ verifying (2.2) and

$$\text{(2.5)} \qquad A\hat{u}^0 = f \text{ in } \bar{\mathcal{O}} \backslash \Gamma_0 \text{ [martingale sense (1.13)]},$$

we set $\hat{u}^n(x)$ as the maximum solution of problem (2.2), (2.3) and

$$\text{(2.6)} \qquad u^n \leqq M\hat{u}^{n-1} \quad \text{on } \bar{\mathcal{O}} \backslash \Gamma_0,$$

This section is divided into two parts. First we solve problem (2.2), (2.3), (2.4) and consider the case where the set of regular points Γ_0 is closed. Next we study the general case and give some complementary results

2.1. Regular case.

THEOREM 2.1. *Let the assumptions* (1.1), (1.20), (1.22) *and* (2.1) *hold. Then the problem* (2.2), (2.3), (2.4) *admits a maximum solution* $\hat{u}$ *which is given by the decreasing limit*

$$\hat{u}(x) = \lim_{n\to\infty} \hat{u}^n(x) \quad \forall x \in \bar{\mathcal{O}}. \tag{2.7}$$

Moreover, the function $\hat{u}(x)$ *is upper semicontinuous and the following estimate is true*:

$$\|u\| \leqq \frac{1}{\alpha}\|f\|, \tag{2.8}$$

where $\|\cdot\|$ *denotes the supremum norm on* $\bar{\mathcal{O}}$.

Proof. Using the monotone property of operator M,

$$\phi \leqq \psi \text{ implies } M\phi \leqq M\psi, \tag{2.9}$$

and knowing that $0 \leqq \hat{u}^1 \leqq \hat{u}^0$, we deduce

$$0 \leqq \hat{u}^{n+1} \leqq \hat{u}^n \leqq \hat{u}^0, \qquad n = 1, 2, \cdots. \tag{2.10}$$

Then, for any solution u of problem (2.2), (2.3), the trivial maximum principle in the martingale formulation implies $u \leqq \hat{u}^0$. Because of (2.4) and (2.9), we obtain

$$u \leqq \hat{u}^n, \qquad n = 1, 2, \cdots. \tag{2.11}$$

So, the function $\hat{u}$ defined by (2.7) is the maximum solution of problem (2.2), (2.3), and (2.4). Since $\hat{u}^n$ is upper semicontinuous (cf. [7]), we conclude the theorem. □

Remark 2.1. If we set $\psi = M\hat{u}$, the maximum solution $\hat{u}$ can also be considered as an optimal stopping time cost, i.e., the maximum solution of problem (2.2), (2.3) and $u \leqq \psi$.

We can also define the sequence $\{\hat{u}^n\}_{n=1}^{\infty}$ as the optimal costs

$$\hat{u}^0(x) = E\left\{\int_0^{\tau^0} f(y^0(t))\, e^{-\alpha t}\, dt\right\}, \tag{2.12}$$

and given $\hat{u}^{n-1}$ we obtain $\hat{u}^n$ by

$$\hat{u}^n(x) = \inf_\theta E\left\{\int_0^{\theta\wedge\tau^0} f(y^0(t))\, e^{-\alpha t}\, dt + M\hat{u}^{n-1}(y^0(\theta))1_{\theta<\tau^0}\, e^{-\alpha\theta}\right\} \tag{2.13}$$

where θ is any stopping time of $\mathscr{F}^t$.

THEOREM 2.2. *Let the conditions* (1.1), (1.20), (1.22), (2.1), *and*

$$f \in C(\bar{\mathcal{O}}), \tag{2.14}$$

$$\Gamma_0 \textit{ closed}, \tag{2.15}$$

hold. Then the maximum solution $\hat{u}$ *of problem* (2.2), (2.3), (2.4) *is continuous. Moreover,* $\hat{u}$ *is given as the optimal cost* (0.3), *and the following estimate is true*:

$$\|\hat{u}^n - \hat{u}\| \leqq \frac{\|f\|^2}{k_0\alpha^2(n+1)}, \qquad n = 0, 1, 2, \cdots. \tag{2.16}$$

Proof. Recalling that, from [7] and Lemma 1.3, $\hat{u}^n$ is continuous, we need only to show the estimate (2.16). Since Γ_0 is closed, we are in the case of Feller processes (cf. A. Bensoussan [1] and M. Robin [11]).

First, we are going to prove that

$$(2.17)^{14} \quad \hat{u}^n(x) = \inf\{J_x(\nu)/\nu \text{ admissible impulse control such that } \theta_i = \infty\ \forall i \geqq n+1\},$$

where the functional cost $J_x(\nu)$ is given by (0.2).

Indeed, from Lemma 1.4, there exist functions $\hat{\xi}^i(x)$, $i = 1, \cdots, n$ verifying (1.25) and

$$(2.18) \qquad [Mu^{n-i}](x) = k(\hat{\xi}^i(x)) + \hat{u}^{n-i}(x + \hat{\xi}^i(x)) \quad \forall x \in \bar{\mathcal{O}}.$$

Thus, we define $\hat{\nu}^n = \{\hat{\theta}_i, \hat{\xi}_i\}_{i=1}^{\infty}$ as follows.

$$(2.19) \qquad \tilde{\theta}^0 = 0;$$

$$(2.20) \qquad \begin{aligned} &d\hat{y}^0(t) = g(\hat{y}^0(t))\,dt + \sigma(\hat{y}^0(t))\,dw(t), \qquad t \geqq 0,\\ &\hat{y}^0(0) = x; \end{aligned}$$

$$(2.21)^{15} \qquad \hat{\tau}^i = \inf\{t \geqq 0/\hat{y}^i(t) \notin \bar{\mathcal{O}}\}, \qquad i = 0, 1, \cdots, n;$$

$$(2.22)^{16} \qquad \tilde{\theta}^{i+1} = \inf\{t \in [\tilde{\theta}^i, \hat{\tau}^i]/\hat{u}^{n-i}(\hat{y}^i(t)) = [M\hat{u}^{n-i-1}](\hat{y}^i(t))\}, \qquad i = 0, 1, \cdots, n-1;$$

$$(2.23)^{17} \qquad \hat{\xi}_{i+1} = \hat{\xi}^i(\hat{y}^i(\tilde{\theta}^{i+1})), \qquad i = 0, 1, \cdots, n-1;$$

$$(2.24) \qquad \begin{aligned} &d\hat{y}^i(t) = g(\hat{y}^i(t))\,dt + \sigma(\hat{y}^i(t))\,dw(t), \qquad t \geqq \tilde{\theta}^i,\\ &\hat{y}^i(\tilde{\theta}^i) = \hat{y}^{i-1}(\tilde{\theta}^i) + \hat{\xi}_i,\\ &\hat{y}^i(t) = \hat{y}^{i-1}(t), \qquad t < \tilde{\theta}^i, \quad i = 1, 2, \cdots, n; \end{aligned}$$

and next

$$(2.25) \qquad \hat{\theta}_i = \begin{cases} \tilde{\theta}_i & \text{if } i \leqq n \text{ and } \tilde{\theta}^i < \hat{\tau}^{i-1}, \\ \infty & \text{otherwise,} \end{cases} \qquad i = 1, 2, \cdots,$$

$$(2.26) \qquad \hat{\xi}_i = 0 \quad \text{if } i \geqq n+1.$$

We have

$$(2.27) \qquad y(t, \hat{\nu}^n) = \hat{y}^n(t), \qquad t \geqq 0,$$

and from Markov's property

$$(2.28) \qquad \hat{u}^n(x) = J_x(\hat{\nu}^n),$$

$$(2.29) \qquad \hat{u}^n(x) \leqq J_x(\nu) \quad \text{if } \nu \text{ has at most } n \text{ impulses.}$$

Then, (2.28) and (2.29) imply (2.17).

Now we are going to show the estimate (2.16).

Let $\nu = \{\theta_i, \xi_i\}_{i=1}^{\infty}$ be any admissible impulse control; setting $\nu^n = \{\theta_1, \xi_1; \cdots; \theta_n, \xi_n; \infty, \xi_{n+1}; \cdots\}$ we have

$$y(t, \nu) = y(t, \nu^n) = y^n(t) \quad \text{if } t < \theta_n \wedge \tau^n.$$

[14] i.e., ν has at the most n impulses.

[15] We set $\hat{\tau}^i = \infty$ if $\hat{y}^i(t) \in \bar{\mathcal{O}}\ \forall t \geqq 0$.

[16] We set $\tilde{\theta}^{i+1} = \hat{\tau}^i$ if the subset is empty.

[17] If $\tilde{\theta}^{i+1} = \infty$ we set $\hat{\xi}_{i+1} = 0$.

Hence, if $\hat{u}$ is given by (0.3), we obtain

$$0 \leqq \hat{u}^n - \hat{u} \leqq \sup_{\nu} E\left\{\int_{\theta_n \wedge \tau^n}^{\tau^n} f(y^n(t))\, e^{-\alpha t}\, dt\right\}. \tag{2.30}$$

Since

$$e^{-\alpha\theta_n} \leqq \frac{1}{k_0(n+1)} \sum_{i=1}^{\infty} k(\xi_i) 1_{\theta_i < \infty}\, e^{-\alpha\theta_i},$$

and since it is possible to take the supremum only over all admissible impulse controls such that

$$E\left\{\sum_{i=1}^{\infty} k(\xi_i) 1_{\theta_i<\infty}\, e^{-\alpha\theta_i}\right\} \leqq \frac{1}{\alpha}\|f\|,$$

the estimate (2.16) follows from (2.30). □

Remark 2.2. The estimate (2.16) can be improved using a probabilistic version of results in B. Hanouzet and J. L. Joly [3]. We have

$$\|\hat{u}^n - \hat{u}\| \leqq Cq^n, \qquad n = 0, 1, 2, \cdots, \tag{2.31}$$

where constants $C > 0$ and $q \in [0, 1[$ depend only on $\|f\|$, α, and k_0. Indeed, we define the operator $S\colon C(\bar{\mathcal{O}}) \to C(\bar{\mathcal{O}})$ by

$$Sv = \inf_{\theta} E\left\{\int_0^{\theta\wedge\tau^0} f(y^0(t))\, e^{-\alpha t}\, dt + Mv(y^0(\theta)) 1_{\theta<\tau^0}\, e^{-\alpha\theta}\right\}, \tag{2.32}$$

where θ is any stopping time of $\mathcal{F}^t$.

Let $\hat{u}^0$ be the function given by (2.12), so using estimate (2.8) and the fact that $k_0 \leqq M(0)$, we deduce

$$\lambda\hat{u}^0 \leqq S(0) \quad \text{if } 0 \leqq \lambda \leqq \frac{\alpha k_0}{\|f\|}. \tag{2.33[18]}$$

Clearly, the operator S is increasing and concave, hence it is easy to prove from (2.33) the following property:

$$\begin{gathered} \forall u, v \in C(\bar{\mathcal{O}}), \quad 0 \leqq u, v \leqq \hat{u}^0 \text{ and satisfying} \\ -rv \leqq u - v \leqq pu, \qquad r, p \in [0, 1], \\ \text{we have} \\ -(1-\lambda) rSv \leqq Su - Sv \leqq (1-\lambda) pSu. \end{gathered} \tag{2.34}$$

Next, we obtain from (2.34)

$$\|S^n\hat{u}^0 - S^m\hat{u}^0\| \leqq (1-\lambda)^{m-n}\|\hat{u}^0\|, \qquad m > n, \tag{2.35}$$

and recalling that $\hat{u}^n = S^n\hat{u}^0$, we have the estimate (2.31) with $C = \|\hat{u}^0\|$ and $q = 1 - \lambda$. □

COROLLARY 2.1. *Let the assumptions be as in Theorem* 2.2. *Then there exists an optimal admissible impulse control* $\hat{\nu} = \hat{\nu}_x$,

$$\hat{u}(x) = J_x(\hat{\nu}), \tag{2.36}$$

where $\hat{u}$ *is given by* (0.3).

[18] We assume that $\lambda \leqq 1$.

Proof. From Theorem 2.2, the function $\hat{u}(x)$ is continuous. Then, from Lemma 1.4, there exists a function $\hat{\xi}(x)$ verifying (1.25) and

$$[M\hat{u}](x)=k(\hat{\xi}(x))+\hat{u}(x+\hat{\xi}(x))\quad \forall x\in\bar{\mathcal{O}}. \tag{2.37}$$

Then, we define $\hat{\nu}=\{\hat{\theta}_i,\hat{\xi}_i\}_{i=1}^{\infty}$ by

$$\tilde{\theta}^0=0; \tag{2.38}$$

$$\begin{aligned}&d\hat{y}^0(t)=g(\hat{y}^0(t))\,dt+\sigma(\hat{y}^0(t))\,dw(t),\qquad t\geqq 0,\\ &\hat{y}^0(0)=x;\end{aligned} \tag{2.39}$$

$$\hat{\tau}^i=\inf\{t\geqq 0/\hat{y}^i(t)\notin\bar{\mathcal{O}}\},\qquad i=0,1,2,\cdots; \tag{2.40}$$

$$\tilde{\theta}^{i+1}=\inf\{t\in[\tilde{\theta}^i,\hat{\tau}^i]/\hat{u}(\hat{y}^i(t))=[M\hat{u}](\hat{y}^i(t))\},\qquad i=0,1,2,\cdots, \tag{2.41}$$

$$\hat{\xi}_{i+1}=\hat{\xi}(\hat{y}^i(\tilde{\theta}^{i+1})),\qquad i=0,1,2,\cdots; \tag{2.42}$$

$$\begin{aligned}&d\hat{y}^i(t)=g(\hat{y}^i(t))\,dt+\sigma(\hat{y}^i(t))\,dw(t),\qquad && t\geqq\tilde{\theta}^i,\\ &\hat{y}^i(\tilde{\theta}^i)=\hat{y}^{i-1}(\tilde{\theta}^i)+\hat{\xi}_i, && i=1,2,\cdots,\\ &\hat{y}^i(t)=\hat{y}^{i-1}(t),\qquad t<\tilde{\theta}^i,\end{aligned} \tag{2.43}$$

and later on,

$$\hat{\theta}_i=\begin{cases}\tilde{\theta}^i & \text{if } \tilde{\theta}^i<\hat{\tau}^{i-1},\\ \infty & \text{otherwise,}\end{cases}\qquad i=1,2,\cdots. \tag{2.44}$$

We have

$$y(t,\hat{\nu})=\hat{y}^n(t)\quad \text{if } 0\leqq t<\hat{\theta}_n, \tag{2.45}$$

and from Markov's property

$$\begin{aligned}\hat{u}(x)=E\Big\{\int_0^{\hat{\theta}_n\wedge\hat{\tau}^{n-1}} f(\hat{y}^n(t))\,e^{-\alpha t}\,dt+\sum_{i=1}^{n}k(\hat{\xi}_i)1_{\hat{\theta}_i<\infty}\,e^{-\alpha\hat{\theta}_i}\Big\}\\ +E\{1_{\hat{\theta}_n<\hat{\tau}^{n-1}}\hat{u}(\hat{y}^n(\hat{\theta}_n))\,e^{-\alpha\hat{\theta}_n}\}.\end{aligned} \tag{2.46}$$

Hence, letting $n\to\infty$ in (2.46) and, using (2.45) and (1.9), we obtain (2.36). □

2.2. Complementary results. Now we omit assumptions (2.14) and (2.15).

THEOREM 2.3. *Let the conditions* (1.1), (1.20), (1.22), *and* (2.1) *hold. Then the maximum solution* $\hat{u}$ *of problem* (2.2), (2.3), (2.4) *is given as the optimal cost* (0.3), *and the estimate* (2.16) *is true.*

Proof. As in Theorem 2.2, we just need to prove (2.17). Moreover, we will only show that

(2.47) $\forall\varepsilon>0$ there exists $\hat{\nu}^{\varepsilon}$, an admissible impulse control which has at most n impulses, such that

$$\hat{u}^n(x)+\varepsilon\geqq J_x(\hat{\nu}^{\varepsilon}).$$

Indeed, given $\varepsilon>0$, from Theorem 3.4 in [7], we can choose a stopping time which is ε-optimal and depends measurably on the initial point, so there exist functions $\hat{\theta}^i_{\varepsilon}(x)$, $i=1,2,\cdots,n$, such that

$$\begin{aligned}&\hat{\theta}^i_{\varepsilon}:\bar{\mathcal{O}}\times\Omega\to[0,\infty]\text{ is Borel measurable,}\\ &\forall x\in\bar{\mathcal{O}},\quad \hat{\theta}^i_{\varepsilon}(x)\text{ is a stopping time;}\end{aligned} \tag{2.48}$$

$$\begin{aligned}\hat{u}^{n-i+1}+\varepsilon 2^{-n-1} \geqq E\bigg\{\int_0^{\hat{\theta}_\varepsilon^i \wedge \tau^0} f(y^0(t))\, e^{-\alpha t}\, dt + 1_{\hat{\theta}_\varepsilon^i<\tau^0}[M\hat{u}^{n-i}]& \\ \cdot (y^0(\hat{\theta}_\varepsilon^i)) \exp(-\alpha\hat{\theta}_\varepsilon^i)\bigg\}&.\end{aligned} \tag{2.49}$$

Also from Lemma 1.4, there exist functions $\hat{\xi}_\varepsilon^i(x)$, $i=1, 2, \cdots, n$, verifying (1.25), and

$$[M\hat{u}^{n-i}](x)+\varepsilon 2^{-n-1} \geqq k(\hat{\xi}_\varepsilon^i(x))+\hat{u}^{n-i}(x+\hat{\xi}_\varepsilon^i(x)) \quad \forall x \in \bar{\mathcal{O}}. \tag{2.50}$$

Thus, defining the admissible impulse control $\hat{\nu}^\varepsilon = \{\hat{\theta}_i, \hat{\xi}_i\}_{i=1}^\infty$ by (2.19), (2.20), and

$$\hat{\tau}^i = \inf\{t \geqq 0/\hat{y}^i(t) \notin \bar{\mathcal{O}}\}, \qquad i=0, 1, \cdots, n, \tag{2.51}$$

$$\tilde{\theta}^i = [\tilde{\theta}^{i-1}+\hat{\theta}_\varepsilon^i(\hat{y}^{i-1}(\tilde{\theta}^{i-1}))] \wedge \hat{\tau}^{i-1}, \qquad i=1, \cdots, n, \tag{2.52}$$

$$\hat{\xi}_i = \hat{\xi}_\varepsilon^{i-1}(\hat{y}^{i-1}(\tilde{\theta}^i)), \qquad i=1, \cdots, n, \tag{2.53}$$

and (2.24), (2.25), (2.26) we deduce assertion (2.47) using Markov's property. □

COROLLARY 2.2. *Let the assumptions be as in Theorem* 2.3. *Then given $\varepsilon > 0$ there exists a function $\hat{\nu}_\varepsilon(x) = \{\hat{\theta}_i(x), \hat{\xi}_i(x)\}_{i=1}^\infty$ such that $\hat{\theta}_i$ and $\hat{\xi}_i$ verify* (2.48) *and* (1.25) *respectively, and*

$$\hat{u}(x)+\varepsilon \geqq J_x(\hat{\nu}_\varepsilon(x)) \quad \forall x \in \bar{\mathcal{O}}, \tag{2.54}$$

where $\hat{u}$ is the optimal cost given by (0.3).

Proof. We just need to combine the methods of Theorem 2.3 and Corollary 2.1. □

Finally, the function $\hat{u}$ is regarded as a distribution in $\mathcal{O}$. Notice that Theorem 0.1 is completely proved.

Recalling that A is the differential operator (0.4) and assuming

$$\frac{\partial^2}{\partial x^2}\sigma\sigma^* \in L^1_{\text{loc}}(\mathcal{O}) \tag{2.55}$$

we can define Au, for any $u \in B(\bar{\mathcal{O}})$, as the following distribution,

$$\langle Au, \phi\rangle = \int_{\mathcal{O}} uA^*\phi\, dx \quad \forall \phi \in \mathcal{D}(\mathcal{O}), \tag{2.56}$$

where A^* is the adjoint of A,

$$A^*\phi = -\frac{1}{2}\operatorname{tr}\left[\frac{\partial^2}{\partial x^2}\sigma\sigma^*\phi\right]+\frac{\partial}{\partial x}g\phi+\alpha\phi. \tag{2.57}$$

THEOREM 2.4. *Assume the boundary Γ is smooth, and conditions* (1.1), (1.20), (1.22), (2.1), *and* (2.55) *hold. Then the optimal cost $\hat{u}$ given by* (0.3) *satisfies*

$$A\hat{u} \leqq f \quad \textit{in } \mathcal{D}'(\mathcal{O}). \tag{2.58}$$

Moreover, if (2.14) *and* (2.15) *are true, we also have*

(2.59)[19] $$A\hat{u} = f \quad \textit{in } \mathcal{D}'([\hat{u} < M\hat{u}]).$$

Proof. We need only to use Theorem 3.6 in [7] and Remark 2.1. □

[19] $[\hat{u} < M\hat{u}]$ denotes the subset of $\mathcal{O}$ such that $\hat{u}(x) < M\hat{u}(x)$.

3. Quasi-variational inequality. Let $a_{ij}(x)$, $a_i(x)$ be functions for $i, j = 1, \cdots, N$ such that

$$\begin{gathered}(a_{ij})_{ij} \text{ is a nonnegative symmetric matrix and} \\ a_{ij} \in C^1(\mathbb{R}^N), \qquad \frac{\partial^2 a_{ij}}{\partial x_k \, \partial x_l} \in L^\infty(\mathbb{R}^N) \quad \forall i, j, k, l = 1, \cdots, N,\end{gathered} \tag{3.1}$$

$$a_i \in C(\mathbb{R}^N), \qquad \frac{\partial a_i}{\partial x_k} \in L^\infty(\mathbb{R}^N) \quad \forall i, k = 1, \cdots, N. \tag{3.2}$$

Define the following differential operator A,

$$A = -\sum_{i,j=1}^N \frac{\partial}{\partial x_i} a_{ij} \frac{\partial}{\partial x_j} + \sum_{i=1}^N a_i \frac{\partial}{\partial x_i} + \alpha, \tag{3.3}$$

where α is a positive constant.

We always identify g and σ given by (1.1) as

$$\begin{gathered}(a_{ij})_{ij} = \tfrac{1}{2}\sigma\sigma^*, \\ a_i = \sum_{j=1}^N \frac{\partial a_{ij}}{\partial x_j} - g_i.\end{gathered} \tag{3.4}$$

Let $\beta_0(x)$ and $\beta_1(x)$ be the weight functions $(1+|x|^2)^{-(\lambda+1)/2}$ and $(1+|x|^2)^{-\lambda/2}$, $\lambda > N/2$, respectively.

Introduce the Hilbert spaces

$$H = \{v / \beta_0 v \in L^2(\mathcal{O})\}, \tag{3.5}$$

with the inner product

$$(u, v) = \int_{\mathcal{O}} (\beta_0 u)(\beta_0 v) \, dx \tag{3.6}$$

and the norm $|\cdot|$;

$$V = \left\{ v \in H / \beta_1 \frac{\partial v}{\partial x_k} \in L^2(\mathcal{O}) \ \forall k = 1, \cdots, N \right\} \tag{3.7}$$

with the norm

$$\|v\| = \left(|v|^2 + \sum_{k=1}^N \int_{\mathcal{O}} \left| \beta_1 \frac{\partial v}{\partial x_k} \right|^2 dx \right)^{1/2}. \tag{3.8}$$

V' denotes the dual space of V and $\langle \cdot , \cdot \rangle$ the duality between V' and V.

We have

$$V \subset H \subset V', \quad L^\infty(\mathcal{O}) \subset H, \quad \left\{ v \Big/ \frac{\partial v}{\partial x_i} \in L^\infty(\mathcal{O}) \quad \forall i = 1, \cdots, N \right\} \subset V. \tag{3.9}$$

Let $a(\cdot , \cdot)$ be the bilinear form associated with the operator A,

$$\begin{aligned} a(u, v) = {} & \sum_{i,j=1}^N \int_{\mathcal{O}} \tilde{a}_{ij} \left(\beta_1 \frac{\partial u}{\partial x_i} \right) \left(\beta_1 \frac{\partial v}{\partial x_j} \right) dx \\ & + \sum_{i=1}^N \int_{\mathcal{O}} \tilde{a}_i \left(\beta_1 \frac{\partial u}{\partial x_i} \right) (\beta_0 v) \, dx + \alpha(u, v), \end{aligned} \tag{3.10}$$

where

$$\tilde{a}_{ij}(x)=(1+|x|^2)^{-1}a_{ij}(x),$$
$$\tilde{a}_i(x)=(1+|x|^2)^{-1/2}a_i(x)-2(\lambda+1)(1+|x|^2)^{-3/2}\sum_{j=1}^{N}a_{ij}(x)x_j. \tag{3.11}$$

Notice that a_{ij}, a_i are not supposed to be bounded, but a_{ij} is at most of quadratic growth, and a_i of linear growth. Then, $\tilde{a}_{ij}$, $\tilde{a}_i$ in (3.11) are bounded.

This section is divided into two parts. First, we consider the case where $\mathcal{O}=\mathbb{R}^N$. Next, we study the general case.

3.1. Case $\mathcal{O}=\mathbb{R}^N$. Assume $\mathcal{O}=\mathbb{R}^N$. After some calculation, we deduce

$$a(u,v)=(Au,v)\qquad \forall u,v\in V, Au\in H, \tag{3.12}$$

(3.13)[20] $$|a(u,v)|\leqq C\|u\|\|v\|\qquad \forall u,v\in V,$$

and if α is large enough there exists $\alpha_0>0$ such that

$$a(u,v)\geqq\alpha_0(u,u)\qquad \forall u\in V. \tag{3.14}$$

Next, from (3.12) and (3.13), it follows that

$$a(u,v)=\langle Au,v\rangle\qquad \forall u,v\in V. \tag{3.15}$$

We recall that M denotes the operator given by (1.21). We define, for any $u\in V\cap L^\infty(\mathbb{R}^N)$, the closed cone $K(u)$ in V by

$$K(u)=\{v\in V/v(x)\leqq[Mu](x)\text{ a.e. in }\mathbb{R}^N\}. \tag{3.16}$$

Let us consider the following quasi-variational inequality,

(3.17) Find $u\in V\cap L^\infty(\mathbb{R}^N)$ such that $u\in K(u)$ and
$$a(u,v-u)\geqq(f,v-u)\qquad \forall v\in K(u),$$

and also the sequence of variational inequalities

(3.18) Find $u^0\in V$ such that $a(u^0,v)=(f,v)\quad \forall v\in V.$

(3.19) Find $u^n\in V\cap L^\infty(\mathbb{R}^N)$ such that $u^n\in K(u^{n-1})$ and
$$a(u^n,v-u^n)\geqq(f,v-u^n)\qquad \forall v\in K(u^{n-1}).$$

We have

THEOREM 3.1. *Let the assumptions* (3.1), (3.2), (1.20), (2.1), *and*

$$\frac{\partial f}{\partial x_k}\in L^\infty(\mathbb{R}^N),\qquad k=1,\cdots,N \tag{3.20}$$

hold. Then the quasi-variational inequality (3.17) *admits a maximum solution $\hat{u}$ which is given as the optimal cost* (0.3). *Moreover, $\hat{u}$ is Lipschitz continuous and the following estimates are true.*

(3.21)[21] $$\left\|\frac{\partial\hat{u}}{\partial x}\right\|_{L^\infty}\leqq\frac{1}{\alpha-\gamma_0}\left\|\frac{\partial f}{\partial x}\right\|_{L^\infty},$$

[20] C denotes a constant.

[21] $\|\partial\hat{u}/\partial x\|_{L^\infty}$ denotes the smallest Lipschitz continuous constant of $\hat{u}$, and γ_0 is given by (1.19).

$$0 \leqq u^n - \hat{u} \leqq C(n+1)^{-1}, \qquad n = 0, 1, \cdots, \tag{3.22}$$

where the constant C depends only on the supremum norm of f and α, k_0.[22]

Proof. First, from Theorem 4.1 in [7], the sequence defined by (3.18), (3.19) coincides with that defined by (2.12), (2.13).

Then, from (2.17), we have

$$|u^n(x) - u^n(x')| \leqq \sup \{|J_x(\nu) - J_{x'}(\nu)| / \nu \text{ an impulse control such that } \theta_i = \infty \ \forall\, i \geqq n+1\}.$$

Hence, Lemma 1.1 and (3.20) imply

$$\left\| \frac{\partial u^n}{\partial x} \right\|_{L^\infty} \leqq \frac{1}{\alpha - \gamma_0} \left\| \frac{\partial f}{\partial x} \right\|_{L^\infty} \qquad \forall n = 0, 1, 2, \cdots. \tag{3.23}$$

Thus, using Theorem 2.2 and classical technique, the proof is completed. □

Remark 3.1. Clearly, using only analytic methods, like B. Hanouzet and J. L. Joly [3], we can prove that (Remark 2.2)

$$0 \leqq u^n - \hat{u} \leqq cq^n, \qquad n = 0, 1, \cdots, \quad \text{with } 0 < q < 1. \qquad \square \tag{3.24}$$

3.2. General case. Now, we come back to the general case, $\mathcal{O}$ an open subset of $\mathbb{R}^N$ with boundary Γ sufficiently smooth.

Define the closed subspace of V,

$$V_0 = \{v \in V / v = 0 \text{ on } \Gamma\}. \tag{3.25}$$

Then, as in the case $\mathcal{O} = \mathbb{R}^N$, if α is large enough there exists a constant $\alpha_0 > 0$ such that

$$a(u, u) \geqq \alpha_0(u, u) \quad \forall u \in V_0, \tag{3.26}$$

and we also have

$$a(u, v) = \langle Au, v \rangle \quad \forall u, v \in V_0. \tag{3.27}$$

For any $u \in V_0 \cap L^\infty(\mathcal{O})$, we define $K_0(u)$, the following closed cone in V_0 by

$$K_0(u) = \{v \in V_0 / v \leqq Mu, \text{ a.e. in } \mathcal{O}\}. \tag{3.28}$$

Let us consider the quasi-variational inequality

$$\begin{aligned} &\text{Find } u \in V_0 \cap L^\infty(\mathcal{O}) \text{ such that } u \in K_0(u) \text{ and} \\ &a(u, v-u) \geqq (f, v-u) \quad \forall v \in K_0(u), \end{aligned} \tag{3.29}$$

and the associated sequence of variational inequalities,

$$\text{Find } u^0 \in V_0 \text{ such that } a(u^0, v) = (f, v) \quad \forall v \in V_0. \tag{3.30}$$

$$\begin{aligned} &\text{Find } u^n \in V_0 \cap L^\infty(\mathcal{O}) \text{ such that } u^n \in K_0(u^{n-1}) \text{ and} \\ &a(u^n, v-u^n) \geqq (f, v-u^n) \quad \forall v \in K_0(u^{n-1}). \end{aligned} \tag{3.31}$$

[22] k_0 is given in (1.20).

Remark 3.2. Assume (2.1). Suppose that $\mathcal{O}$ is bounded and satisfies the uniform exterior sphere condition of radius $\rho > 0$, and that

$$\begin{aligned} \Gamma = \{x \in \Gamma / |\sigma(x) n(x)| > 0\} \\ \cup \{x \in \Gamma / 2g(x)n(x) < -\mathrm{tr}\,(\sigma(x)\sigma^*(x))\}, \\ n(x) \text{ is the inner normal with modulus } \rho. \end{aligned} \tag{3.32}$$

Then, there exists a Lipschitz continuous subsolution

$$\begin{aligned} &\bar{w} \in C(\bar{\mathcal{O}});\ \bar{w}, \frac{\partial \bar{w}}{\partial x_i} \in L^\infty(\mathcal{O}), \qquad i = 1, \cdots, N, \\ &A\bar{w} \leqq -f \text{ in } \mathcal{O}, \qquad \bar{w}(x) = 0 \quad \forall x \in \Gamma. \end{aligned} \tag{3.33$)^{23}$}$$

Indeed, we only need to use Lemma 1.5 in [7].

THEOREM 3.2. *Let the conditions* (3.1), (3.2), (1.20), (1.22), (2.1), (3.33) *and*[24]

$$\frac{\partial f}{\partial x_k} \in L^\infty(\mathcal{O}), \qquad k = 1, \cdots, N, \tag{3.34}$$

hold. Then the quasi-variational inequality (3.29) *admits a maximum solution $\hat{u}$ which is given as the optimal cost* (0.3). *Moreover, $\hat{u}$ is Lipschitz continuous and the estimates* (3.22) *and*

$$\left\|\frac{\partial \hat{u}}{\partial x}\right\|_{L^\infty} \leqq \frac{1}{\alpha - \gamma_0} \left\|\frac{\partial f}{\partial x}\right\|_{L^\infty} + \left\|\frac{\partial \bar{w}}{\partial x}\right\|_{L^\infty} \tag{3.35}$$

are true.

Proof. As for Theorem 3.1, we just need to prove the following estimate,

$$\left\|\frac{\partial u^n}{\partial x}\right\|_{L^\infty} \leqq \frac{1}{\alpha - \gamma_0} \left\|\frac{\partial f}{\partial x}\right\|_{L^\infty} + \left\|\frac{\partial \bar{w}}{\partial x}\right\|_{L^\infty}, \qquad n = 0, 1, \cdots. \tag{3.36}$$

Indeed, starting at

$$u^n(x) - u^n(x') = \sup_{\nu'} \inf_{\nu} [J_x(\nu) - J_{x'}(\nu')], \tag{3.37}$$

we set, for any $\nu' = \{\theta'_i, \xi'_i\}_{i=1}^\infty$, the impulse control $\nu = \{\theta_i, \xi_i\}_{i=1}^\infty$ defined by (1.2) and

$$\tau_x^i = \inf\{t \geqq 0 / y_x^i(t) \notin \bar{\mathcal{O}}\}, \qquad i = 0, 1, \cdots; \tag{3.38}$$

$$\theta_i = \begin{cases} \theta'_i & \text{if } \theta'_i < \tau_x^{i-1} \wedge \tau_{x'}^{i-1}, \\ \infty & \text{otherwise}; \end{cases} \tag{3.39$)^{25}$}$$

$$\xi_i = \begin{cases} \xi'_i & \text{if } \theta_i < \infty \quad \text{and} \quad \xi'_i + y_x^{i-1}(\theta_i) \in \bar{\mathcal{O}}, \\ 0 & \text{if } \theta_i = \infty, \\ \lambda \xi'_i & \text{if } \theta_i < \infty \quad \text{and} \quad \lambda \xi'_i + y_x^{i-1}(\theta_i) \in \Gamma; \end{cases} \tag{3.40}$$

$$\begin{aligned} &dy^i(t) = g(y^i(t))\,dt + \sigma(y^i(t))\,dw(t), \qquad t \geqq \theta_i, \\ &y^i(\theta_i) = y^{i-1}(\theta_i) + \xi_i, \\ &y^i(t) = y^{i-1}(t), \qquad t < \theta_i. \end{aligned} \tag{3.41}$$

[23] In the martingale sense with α large enough.

[24] We also assume α large enough and $k(\lambda\xi) \leqq k(\xi)$, $\forall \xi \in \Lambda$, $\lambda \in [0, 1]$.

[25] $\tau_{x'}^i$ is given as τ_x^i in (3.38).

Notice that ξ_i is well defined, because if $\xi_i' + y_x^{i-1}(\theta_i) \notin \bar{\mathcal{O}}$ and $\theta_i < \infty$ we have $y^{i-1}(\theta_i) \in \bar{\mathcal{O}}$, and so there exists $\lambda \in [0, 1]$ such that $\lambda \xi_i' + y_x^{i-1}(\theta_i) \in \Gamma = \Gamma_0$.

Thus, ν is an admissible impulse control for x, and choosing ν as above in (3.37), we deduce

$$\begin{aligned} u^n(x) - u^n(x') \leqq \sup_{\nu'} E\left\{ \int_{\tau_x \wedge \tau_{x'}}^{\tau_x} f(y_x(t, \nu))\, e^{-\alpha t}\, dt \right\} \\ + \sup_{\nu'} E\left\{ \int_0^{\tau_x \wedge \tau_{x'}} |f(y_x(t, \nu)) - f(y_{x'}(t, \nu'))|\, e^{-\alpha t}\, dt \right\}, \end{aligned} \tag{3.42}$$

where the supremum is taken over all admissible impulse controls ν'.

Finally, from Lemma 1.2 and the fact that

$$y_x(t, \nu) = y_x(t, \nu'), \quad \text{a.s. in } [0, \tau_x \wedge \tau_{x'}[, \tag{3.43}$$

the estimate (3.36) follows from (3.42). □

THEOREM 3.3. *Under the conditions of Theorem* 3.2, *the following quasi-variational inequality*

$$\begin{aligned} &\hat{u} \in W_0^{1,\infty}(\mathcal{O}), \qquad \hat{u} \leqq M\hat{u} \text{ in } \mathcal{O}, \\ &A\hat{u} \leqq f \text{ in } \mathcal{D}'(\mathcal{O}), \qquad A\hat{u} = f \text{ in } \mathcal{D}'([\hat{u} < M\hat{u}]), \end{aligned} \tag{3.44}$$

has one and only one solution $\hat{u}$. *Moreover*, $\hat{u}$ *is given as the optimal cost* (0.3).

Proof. We only need to prove the uniqueness of problem (3.44). Moreover, it suffices to show that any solution of (3.44) is a solution of (2.46).

Indeed, using a classical technique (cf. D. W. Stroock and S. R. S. Varadhan [12]), we can prove that if $\hat{u}$ verifies

$$\hat{u} \in W_0^{1,\infty}(\mathcal{O}), \qquad A\hat{u} = f \quad \text{in } \mathcal{D}'([\hat{u} < M\hat{u}]),$$

then we also have

$$A\hat{u} = f \text{ in the martingale sense on } [\hat{u} < M\hat{u}].$$

Therefore, as in Corollary 2.1, we obtain the equality (2.46) and the theorem is established. □

Remark 3.3. It is possible to consider a function $a_0(x)$ instead of the constant α for the definition of cost (0.2). Moreover, we can also consider f not necessarily bounded and $k = k(x, \xi)$.

Remark 3.4. All these results can be extended to the parabolic case.

Remark 3.5. In [9], we give an application to the impulse control problems with partial information.

Final Remark. In a separate paper (cf. [8]) the stopping time and impulse control problems for degenerate diffusions with boundary conditions will be studied.

REFERENCES

[1] A. BENSOUSSAN, *On the semigroup formulation of variational inequalities and quasi-variational inequalities*, First Franco-Southeast Asian Mathematical Conference, Singapore, May-1979.

[2] A. BENSOUSSAN AND J. L. LIONS, *Contrôle impulsionnel et inéquations quasi-variationnelles*, Dunod, Paris, to be published.

[3] B. HANOUZET AND J. L. JOLY, *Convergence uniforme des itérés définissant la solution d'une inéquation quasi-variationnelle abstraite*, C.R. Acad. Sci. Paris Sér. A, 286 (1978), pp. 735–738.

[4] J. P. LEPELTIER AND B. MARCHAL, *Techniques probabilistes dans le contrôle impulsionnel*, Stochastics, 2 (1979), pp. 243–286.

[5] J. L. MENALDI, *Sur le problème de contrôle impulsionnel et l'inéquation quasi-variationnelle dégénérée associée*, C.R. Acad. Sci. Paris Sér. A, 284 (1977), pp. 1499–1502.

[6] ———, *Le problème de contrôle impulsionnel déterministe et l'inéquation quasi-variationnelle de premier ordre associée*, Appl. Math. Optim., to appear.

[7] ——, *On the optimal stopping time problem for degenerate diffusions*, this Journal, this issue, pp. 697–721.

[8] ———, *On the degenerate variational inequality with Neumann boundary conditions*, submitted to J. Optim. Theory Appl.

[9] ——, *The separation principle for impulse control problems*, Proc. Amer. Math. Soc., to appear.

[10] F. MIGNOT AND J. P. PUEL, *Inéquations variationnelles et quasi-variationnelles hyperboliques du premier ordre*, J. Math. Pures et Appl., 55 (1976), pp. 353–378.

[11] M. ROBIN, *Contrôle impulsionnel des processus de Markov*, Thèse d'Etat, Université de Paris IX, Paris, 1977.

[12] D. W. STROOCK AND S. R. S. VARADHAN, *On degenerate elliptic-parabolic operators of second order and their associated diffusions*, Comm. Pure Appl. Math., 25 (1972), pp. 651–713.

[13] L. TARTAR, *Inéquations quasi-variationnelles abstraites*, C.R. Acad. Sci. Paris Sér. A, 278 (1974), pp. 1193–1196.